About

CS 189 is the Machine Learning course at UC Berkeley. In this guide we have created a comprehensive course guide in order to share our knowledge with students and the general public, and hopefully draw the interest of students from other universities to Berkeley’s Machine Learning curriculum.

We owe gratitude to Professor Anant Sahai and Professor Stella Yu, as this book is heavily inspired from their lectures. In addition, we are indebted to Professor Jonathan Shewchuk for his machine learning notes, from which we drew inspiration.

The latest version of this document can be found at http://snasiriany.me/cs189/. Please report any mistakes to snasiriany@berkeley.edu. Please contact the authors if you wish to redistribute this document.

Notation

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{R}$</td>
<td>set of real numbers</td>
</tr>
<tr>
<td>$\mathbb{R}^n$</td>
<td>set (vector space) of $n$-tuples of real numbers, endowed with the usual inner product</td>
</tr>
<tr>
<td>$\mathbb{R}^{m \times n}$</td>
<td>set (vector space) of $m$-by-$n$ matrices</td>
</tr>
<tr>
<td>$\delta_{ij}$</td>
<td>Kronecker delta, i.e. $\delta_{ij} = 1$ if $i = j$, 0 otherwise</td>
</tr>
<tr>
<td>$\nabla f(x)$</td>
<td>gradient of the function $f$ at $x$</td>
</tr>
<tr>
<td>$\nabla^2 f(x)$</td>
<td>Hessian of the function $f$ at $x$</td>
</tr>
<tr>
<td>$p(X)$</td>
<td>distribution of random variable $X$</td>
</tr>
<tr>
<td>$p(x)$</td>
<td>probability density/mass function evaluated at $x$</td>
</tr>
<tr>
<td>$\mathbb{E}[X]$</td>
<td>expected value of random variable $X$</td>
</tr>
<tr>
<td>$\text{Var}(X)$</td>
<td>variance of random variable $X$</td>
</tr>
<tr>
<td>$\text{Cov}(X,Y)$</td>
<td>covariance of random variables $X$ and $Y$</td>
</tr>
</tbody>
</table>

Other notes:

- Vectors and matrices are in bold (e.g. $\mathbf{x}, \mathbf{A}$). This is true for vectors in $\mathbb{R}^n$ as well as for vectors in general vector spaces. We generally use Greek letters for scalars and capital Roman letters for matrices and random variables.

- We assume that vectors are column vectors, i.e. that a vector in $\mathbb{R}^n$ can be interpreted as an $n$-by-1 matrix. As such, taking the transpose of a vector is well-defined (and produces a row vector, which is a 1-by-$n$ matrix).
## 5 Classification 101

5.1 Generative vs. Discriminative Classification ........................................ 101
5.2 Least Squares Support Vector Machine ................................................... 103
5.3 Logistic Regression .............................................................................. 107
5.4 Gaussian Discriminant Analysis .............................................................. 115
5.5 Support Vector Machines ...................................................................... 121
5.6 Generalization and Stability .................................................................. 127
5.7 Duality .................................................................................................... 128
5.8 Nearest Neighbor Classification .............................................................. 139

## 6 Clustering 145

6.1 K-means Clustering ............................................................................... 146
6.2 Mixture of Gaussians ............................................................................ 149
6.3 Expectation Maximization (EM) Algorithm .......................................... 150

## 7 Decision Tree Learning 157

7.1 Decision Trees ........................................................................................ 157
7.2 Random Forests ..................................................................................... 162
7.3 Boosting .................................................................................................. 163

## 8 Deep Learning 169

8.1 Convolutional Neural Networks .............................................................. 169
8.2 CNN Architectures ............................................................................... 175
8.3 Visualizing and Understanding CNNs .................................................... 178
Chapter 1

Regression I

Our goal in machine learning is to extract a relationship from data. In regression tasks, this relationship takes the form of a function \( y = f(x) \), where \( y \in \mathbb{R} \) is some quantity that can be predicted from an input \( x \in \mathbb{R}^d \), which should for the time being be thought of as some collection of numerical measurements. The true relationship \( f \) is unknown to us, and our aim is to recover it as well as we can from data. Our end product is a function \( \hat{y} = h(x) \), called the hypothesis, that should approximate \( f \). We assume that we have access to a dataset \( D = \{(x_i, y_i)\}_{i=1}^n \), where each pair \((x_i, y_i)\) is an example (possibly noisy or otherwise approximate) of the input-output mapping to be learned. Since learning arbitrary functions is intractable, we restrict ourselves to some hypothesis class \( \mathcal{H} \) of allowable functions. More specifically, we typically employ a parametric model, meaning that there is some finite-dimensional vector \( w \in \mathbb{R}^d \), the elements of which are known as parameters or weights, that controls the behavior of the function. That is,

\[
    h_w(x) = g(x, w)
\]

for some other function \( g \). The hypothesis class is then the set of all functions induced by the possible choices of the parameters \( w \):

\[
    \mathcal{H} = \{h_w \mid w \in \mathbb{R}^d\}
\]

After designating a cost function \( L \), which measures how poorly the predictions \( \hat{y} \) of the hypothesis match the true output \( y \), we can proceed to search for the parameters that best fit the data by minimizing this function:

\[
    w^* = \arg \min_w L(w)
\]

1.1 Ordinary Least Squares

Ordinary least squares (OLS) is one of the simplest regression problems, but it is well-understood and practically useful. It is a linear regression problem, which means that we take \( h_w \) to be of the form \( h_w(x) = x^\top w \). We want

\[
    y_i \approx \hat{y}_i = h_w(x_i) = x_i^\top w
\]
for each \( i = 1, \ldots, n \). This set of equations can be written in matrix form as

\[
\begin{bmatrix}
    y_1 \\
    \vdots \\
    y_n
\end{bmatrix}
\approx
\begin{bmatrix}
    x_1^\top \\
    \vdots \\
    x_n^\top
\end{bmatrix}
\begin{bmatrix}
    w_1 \\
    \vdots \\
    w_d
\end{bmatrix}
\]

In words, the matrix \( X \in \mathbb{R}^{n \times d} \) has the input datapoint \( x_i \) as its \( i \)th row. This matrix is sometimes called the design matrix. Usually \( n \geq d \), meaning that there are more datapoints than measurements.

There will in general be no exact solution to the equation \( y = Xw \) (even if the data were perfect, consider how many equations and variables there are), but we can find an approximate solution by minimizing the sum (or equivalently, the mean) of the squared errors:

\[
L(w) = \sum_{i=1}^{n} (x_i^\top w - y_i)^2 = \min_w \|Xw - y\|_2^2
\]

Now that we have formulated an optimization problem, we want to go about solving it. We will see that the particular structure of OLS allows us to compute a closed-form expression for a globally optimal solution, which we denote \( w_{\text{ols}}^* \).

**Approach 1: Vector calculus**

Calculus is the primary mathematical workhorse for studying the optimization of differentiable functions. Recall the following important result: if \( L : \mathbb{R}^d \to \mathbb{R} \) is continuously differentiable, then any local optimum \( w^* \) satisfies \( \nabla L(w^*) = 0 \). In the OLS case,

\[
L(w) = \|Xw - y\|_2^2 = (Xw - y)^\top (Xw - y) = (Xw)^\top Xw - (Xw)^\top y - y^\top Xw + y^\top y = w^\top X^\top Xw - 2w^\top X^\top y + y^\top y
\]

Using the following results from matrix calculus

\[
\nabla_x (a^\top x) = a \\
\nabla_x (x^\top Ax) = (A + A^\top)x
\]

the gradient of \( L \) is easily seen to be

\[
\nabla L(w) = \nabla_w (w^\top X^\top Xw - 2w^\top X^\top y + y^\top y) = \nabla_w (w^\top X^\top Xw) - 2\nabla_w (w^\top X^\top y) + \underbrace{\nabla_w (y^\top y)}_{0} = 2X^\top Xw - 2X^\top y
\]

where in the last line we have used the symmetry of \( X^\top X \) to simplify \( X^\top X + (X^\top X)^\top = 2X^\top X \).

Setting the gradient to 0, we conclude that any optimum \( w_{\text{ols}}^* \) satisfies

\[
X^\top Xw_{\text{ols}}^* = X^\top y
\]
If $X$ is full rank, then $X^\top X$ is as well (assuming $n \geq d$), so we can solve for a unique solution

$$w_{\text{ols}}^* = (X^\top X)^{-1}X^\top y$$

Note: Although we write $(X^\top X)^{-1}$, in practice one would not actually compute the inverse; it is more numerically stable to solve the linear system of equations above (e.g. with Gaussian elimination).

In this derivation we have used the condition $\nabla L(w^*) = 0$, which is a necessary but not sufficient condition for optimality. We found a critical point, but in general such a point could be a local minimum, a local maximum, or a saddle point. Fortunately, in this case the objective function is convex, which implies that any critical point is indeed a global minimum. To show that $L$ is convex, it suffices to compute the Hessian of $L$, which in this case is

$$\nabla^2 L(w) = 2X^\top X$$

and show that this is positive semi-definite:

$$\forall w, \ w^\top (2X^\top X)w = 2(Xw)^\top Xw = 2\|Xw\|^2 \geq 0$$

**Approach 2: Orthogonal projection**

There is also a linear algebraic way to arrive at the same solution: orthogonal projections.

Recall that if $V$ is an inner product space and $S$ a subspace of $V$, then any $v \in V$ can be decomposed uniquely in the form

$$v = v_S + v_\perp$$

where $v_S \in S$ and $v_\perp \in S^\perp$. Here $S^\perp$ is the orthogonal complement of $S$, i.e. the set of vectors that are perpendicular to every vector in $S$.

The orthogonal projection onto $S$, denoted $P_S$, is the linear operator that maps $v$ to $v_S$ in the decomposition above. An important property of the orthogonal projection is that

$$\|v - P_S v\| \leq \|v - s\|$$

for all $s \in S$, with equality if and only if $s = P_s v$. That is,

$$P_S = \arg \min_{s \in S} \|v - s\|$$

**Proof.** By the Pythagorean theorem,

$$\|v - s\|^2 = \|v - P_S v + P_S v - s\|^2 = \|v - P_S v\|^2 + \|P_S v - s\|^2 \geq \|v - P_S v\|^2$$

with equality holding if and only if $\|P_S v - s\|^2 = 0$, i.e. $s = P_S v$. Taking square roots on both sides gives $\|v - s\| \geq \|v - P_S v\|$ as claimed (since norms are nonnegative). 

Here is a visual representation of the argument above:
In the OLS case,

\[ w_{\text{OLS}}^* = \arg \min_w \|Xw - y\|_2^2 \]

But observe that the set of vectors that can be written \( Xw \) for some \( w \in \mathbb{R}^d \) is precisely the range of \( X \), which we know to be a subspace of \( \mathbb{R}^n \), so

\[ \min_{z \in \text{range}(X)} \|z - y\|_2^2 = \min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 \]

By pattern matching with the earlier optimality statement about \( P_S \), we observe that \( P_{\text{range}(X)}y = Xw_{\text{OLS}}^* \), where \( w_{\text{OLS}}^* \) is any optimum for the right-hand side. The projected point \( Xw_{\text{OLS}}^* \) is always unique, but if \( X \) is full rank (again assuming \( n \geq d \)), then the optimum \( w_{\text{OLS}}^* \) is also unique (as expected). This is because \( X \) being full rank means that the columns of \( X \) are linearly independent, in which case there is a one-to-one correspondence between \( w \) and \( Xw \).

To solve for \( w_{\text{OLS}}^* \), we need the following fact\(^1\):

\[ \text{null}(X^\top) = \text{range}(X)^\perp \]

Since we are projecting onto \( \text{range}(X) \), the orthogonality condition for optimality is that \( y - Py \perp \text{range}(X) \), i.e. \( y - Xw_{\text{OLS}}^* \in \text{null}(X^\top) \). This leads to the equation

\[ X^\top(y - Xw_{\text{OLS}}^*) = 0 \]

which is equivalent to

\[ X^\top Xw_{\text{OLS}}^* = X^\top y \]

as before.

### 1.2 Ridge Regression

While Ordinary Least Squares can be used for solving linear least squares problems, it falls short due to numerical instability and generalization issues. Numerical instability arises when the features of the data are close to collinear (leading to linearly dependent feature columns), causing the input

\(^1\) This result is often stated as part of the Fundamental Theorem of Linear Algebra.
matrix $X$ to lose its rank or have singular values that very close to 0. Why are small singular values bad? Let us illustrate this via the singular value decomposition (SVD) of $X$:

$$X = U\Sigma V^\top$$

where $U \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{n \times d}$, $V \in \mathbb{R}^{d \times d}$. In the context of OLS, we must have that $X^\top X$ is invertible, or equivalently, $\text{rank}(X) = \text{rank}(X^\top) = \text{rank}(X) = d$. Assuming that $X$ and $X^\top$ are full column rank $d$, we can express the SVD of $X$ as

$$X = U \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix} V^\top$$

where $\Sigma_d \in \mathbb{R}^{d \times d}$ is a diagonal matrix with strictly positive entries. Now let’s try to expand the $(X^\top X)^{-1}$ term in OLS using the SVD of $X$:

$$(X^\top X)^{-1} = (V \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix} U^\top \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix} V^\top)^{-1}$$

$$= (V \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix} 0 \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix} V^\top)^{-1}$$

$$= (V \Sigma_d V^\top)^{-1} = (V^\top)^{-1} (\Sigma_d^2)^{-1} V^{-1} = V \Sigma_d^{-2} V^\top$$

This means that $(X^\top X)^{-1}$ will have singular values that are the squared inverse of the singular values of $X$, potentially leading to extremely large singular values when the singular value of $X$ are close to 0. Such excessively large singular values can be very problematic for numerical stability purposes. In addition, abnormally high values to the optimal $w$ solution would prevent OLS from generalizing to unseen data.

There is a very simple solution to these issues: penalize the entries of $w$ from becoming too large. We can do this by adding a penalty term constraining the norm of $w$. For a fixed, small scalar $\lambda > 0$, we now have:

$$\min_w \|Xw - y\|^2_2 + \lambda \|w\|^2_2$$

Note that the $\lambda$ in our objective function is a hyperparameter that measures the sensitivity to the values in $w$. Just like the degree in polynomial features, $\lambda$ is a value that we must choose arbitrarily through validation. Let’s expand the terms of the objective function:

$$L(w) = \|Xw - y\|^2_2 + \lambda \|w\|^2_2$$

$$= w^\top X^\top Xw - 2w^\top X^\top y + y^\top y + \lambda w^\top w$$

Finally take the gradient of the objective and find the value of $w$ that achieves 0 for the gradient:

$$\nabla_w L(w) = 0$$

$$2X^\top Xw - 2X^\top y + 2\lambda w = 0$$

$$(X^\top X + \lambda I)w = X^\top y$$

$$w^*_{\text{ridge}} = (X^\top X + \lambda I)^{-1}X^\top y$$

This value is guaranteed to achieve the (unique) global minimum, because the objective function is strongly convex. To show that $f$ is strongly convex, it suffices to compute the Hessian of $f$, which in this case is

$$\nabla^2 L(w) = 2X^\top X + 2\lambda I$$
and show that this is positive definite (PD):

\[ \forall w \neq 0, \ w^\top (X^\top X + \lambda I) w = (Xw)^\top Xw + \lambda w^\top w = \|Xw\|_2^2 + \lambda \|w\|_2^2 > 0 \]

Since the Hessian is positive definite, we can equivalently say that the eigenvalues of the Hessian are strictly positive and that the objective function is strongly convex. A useful property of strongly convex functions is that they have a unique optimum point, so the solution to ridge regression is unique. We cannot make such guarantees about ordinary least squares, because the corresponding Hessian could have eigenvalues that are 0. Let us explore the case in OLS when the Hessian has a 0 eigenvalue. In this context, the term \(X^\top X\) is not invertible, but this does not imply that no solution exists! In OLS, there always exists a solution, and when the Hessian is PD that solution is unique; when the Hessian is PSD, there are infinitely many solutions. (There always exists a solution to the expression \(X^\top Xw = X^\top y\), because the range of \(X^\top X\) and the range space of \(X^\top\) are equivalent; since \(X^\top y\) lies in the range of \(X^\top\), it must equivalently lie in the range of \(X^\top X\) and therefore there always exists a \(w\) that satisfies the equation \(X^\top Xw = X^\top y\).)

The technique we just described is known as ridge regression. Note that now the expression \(X^\top X + \lambda I\) is invertible, regardless of rank of \(X\). Let’s find \((X^\top X + \lambda I)^{-1}\) through SVD:

\[
(X^\top X + \lambda I)^{-1} = \left( V \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} U^\top U \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} V^\top + \lambda I \right)^{-1} \\
= \left( V \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} V^\top + \lambda I \right)^{-1} \\
= \left( V \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} V^\top + V(\lambda I)V^\top \right)^{-1} \\
= \left( V \begin{bmatrix} \Sigma_r^2 & 0 \\ 0 & 0 \end{bmatrix} + \lambda I \right)^{-1} \\
= \left( V \begin{bmatrix} \Sigma_r^2 + \lambda I & 0 \\ 0 & \lambda I \end{bmatrix} V^\top \right)^{-1} \\
= (V^\top)^{-1} \begin{bmatrix} \Sigma_r^2 + \lambda I & 0 \\ 0 & \lambda I \end{bmatrix}^{-1} V^{-1} \\
= V \begin{bmatrix} (\Sigma_r^2 + \lambda I)^{-1} & 0 \\ 0 & \frac{1}{\lambda} I \end{bmatrix} V^\top
\]

Now with our slight tweak, the matrix \(X^\top X + \lambda I\) has become full rank and thus invertible. The singular values have become \(\frac{1}{\sigma^2 + \lambda}\) and \(\frac{1}{\lambda}\), meaning that the singular values are guaranteed to be at most \(\frac{1}{\lambda}\), solving our numerical instability issues. Furthermore, we have partially solved the overfitting issue. By penalizing the norm of \(x\), we encourage the weights corresponding to relevant features that capture the main structure of the true model, and penalize the weights corresponding to complex features that only serve to fine tune the model and fit noise in the data.
1.3 Feature Engineering

We’ve seen that the least-squares optimization problem

$$\min_w \|Xw - y\|_2^2$$

represents the “best-fit” linear model, by projecting $y$ onto the subspace spanned by the columns of $X$. However, the true input-output relationship $y = f(x)$ may be nonlinear, so it is useful to consider nonlinear models as well. It turns out that we can still do this under the framework of linear least-squares, by augmenting the data with new features. In particular, we devise some function $\phi : \mathbb{R}^\ell \to \mathbb{R}^d$, called a feature map, that maps each raw data point $x \in \mathbb{R}^\ell$ into a vector of features $\phi(x)$. The hypothesis function then writes

$$h_w(x) = \sum_{j=1}^d w_j \phi_j(x) = w^T \phi(x)$$

Note that the resulting model is still linear with respect to the features, but it is nonlinear with respect to the original data if $\phi$ is nonlinear. The component functions $\phi_j$ are sometimes called basis functions because our hypothesis is a linear combination of them. In the simplest case, we could just use the components of $x$ as features (i.e. $\phi_j(x) = x_j$), but in general it is helpful to disambiguate the features of an example from the example’s entries.

We can then use least-squares to estimate the weights $w$, just as before. To do this, we replace the original data matrix $X \in \mathbb{R}^{n \times \ell}$ by $\Phi \in \mathbb{R}^{n \times d}$, which has $\phi(x_i)^T$ as its $i$th row:

$$\min_w \|\Phi w - y\|_2^2$$

Example: Fitting Ellipses

Let’s use least-squares to estimate the parameters of an ellipse from data.

Assume that we have $n$ data points $D = \{(x_{1,i}, x_{2,i})\}_{i=1}^n$, which may be noisy (i.e. could be off the actual orbit). Our goal is to determine the relationship between $x_1$ and $x_2$.

We assume that the ellipse from which the points were generated has the form

$$w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_1 + w_5 x_2 = 1$$

where the coefficients $w_1, \ldots, w_5$ are the parameters we wish to estimate.

We formulate the problem with least-squares:

$$\min_w \|\Phi w - 1\|_2^2$$

where

$$\Phi = \begin{bmatrix}
  x_{1,1} & x_{2,1} & x_{1,1} x_{2,1} & x_{1,1} & x_{2,1} \\
  x_{1,2} & x_{2,2} & x_{1,2} x_{2,2} & x_{1,2} & x_{2,2} \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  x_{1,n} & x_{2,n} & x_{1,n} x_{2,n} & x_{1,n} & x_{2,n}
\end{bmatrix}$$

In this case, the feature map $\phi$ is given by

$$\phi(x) = (x_1^2, x_2^2, x_1 x_2, x_1, x_2)$$

Note that there is no “target” vector $y$ here, so this is not a traditional regression problem, but it still fits into the framework of least-squares.
Polynomial Features

The example above demonstrates an important class of features known as polynomial features. Remember that a polynomial is a linear combination of monomial basis terms. Monomials can be classified in two ways, by their degree and dimension:

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Degree</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (univariate)</td>
<td>1</td>
<td>$x$</td>
<td>$x^2$</td>
<td>$x^3$</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>2 (bivariate)</td>
<td>1</td>
<td>$x_1$, $x_2$</td>
<td>$x_1^2$, $x_2^2$, $x_1x_2$</td>
<td>$x_1^3$, $x_2^3$, $x_1^2x_2$, $x_1x_2^2$</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

A big reason we care polynomial features is that any smooth function can be approximated arbitrarily closely by some polynomial. For this reason, polynomials are said to be universal approximators.

One downside of polynomials is that as their degree increases, their number of terms increases rapidly. Specifically, one can use a “stars and bars” style combinatorial argument to show that a polynomial of degree $d$ in $\ell$ variables has

$$\binom{\ell + d}{\ell} = \frac{(\ell + d)!}{\ell!d!}$$

terms. To get an idea for how quickly this quantity grows, consider a few examples:

<table>
<thead>
<tr>
<th>$\ell$</th>
<th>$d$</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>11</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>20</td>
<td>56</td>
<td>286</td>
<td>3276</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>56</td>
<td>252</td>
<td>3003</td>
<td>142506</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>286</td>
<td>3003</td>
<td>184756</td>
<td>183579396</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>26</td>
<td>3276</td>
<td>142506</td>
<td>183579396</td>
<td>126410606437752</td>
<td></td>
</tr>
</tbody>
</table>

Later we will learn about the kernel trick, a clever mathematical method that allows us to circumvent this rapidly growing cost in certain cases.

### 1.4 Hyperparameters and Validation

As above, consider a hypothesis of the form

$$h_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{d} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

2 Taylor's theorem gives more precise statements about the approximation error.
3 We count the number of distinct monomials of degree at most $d$ in $\ell$ variables $x_1, \ldots, x_\ell$, or equivalently, the number of distinct monomials of degree exactly $d$ in $\ell + 1$ variables $x_0, x_1, \ldots, x_\ell$. Every monomial has the form $x_0^{k_0} \cdots x_\ell^{k_\ell}$ where $k_0 + \cdots + k_\ell = d$. This corresponds to an arrangement of $d$ stars and $\ell$ bars, where the number of stars between consecutive bars (or the ends of the expression) gives the degree of that ordered variable. For example,

$$\begin{align*}
*|**|*|** & \leftrightarrow & x_0^3x_1x_2^2
\end{align*}$$

The number of unique ways to arrange these stars and bars is the number of ways to choose the positions of the $\ell$ bars out of the total $\ell + d$ slots, i.e. $\ell + d$ choose $\ell$. (You could also pick the positions of the $d$ stars out of the total $\ell + d$ slots; the expression is symmetric in $\ell$ and $d$.)
Observe that the model order $d$ is not one of the decision variables being optimized when we fit to the data. For this reason $d$ is called a hyperparameter. We might say more specifically that it is a model hyperparameter, since it determines the structure of the model.

For another example, recall ridge regression, in which we add an $\ell^2$ penalty on the parameters $w$:

$$\min_w \|Xw - y\|_2^2 + \lambda\|w\|_2^2$$

The regularization weight $\lambda$ is also a hyperparameter, as it is fixed during the minimization above. However $\lambda$, unlike the previously discussed hyperparameter $d$, is not a part of the model. Rather, it is an aspect of the optimization procedure used to fit the model, so we say it is an optimization hyperparameter. Hyperparameters tend to fall into one of these two categories.

Since hyperparameters are not determined by the data-fitting optimization procedure, how should we choose their values? A suitable answer to this question requires some discussion of the different types of error at play.

Types of Error

We have seen that it is common to minimize some measure of how poorly our hypothesis fits the data we have, but what we actually care about is how well the hypothesis predicts future data. Let us try to formally distinguish the various types of error. Assume that the data are distributed according to some (unknown) distribution $D$, and that we have a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, which is to measure the error between the true output $y$ and our estimate $\hat{y} = h(x)$. The risk (or true error) of a particular hypothesis $h \in \mathcal{H}$ is the expected loss over the whole data distribution:

$$R(h) = \mathbb{E}_{(x,y) \sim D}[\ell(h(x), y)]$$

Ideally, we would find the hypothesis that minimizes the risk, i.e.

$$h^* = \arg \min_{h \in \mathcal{H}} R(h)$$

However, computing this expectation is impossible because we do not have access to the true data distribution. Rather, we have access to samples $(x_i, y_i) \overset{iid}{\sim} D$. These enable us to approximate the real problem we care about by minimizing the empirical risk (or training error)

$$\hat{R}_\text{TRAIN}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(x_i), y_i)$$

But since we have a finite number of samples, the hypothesis that performs the best on the training data is not necessarily the best on the whole data distribution. In particular, if we both train and evaluate the hypothesis using the same data points, the training error will be a very biased estimate of the true error, since the hypothesis has been chosen specifically to perform well on those points. This phenomenon is sometimes referred to as “data incest”.

A common solution is to set aside some portion (say 30%) of the data, to be called the validation set, which is disjoint from the training set and not allowed to be used when fitting the model:
We can use this validation set to estimate the true error by the validation error

\[ \hat{R}_{\text{val}}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i^{\text{val}}), y_i^{\text{val}}) \]

With this estimate, we have a simple method for choosing hyperparameter values: try a bunch of configurations of the hyperparameters and choose the one that yields the lowest validation error.

**The effect of hyperparameters on error**

Note that as we add more features to a linear model, training error can only decrease. This is because the optimizer can set \( w_i = 0 \) if feature \( i \) cannot be used to reduce training error.

Adding more features tends to reduce true error as long as the additional features are useful predictors of the output. However, if we keep adding features, these begin to fit noise in the training data instead of the true signal, causing true error to actually increase. This phenomenon is known as **overfitting**.

The validation error tracks the true error reasonably well as long as the validation set is sufficiently large. The regularization hyperparameter \( \lambda \) has a somewhat different effect on training error. Observe that if \( \lambda = 0 \), we recover the exact OLS problem, which is directly minimizing the training error. As \( \lambda \) increases, the optimizer places less emphasis on the training error and more emphasis on reducing the magnitude of the parameters. This leads to a degradation in training error as \( \lambda \) grows:
Cross-validation

Setting aside a validation set works well, but comes at a cost, since we cannot use the validation data for training. Since having more data generally improves the quality of the trained model, we may prefer not to let that data go to waste, especially if we have little data to begin with and/or collecting more data is expensive. Cross-validation is an alternative to having a dedicated validation set.

$k$-fold cross-validation works as follows:

1. Shuffle the data and partition it into $k$ equally-sized (or as equal as possible) blocks.
2. For $i = 1, \ldots, k$,
   - Train the model on all the data except block $i$.
   - Evaluate the model (i.e. compute the validation error) using block $i$.

\[
\begin{array}{cccccccc}
1 & 2 & 3 & 4 & 5 & 6 & \cdots & k \\
\hline
\text{validate} & \text{train} \\
\text{train} & \text{validate} & \text{train} \\
\text{train} & \text{validate} & \text{train} \\
\vdots \\
\end{array}
\]

3. Average the $k$ validation errors; this is our final estimate of the true error.

Observe that, although every datapoint is used for evaluation at some time or another, the model is always evaluated on a different set of points than it was trained on, thereby cleverly avoiding the “data incest” problem mentioned earlier.

Note also that this process (except for the shuffling and partitioning) must be repeated for every hyperparameter configuration we wish to test. This is the principle drawback of $k$-fold cross-validation as compared to using a held-out validation set – there is roughly $k$ times as much computation required. This is not a big deal for the relatively small linear models that we’ve seen so far, but it can be prohibitively expensive when the model takes a long time to train, as is the case in the Big Data regime or when using neural networks.
Chapter 2

Regression II

2.1 MLE and MAP for Regression (Part I)

So far, we’ve explored two approaches of the regression framework, Ordinary Least Squares and Ridge Regression:

\[
\hat{w}_{\text{OLS}} = \arg\min \|y - Xw\|_2^2 \\
\hat{w}_{\text{RIDGE}} = \arg\min \|y - Xw\|_2^2 + \lambda \|w\|_2^2
\]

One question that arises is why we specifically use the \(\ell^2\) norm to measure the error of our predictions, and to penalize the model parameters. We will justify this design choice by exploring the statistical interpretations of regression — namely, we will employ Gaussians, MLE and MAP to validate what we’ve done so far through a different lens.

Probabilistic Model

In the context of supervised learning, we assume that there exists a true underlying model mapping inputs to outputs:

\[ f : \mathbb{R}^d \rightarrow f(\mathbb{R}) \]

The true model is unknown to us, and our goal is to find a hypothesis model that best represents the true model. The only information that we have about the true model is via a dataset

\[ D = \{(x_i, y_i)\}_{i=1}^n \]

where \(x_i \in \mathbb{R}^d\) is the input and \(y_i \in \mathbb{R}\) is the observation, a noisy version of the true output \(f(x_i)\):

\[ Y_i = f(x_i) + Z_i \]

We assume that \(x_i\) is a fixed value (which implies that \(f(x_i)\) is fixed as well), while \(Z_i\) is a random variable (which implies that \(Y_i\) is a random variable as well). We always assume that \(Z_i\) has zero mean, because otherwise there would be systematic bias in our observations. The \(Z_i\)'s could be Gaussian, uniform, Laplacian, etc... In most contexts, we us assume that they are independent identically distributed (i.i.d) Gaussians: \(Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)\). We can therefore say that \(Y_i\) is a random variable whose probability distribution is given by

\[ Y_i \overset{\text{iid}}{\sim} \mathcal{N}(f(x_i), \sigma^2) \]
Now that we have defined the model and data, we wish to find a hypothesis model $h_\theta$ (parameterized by $\theta$) that best captures the relationships in the data, while possibly taking into account prior beliefs that we have about the true model. We can represent this as a probability problem, where the goal is to find the optimal model that maximizes our probability.

**Maximum Likelihood Estimation**

In Maximum Likelihood Estimation (MLE), the goal is to find the hypothesis model that maximizes the probability of the data. If we parameterize the set of hypothesis models with $\theta$, we can express the problem as

$$\hat{\theta}_{\text{MLE}} = \arg \max_\theta \mathcal{L}(\theta; \mathcal{D}) = p(\text{data} = \mathcal{D} \mid \text{true model} = h_\theta)$$

The quantity $\mathcal{L}(\theta)$ that we are maximizing is also known as the **likelihood**, hence the term MLE. Substituting our representation of $\mathcal{D}$ we have

$$\hat{\theta}_{\text{MLE}} = \arg \max_\theta \mathcal{L}(\theta; X, y) = p(y_1, \ldots, y_n \mid x_1, \ldots, x_n, \theta)$$

Note that we implicitly condition on the $x_i$'s, because we treat them as fixed values of the data. The only randomness in our data comes from the $y_i$'s (since they are noisy versions of the true values $f(x_i)$). We can further simplify the problem by working with the log likelihood $\ell(\theta; X, y) = \log \mathcal{L}(\theta; X, y)$

$$\hat{\theta}_{\text{MLE}} = \arg \max_\theta \mathcal{L}(\theta; X, y) = \arg \max_\theta \ell(\theta; X, y)$$

With logs we are still working with the same problem, because logarithms are monotonic functions. In other words we have that:

$$P(A) < P(B) \iff \log P(A) < \log P(B)$$

Let’s decompose the log likelihood:

$$\ell(\theta; X, y) = \log p(y_1, \ldots, y_n \mid x_1, \ldots, x_n, \theta) = \log \prod_{i=1}^n p(y_i \mid x_i, \theta) = \sum_{i=1}^n \log [p(y_i \mid x_i, \theta)]$$

We decoupled the probabilities from each datapoints because their corresponding noise components are independent. Note that the logs allow us to work with sums rather products, simplifying the problem — one reason why the log likelihood is such a powerful tool. Each individual term $p(y_i \mid x_i, \theta)$ comes from a Gaussian

$Y_i \mid \theta \sim \mathcal{N}(h_\theta(x_i), \sigma^2)$

Continuing with logs:

$$\hat{\theta}_{\text{MLE}} = \arg \max_\theta \ell(\theta; X, y)$$

$$= \arg \max_\theta \sum_{i=1}^n \log [p(y_i \mid x_i, \theta)]$$

$$= \arg \max_\theta -\left(\sum_{i=1}^n \frac{(y_i - h_\theta(x_i))^2}{2\sigma^2}\right) - n \log \sqrt{2\pi\sigma}$$
2.1. MLE AND MAP FOR REGRESSION (PART I)

\[= \arg \min_{\theta} \left( \sum_{i=1}^{n} \frac{(y_i - h_\theta(x_i))^2}{2\sigma^2} \right) + n \log \sqrt{2\pi\sigma} \quad (2.4)\]

\[= \arg \min_{\theta} \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 \quad (2.5)\]

Note that in step (4) we turned the problem from a maximization problem to a minimization problem by negating the objective. In step (5) we eliminated the second term and the denominator in the first term, because they do not depend on the variables we are trying to optimize over.

Now let’s look at the case of regression — our hypothesis has the form \( h_\theta(x_i) = x_i^\top \theta \), where \( \theta \in \mathbb{R}^d \), where \( d \) is the number of dimensions of our featurized datapoints. For this specific setting, the problem becomes:

\[\hat{\theta}_{\text{MLE}} = \arg \min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - x_i^\top \theta)^2\]

This is just the Ordinary Least Squares (OLS) problem! We just proved that OLS and MLE for regression lead to the same answer! We conclude that MLE is a probabilistic justification for why using squared error (which is the basis of OLS) is a good metric for evaluating a regression model.

**Maximum a Posteriori**

In Maximum a Posteriori (MAP) Estimation, the goal is to find the model, for which the data maximizes the probability of the model:

\[\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\text{true model} = h_\theta \mid \text{data} = \mathcal{D})\]

The probability distribution that we are maximizing is known as the **posterior**. Maximizing this term directly is often infeasible, so we we use **Bayes’ Rule** to re-express the objective.

\[\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\text{true model} = h_\theta \mid \text{data} = \mathcal{D})\]

\[= \arg \max_{\theta} \frac{p(\text{data} = \mathcal{D} \mid \text{true model} = h_\theta) \cdot p(\text{true model} = h_\theta)}{p(\text{data} = \mathcal{D})}\]

\[= \arg \max_{\theta} p(\text{data} = \mathcal{D} \mid \text{true model} = h_\theta) \cdot p(\text{true model} = h_\theta)\]

\[= \arg \max_{\theta} \log p(\text{data} = \mathcal{D} \mid \text{true model} = h_\theta) + \log p(\text{true model} = h_\theta)\]

\[= \arg \min_{\theta} - \log p(\text{data} = \mathcal{D} \mid \text{true model} = h_\theta) - \log p(\text{true model} = h_\theta)\]

We treat \( p(\text{data} = \mathcal{D}) \) as a constant value because it does not depend on the variables we are optimizing over. Notice that MAP is just like MLE, except we add a term \( p(\text{true model} = h_\theta) \) to our objective. This term is the **prior** over our true model. Adding the prior has the effect of favoring certain models over others **a priori**, regardless of the dataset. Note the MLE is a special case of MAP, when the prior does not treat any model more favorably over other models. Concretely, we have that

\[\hat{\theta}_{\text{MAP}} = \arg \min_{\theta} - \left( \sum_{i=1}^{n} \log[p(y_i \mid x_i, \theta)] \right) - \log[p(\theta)]\]
Again, just as in MLE, notice that we implicitly condition on the $x_i$'s because we treat them as constants. Also, let us assume as before that the noise terms are i.i.d. Gaussians: $N_i \sim N(0, \sigma^2)$. For the prior term $P(\Theta)$, we assume that the components $\theta_j$ are i.i.d. Gaussians:

$$\theta_j \sim N(\theta_j^0, \sigma^2_h)$$

Using this specific information, we now have:

$$\hat{\theta}_{\text{MAP}} = \arg \min_{\theta} \left( \frac{\sum_{i=1}^{n} (y_i - h_\theta(x_i))^2}{2\sigma^2} + \frac{\sum_{j=1}^{d} (\theta_j - \theta_j^0)^2}{2\sigma^2_{h_j}} \right)$$

Let’s look again at the case for linear regression to illustrate the effect of the prior term when $\theta_j^0 = 0$. In this context, we refer to the linear hypothesis function $h_\theta(x) = \theta^\top x$.

$$\hat{\theta}_{\text{MAP}} = \arg \min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - x_i^\top \theta)^2 + \sigma^2 \sum_{j=1}^{d} \theta_j^2$$

This is just the Ridge Regression problem! We just proved that Ridge Regression and MAP for regression lead to the same answer! We can simply set $\lambda = \frac{\sigma^2}{\sigma^2_h}$. We conclude that MAP is a probabilistic justification for adding the penalized ridge term in Ridge Regression.

**MLE vs. MAP**

Based on our analysis of Ordinary Least Squares Regression and Ridge Regression, we should expect to see MAP perform better than MLE. But is that always the case? Let us visit a simple 2D problem where

$$f(x) = \text{slope} \cdot x + \text{intercept}$$

Suppose we already know the true underlying model parameters:

$$(\text{slope}^*, \text{intercept}^*) = (0.5, 1.0)$$

we would like to know what parameters MLE and MAP will select, after providing them with some dataset $\mathcal{D}$. Let’s start with MLE:
2.1. MLE AND MAP FOR REGRESSION (PART I)

The diagram above shows the contours of the likelihood distribution in model space. The gray dot represents the true underlying model. MLE chooses the point that maximizes the likelihood, which is indicated by the green dot. As we can see, MLE chooses a reasonable hypothesis, but this hypothesis lies in a region on high variance, which indicates a high level of uncertainty in the predicted model. A slightly different dataset could significantly alter the predicted model.

Now, let’s take a look at the hypothesis model from MAP. One question that arises is where the prior should be centered and what its variance should be. This depends on our belief of what the true underlying model is. If we have reason to believe that the model weights should all be small, then the prior should be centered at zero with a small variance. Let’s look at MAP for a prior that is centered at zero:

For reference, we have marked the MLE estimation from before as the green point and the true model as the gray point. The prior distribution is indicated by the diagram on the left, and
the posterior distribution is indicated by the diagram on the right. MAP chooses the point that maximizes the posterior probability, which is approximately \((0.70, 0.25)\). Using a prior centered at zero leads us to skew our prediction of the model weights toward the origin, leading to a less accurate hypothesis than MLE. However, the posterior has significantly less variance, meaning that the point that MAP chooses is less likely to overfit to the noise in the dataset.

Let’s say in our case that we have reason to believe that both model weights should be centered around the 0.5 to 1 range.

Our prediction is now close to that of MLE, with the added benefit that there is significantly less variance. However, if we believe the model weights should be centered around the -0.5 to -1 range, we would make a much poorer prediction than MLE.

As always, in order to compare our beliefs to see which prior works best in practice, we should use cross validation!
2.2 Bias-Variance Tradeoff

Recall from our previous discussion on supervised learning, that for a fixed input $x$ the corresponding measurement $Y$ is a noisy measurement of the true underlying response $f(x)$:

$$Y = f(x) + Z$$

Where $Z$ is a zero-mean random variable, and is typically represented as a Gaussian distribution. Our goal in regression is to recover the underlying model $f(.)$ as closely as possible. We previously mentioned MLE and MAP as two techniques that try to find of reasonable approximation to $f(.)$ by solving a probabilistic objective. We briefly compared the effectiveness of MLE and MAP, and noted that the effectiveness of MAP is in large part dependent on the prior over the parameters we optimize over. One question that naturally arises is: how exactly can we measure the effectiveness of a hypothesis model? In this section, we would like to form a theoretical metric that can exactly measure the effectiveness of a hypothesis function $h$. Keep in mind that this is only a theoretical metric that cannot be measured in real life, but it can be approximated via empirical experiments — more on this later.

Before we introduce the metric, let’s make a few subtle statements about the data and hypothesis. As you may recall from our previous discussion on MLE and MAP, we had a dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$$

In that context, we treated the $x_i$’s in our dataset $\mathcal{D}$ as fixed values. In this case however, we treat the $x_i$’s as values sampled from random variables $X_i$. That is, $\mathcal{D}$ is a random variable, consisting of random variables $X_i$ and $Y_i$. For some arbitrary test input $x$, $h(x; \mathcal{D})$ depends on the random variable $\mathcal{D}$ that was used to train $h$. Since $\mathcal{D}$ is random, we will have a slightly different hypothesis model $h(x; \mathcal{D})$ every time we use a new dataset. Note that $x$ and $\mathcal{D}$ are completely independent from one another — $x$ is a test point, while $\mathcal{D}$ consists of the training data.

Metric

Our objective is to, for a fixed test point $x$, evaluate how closely the hypothesis can estimate the noisy observation $Y$ corresponding to $x$. Note that we have denoted $x$ here as a lowercase letter because we are treating it as a fixed constant, while we have denoted the $Y$ and $\mathcal{D}$ as uppercase letters because we are treating them as random variables. $Y$ and $\mathcal{D}$ as independent random variables, because our $x$ and $Y$ have no relation to the set of $X_i$’s and $Y_i$’s in $\mathcal{D}$. Again, we can view $\mathcal{D}$ as the training data, and $(x, Y)$ as a test point — the test point $x$ is probably not even in the training set $\mathcal{D}$! Mathematically, we express our metric as the expected squared error between the hypothesis and the observation $Y = f(x) + Z$:

$$\varepsilon(x; h) = \mathbb{E}[(h(x; \mathcal{D}) - Y)^2]$$

The expectation here is over two random variables, $\mathcal{D}$ and $Y$:

$$\mathbb{E}_{\mathcal{D}, Y}[(h(x; \mathcal{D}) - Y)^2] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_Y((h(x; \mathcal{D}) - Y)^2 | \mathcal{D})]$$

Note that the error is w.r.t the observation $Y$ and not the true underlying model $f(x)$, because we do not know the true model and only have access to the noisy observations from the true model.
Bias-Variance Decomposition

The error metric is difficult to interpret and work with, so let’s try to decompose it into parts that are easier to understand. Before we start, let’s find the expectation and variance of $Y$:

$$
E[Y] = E[f(x) + Z] = f(x) + E[Z] = f(x)
$$

$$
\text{Var}(Y) = \text{Var}(f(x) + Z) = \text{Var}(Z)
$$

Also, in general for any random variable $X$, we have that

$$
\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2 \implies E[X^2] = \text{Var}(X) + E[X]^2
$$

Let’s use these facts to decompose the error:

$$
\varepsilon(x; h) = E[(h(x; D) - Y)^2] = E[h(x; D)^2] + E[Y^2] - 2E[h(x; D) \cdot Y]
$$

$$
= \left( \text{Var}(h(x; D)) + E[h(x; D)]^2 \right) + \left( \text{Var}(Y) + E[Y]^2 \right) - 2E[h(x; D)] \cdot E[Y]
$$

$$
= \left( E[h(x; D)]^2 - 2E[h(x; D)] \cdot E[Y] + E[Y]^2 \right) + \text{Var}(h(x; D)) + \text{Var}(Y)
$$

$$
= \left( E[h(x; D)] - E[Y] \right)^2 + \text{Var}(h(x; D)) + \text{Var}(Y)
$$

Recall that for any two independent random variables $D$ and $Y$, $g_1(D)$ and $g_2(Y)$ are also independent, for any functions $g_1, g_2$. This implies that $h(x; D)$ and $Y$ are independent, allowing us to express $E[h(x; D) \cdot Y] = E[h(x; D)] \cdot E[Y]$ in the second line of the derivation. The final decomposition, also known as the **bias-variance decomposition**, consists of three terms:

- **Bias$^2$ of method**: Measures how well the *average* hypothesis (over all possible training sets) can come close to the true underlying value $f(x)$, for a fixed value of $x$. A low bias means that on average the regressor $h(x)$ accurately estimates $f(x)$.

- **Variance of method**: Measures the variance of the hypothesis (over all possible training sets), for a fixed value of $x$. A low variance means that the prediction does not change much as the training set varies. An un-biased method (bias = 0) could have a large variance.

- **Irreducible error**: This is the error in our model that we cannot control or eliminate, because it is due to errors inherent in our noisy observation $Y$.

The decomposition allows us to measure the error in terms of bias, variance, and irreducible error. Irreducible error has no relation with the hypothesis model, so we can fully ignore it in theory when minimizing the error. As we have discussed before, models that are very complex have very little bias because on *average* they can fit the true underlying model value $f(x)$ very well, but have very high variance and are very far off from $f(x)$ on an individual basis.

Note that the error above is only for a fixed input $x$, but in regression our goal is to minimize the average error over all possible values of $X$. If we know the distribution for $X$, we can find the effectiveness of a hypothesis model as a whole by taking an expectation of the error over all possible values of $x$: $E_X[\varepsilon(x; h)]$.
2.2. BIAS-VARIANCE TRADEOFF

Alternative Decomposition

The previous derivation is short, but may seem somewhat arbitrary. Let’s explore an alternative derivation. At its core, it uses the technique that $E[(Z - Y)^2] = E[(Z - E[Z]) + (E[Z] - Y)]^2$ which decomposes to easily give us the variance of $Z$ and other terms.

$$
\varepsilon(x; h) = E[(h(x; D) - Y)^2]
$$
$$
= E\left[ (h(x; D) - E[h(x; D)])^2 + E[h(x; D)] - Y \right]^2
$$
$$
= E\left[ (h(x; D) - E[h(x; D)])^2 \right] + 2E\left[ (h(x; D) - E[h(x; D)]) \cdot (E[h(x; D)] - Y) \right] + E\left[ (E[h(x; D)] - Y)^2 \right]
$$
$$
= \text{Var}(h(x; D)) + \left( \text{Var}(E[h(x; D)]) + \text{Var}(Y) \right)
$$
$$
= \text{Var}(h(x; D)) + \text{Var}(E[h(x; D)]) - \text{Var}(Y) + \text{Var}(Y)
$$
$$
= \text{Var}(h(x; D)) + \text{Var}(E[h(x; D)]) - \text{Var}(Y) + \text{Var}(Y)
$$

Experiments

Let’s confirm the theory behind the bias-variance decomposition with an empirical experiment that measures the bias and variance for polynomial regression with 0 degree, 1st degree, and 2nd degree polynomials. In our experiment, we will repeatedly fit our hypothesis model to a random training set. We then find the expectation and variance of the fitted models generated from these training sets.

Let’s first look at a 0 degree (constant) regression model. We repeatedly fit an optimal constant line to a training set of 10 points. The true model is denoted by gray and the hypothesis is denoted by red. Notice that at each time the red line is slightly different due to the different training set used.
Let’s combine all of these hypotheses together into one picture to see the bias and variance of our model.

On the top left diagram we see all of our hypotheses and all training sets used. The bottom left diagram shows the average hypothesis in cyan. As we can see, this model has low bias for x’s in
the center of the graph, but very high bias for \( x \)'s that are away from the center of the graph. The diagram in the bottom right shows that the variance of the hypotheses is quite high, for all values of \( x \).

Now let’s look at a 1st degree (linear) regression model.

The bias is now very low bias for all \( x \)'s. The variance is low for \( x \)'s in the middle of the graph,
but higher for $x$’s that are away from the center of the graph. Finally, let’s look at a 2nd degree (quadratic) regression model.

Fitting A Model over Multiple Datasets: $p = 2$

Bias and Variance in Model Selection: $p = 2$

The bias is still very low for all $x$’s. However, the variance is much higher for all values of $x$. Let’s summarize our results. We find the bias and the variance empirically and graph them for all values of $x$, as shown in the first two graphs. Finally, we take an expectation over the bias and
2.2. **BIAS-VARIANCE TRADEOFF**

variance over all values of $x$, as shown in the third graph.

The bias-variance decomposition confirms our understanding that the true model is linear. While a quadratic model achieves the same theoretical bias as a linear model, it overfits to the data, as indicated by its high variance. On the other hand a constant model underfits the data, as indicated by its high bias. In the process of training our model, we can tell that a constant model is a poor choice, because its high bias is reflected in poor training error. However we cannot tell that a quadratic model is poor, because its high variance is not reflected in the training error. This is the reason why we use validation data and cross-validation as a means to measure the performance of our hypothesis model on unseen data.

**Takeaways**

Let us conclude by stating some implications of the Bias-Variance Decomposition:
1. Underfitting is equivalent to high bias; most overfitting correlates to high variance.

2. Training error reflects bias but not variance. Test error reflects both. In practice, if the training error is much smaller than the test error, then there is overfitting.

3. Variance $\to 0$ as $n \to \infty$.

4. If $f$ is in the set of hypothesis functions, bias will decrease with more data. If $f$ is not in the set of hypothesis functions, then there is an underfitting problem and more data won’t help.

5. Adding good features will decrease the bias, but adding a bad feature rarely increase the bias. (just set the coefficient of the feature to 0)

6. Adding a feature usually increase the variance, so a feature should only be added if it decreases bias more than it increases variance.

7. Irreducible error can not be reduced.

8. Noise in the test set only affects $\text{Var}(Z)$, but noise in the training set also affects bias and variance.

9. For real-world data, $f$ is rarely known, and the noise model might be wrong, so we can’t calculate bias and variance. But we can test algorithms over synthetic data.

### 2.3 Multivariate Gaussians

So far in our discussion of MLE and MAP in regression, we considered a set of Gaussian random variables $Z_1, Z_2, \ldots, Z_k$, which can represent anything from the noise in data to the parameters of a model. One critical assumption we made is that these variables are independent and identically distributed. However, what about the case when these variables are dependent and/or non-identical? For example, in time series data we have the relationship

$$Z_{i+1} = rZ_i + U_i$$

where $U_i \sim \mathcal{N}(0, 1)$ and $-1 \leq r \leq 1$ (so that it doesn’t blow up)

Here’s another example: consider the “sliding window” (like echo of audio)

$$Z_i = \Sigma r_j U_{i-j}$$

where $U_i \sim \mathcal{N}(0, 1)$

In general, if we can represent the random vector $Z = (Z_1, Z_2, \ldots, Z_k)$ as

$$Z = RU$$

where $Z \in \mathbb{R}^n$, $R \in \mathbb{R}^{n \times n}$, $U \in \mathbb{R}^n$, and $U_i \sim \mathcal{N}(0, 1)$, we refer to $Z$ as a **Jointly Gaussian Random Vector**. Our goal now is to derive its probability density formula.

**Definition**

There are three equivalent definitions of a jointly Gaussian (JG) random vector:

1. A random vector $Z = (Z_1, Z_2, \ldots, Z_k)$ is JG if there exists a base random vector $U = (U_1, U_2, \ldots, U_l)$ whose components are independent standard normal random variables, a transition matrix $R \in \mathbb{R}^{k \times l}$, and a mean vector $\mu \in \mathbb{R}^k$, such that $Z = RU + \mu$. 

2.3. MULTIVARIATE GAUSSIANS

2. A random vector $Z = (Z_1, Z_2, \ldots, Z_k)^\top$ is JG if $\sum_{i=1}^{k} a_i Z_i$ is normally distributed for every $a = (a_1, a_2, \ldots, a_k)^\top \in \mathbb{R}^k$.

3. (Non-degenerate case only) A random vector $Z = (Z_1, Z_2, \ldots, Z_k)^\top$ is JG if

$$f_Z(z) = \frac{1}{\sqrt{|\text{det}(\Sigma)|}} \frac{1}{(\sqrt{2\pi})^k} e^{-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu)}$$

Where $\Sigma = \mathbb{E}[(Z - \mu)(Z - \mu)^\top] = \mathbb{E}[(RU)(RU)^\top] = \mathbb{E}R(UU^\top)R^\top = RR^\top$

$\Sigma$ is also called the covariance matrix of $Z$.

Note that all of these conditions are equivalent. In this note we will start by showing a proof that $(1) \implies (3)$. We will leave it as an exercise to prove the rest of the implications needed to show that the three conditions are in fact equivalent.

**Proving $(1) \implies (3)$**

In the context of the noise problem we defined earlier, we are starting with condition $(1)$, i.e. $Z = RU$ (in this case $k = l = n$), and we would like to derive the probability density of $Z$. Note that here we removed the $\mu$ from consideration because in machine learning we always assume that the noise has a mean of 0. We leave it as an exercise for the reader to prove the case for an arbitrary $\mu$.

We will first start by relating the probability density function of $U$ to that of $Z$. Denote $f_U(u)$ as the probability density for $U = u$, and similarly denote $f_Z(z)$ as the probability density for $Z = z$.

One may initially believe that $f_U(u) = f_Z(Ru)$, but this is NOT true. Remember that since there is a change of variables from $U$ to $Z$, we must make sure to incorporate the change of variables constant, which in this case is the absolute value of the determinant of $R$. Incorporating this constant, we will have the correct formula:

$$f_U(u) = |\text{det}(R)|f_Z(Ru)$$

Let’s see why this is true, with a simple 2D geometric explanation. Define $U$ space to be the 2D space with axes $U_1$ and $U_2$. Now take any arbitrary region $R'$ in $U$ space (note that this $R'$ is different from the matrix $R$ that relates $U$ to $Z$). As shown in the diagram below, we have some off-centered circular region $R'$ and we would like to approximate the probability that $U$ takes a value in this region. We can do so by taking a Riemann sum of the density function $f_U(.)$ over smaller and smaller squares that make up the region $R'$:
Mathematically, we have that

$$P(U \subseteq R') = \int \int_{R'} f_U(u_1, u_2) du_1 du_2 \approx \sum \sum_{R'} f_U(u_1, u_2) \Delta u_1 \Delta u_2$$

Now, let’s apply the linear transformation $Z = RU$, mapping the region $R'$ in $U$ space, to the region $T(R')$ in $Z$ space.

The graph on the right is now $Z$ space, the 2D space with axes $Z_1$ and $Z_2$. Assuming that the matrix $R$ is invertible, there is a one-to-one correspondence between points in $U$ space to points in $Z$ space. As we can note in the diagram above, each unit square in $U$ space maps to a parallelogram in $Z$ space (in higher dimensions, we would use the terms hypercube and parallelepiped). Recall the relationship between each unit hypercube and the parallelepiped it maps to:

$$\text{Area(parallelepiped)} = |\det(R)| \cdot \text{Area(hypercube)}$$

In this 2D example, if we denote the area of each unit square as $\Delta u_1 \Delta u_2$, and the area of each unit parallelepiped as $\Delta A$, we say that

$$\Delta A = |\det(R)| \cdot \Delta u_1 \Delta u_2$$

Now let’s take a Riemann sum to find the probability that $Z$ takes a value in $T(R')$:

$$P(Z \subseteq T(R')) = \int \int_{T(R')} f_Z(z_1, z_2) dz_1 dz_2
\approx \sum \sum_{T(R')} f_Z(z) \Delta A
= \sum \sum_{R'} f_Z(Ru) |\det(R)| \Delta u_1 \Delta u_2$$

Note the change of variables in the last step: we sum over the squares in $U$ space, instead of parallelograms in $R$ space.

So far, we have shown that (for any dimension $n$)

$$P(U \subseteq R') = \int \ldots \int_{R'} f_U(u) du_1 du_2 \ldots du_n$$
and
\[ P(Z \subseteq T(R')) = \int \ldots \int_{R'} f_Z(Ru) \, |\det(R)| \, du_1 \, du_2 \ldots \, du_n \]

Notice that these two probabilities are equivalent! The probability that \( U \) takes value in \( R' \) must equal the probability that the transformed random vector \( Z \) takes a value in the transformed region \( T(R') \).

Therefore, we can say that
\[ P(U \subseteq R') = \int \ldots \int_{R'} f_U(u) \, du_1 \, du_2 \ldots \, du_n \]
\[ = \int \ldots \int_{R'} f_Z(Ru) \, |\det(R)| \, du_1 \, du_2 \ldots \, du_n \]
\[ = P(Z \subseteq T(R')) \]

We conclude that
\[ f_U(u) = f_Z(Ru) \, |\det(R)| \]
An almost identical argument will allow us to state that
\[ f_Z(z) = f_U(R^{-1}z) \, |\det(R^{-1})| = \frac{1}{|\det(R)|} f_U(R^{-1}z) \]

Since the densities for all the \( U_i \)'s are i.i.d, and \( U = R^{-1}Z \), we can write the joint density function of \( Z \) as
\[ f_Z(z) = \frac{1}{|\det(R)|} f_U(R^{-1}z) \]
\[ = \frac{1}{|\det(R)|} \prod_{i=1}^{n} f_U((R^{-1}z)_i) \]
\[ = \frac{1}{|\det(R)|} \left( \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}(R^{-1}z)\top(R^{-1}z)} \right) \]
\[ = \frac{1}{|\det(R)|} \left( \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}z\top R^{-\top}R^{-1}z} \right) \]
\[ = \frac{1}{|\det(R)|} \left( \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}z\top (RR^\top)^{-1}z} \right) \]

Note that \((RR^\top)^{-1}\) is simply the covariance matrix for \( Z \):
\[ \text{Cov}[Z] = E[ZZ^\top] = E[RUU^\top R^\top] = RE[UU^\top]R^\top = RIR^\top = RR^\top \]

Thus the density function of \( Z \) can be written as
\[ f_Z(z) = \frac{1}{|\det(R)|} \left( \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}z\top \Sigma Z^{-1}z} \right) \]

Furthermore, we know that
\[ |\det(\Sigma_Z)| = |\det(\Sigma R^\top)| \]
\[ = |\det(R) \cdot \det(R^\top)| \]
\[ \det(\mathbf{R}) \cdot \det(\mathbf{R}) = \det(\mathbf{R})^2 \]

and therefore

\[ f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{\sqrt{\det(\Sigma_Z)}} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \mathbf{z}^T \Sigma_Z^{-1} \mathbf{z}} \]

**Estimating Gaussians from Data**

For a particular multivariate Gaussian distribution \( f(\cdot) \), if we do not have the true means and covariances \( \mu, \Sigma \), then our best bet is to use MLE to estimate them empirically with i.i.d. samples \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \):

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \\
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T
\]

Note that the above formulas are not necessarily trivial and must be formally proven using MLE. Just to present a glimpse of the process, let’s prove that these formulas hold for the case where we are dealing with 1-d data points. For notation purposes, assume that \( \mathcal{D} = \{x_1, x_2, \ldots, x_n\} \) is the set of all training data points that belong to class \( k \). Note that the data points are i.i.d. Our goal is to solve the following MLE problem:

\[
\hat{\mu}, \hat{\sigma}^2 = \arg \max_{\mu, \sigma^2} P(x_1, x_2, \ldots, x_n \mid \mu, \sigma^2) \\
= \arg \max_{\mu, \sigma^2} \ln P(x_1, x_2, \ldots, x_n \mid \mu, \sigma^2) \\
= \arg \max_{\mu, \sigma^2} \sum_{i=1}^{n} \ln P(x_i \mid \mu, \sigma^2) \\
= \arg \max_{\mu, \sigma^2} \frac{-\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} - \ln(\sigma) - \frac{1}{2} \ln(2\pi) \\
= \arg \min_{\mu, \sigma^2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} + \ln(\sigma)
\]

Note that the objective above is not jointly convex, so we cannot simply take derivatives and set them to 0! Instead, we decompose the minimization over \( \sigma^2 \) and \( \mu \) into a nested optimization problem:

\[
\min_{\mu, \sigma^2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} + \ln(\sigma) = \min_{\sigma^2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} + \ln(\sigma)
\]

The optimization problem has been decomposed into an inner problem that optimizes for \( \mu \) given a fixed \( \sigma^2 \), and an outer problem that optimizes for \( \sigma^2 \) given the optimal value \( \hat{\mu} \). Let’s first solve the inner optimization problem. Given a fixed \( \sigma^2 \), the objective is convex in \( \mu \), so we can simply take a partial derivative w.r.t \( \mu \) and set it equal to 0:

\[
\frac{\partial}{\partial \mu} \left( \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} + \ln(\sigma) \right) = \sum_{i=1}^{n} \frac{-(x_i - \mu)}{\sigma^2} = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]
Having solved the inner optimization problem, we now have that

$$\min_{\sigma^2} \min_{\mu} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} + \ln(\sigma) = \min_{\sigma^2} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{2\sigma^2} + \ln(\sigma)$$

Note that this objective is not convex in $\sigma$, so we must instead find the critical point of the objective that minimizes the objective. Assuming that $\sigma \geq 0$, the critical points are:

- $\sigma = 0$: assuming that not all of the points $x_i$ are equal to $\hat{\mu}$, there are two terms that are at odds with each other: a $1/\sigma^2$ term that blows off to $\infty$, and a $\ln(\sigma)$ term that blows off to $-\infty$ as $\sigma \to 0$. Note that the $1/\sigma^2$ term blows off at a faster rate, so we conclude that
  $$\lim_{\sigma \to 0} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{2\sigma^2} + \ln(\sigma) = \infty$$

- $\sigma = \infty$: this case does not lead to the solution, because it gives a maximum, not a minimum.
  $$\lim_{\sigma \to \infty} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{2\sigma^2} + \ln(\sigma) = \infty$$

- Points at which the derivative w.r.t $\sigma$ is 0
  $$\frac{\partial}{\partial \sigma} \left( \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{2\sigma^2} + \ln(\sigma) \right) = \sum_{i=1}^{n} \frac{-(x_i - \hat{\mu})^2}{\sigma^3} + \frac{1}{\sigma} = 0 \implies \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

We conclude that the optimal point is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

**Isocontours**

Let’s try to understand in detail how to visualize a multivariate Gaussian distribution. For simplicity, let’s consider a zero-mean Gaussian distribution $\mathcal{N}(0, \Sigma)$, which just leaves us with the covariance matrix $\Sigma$. Since $\Sigma$ is a symmetric, positive semidefinite matrix, we can decompose it by the spectral theorem into $\Sigma = V \Lambda V^T$, where the columns of $V$ form an orthonormal basis in $\mathbb{R}^d$, and $\Lambda$ is a diagonal matrix with real, non-negative values. We wish to find its level set

$$f(x) = k$$

or simply the set of all points $x$ such that the probability density $f(x)$ evaluates to a fixed constant $k$. This is equivalent to the level set $\ln(f(x)) = \ln(k)$ which further reduces to

$$x^T \Sigma^{-1} x = c$$

for some constant $c$. Without loss of generality, assume that this constant is 1. The level set $x^T \Sigma^{-1} x = 1$ is an ellipsoid with axes $v_1, v_2, \ldots, v_d$, with lengths $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_d}$, respectively. Each axis of the ellipsoid is the vector $\sqrt{\lambda_i} v_i$, and we can verify that

$$(\sqrt{\lambda_i} v_i)^T \Sigma^{-1} (\sqrt{\lambda_i} v_i) = \lambda_i v_i^T \Sigma^{-1} v_i = \lambda_i v_i^T (\Sigma^{-1} v_i) = \lambda_i v_i^T (\lambda_i^{-1} v_i) = v_i^T v_i = 1$$
The entries of $\mathbf{\Lambda}$ dictate how elongated or shrunk the distribution is along each direction. In the case of isotropic distributions, the entries of $\mathbf{\Lambda}$ are all identical, meaning the axes of the ellipsoid form a circle. In the case of anisotropic distributions, the entries of $\mathbf{\Lambda}$ are not necessarily identical, meaning that the resulting ellipsoid may be elongated/shrunken and also rotated.

Figure 2.1: Isotropic (left) vs Anisotropic (right) contours are ellipsoids with axes $\sqrt{\lambda_i}v_i$. Images courtesy Professor Shewchuk’s notes

**Properties**

Let’s state some well-known properties of Multivariate Gaussians. Given a JG random vector $\mathbf{Z} \sim \mathcal{N}(\mu_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$, the linear transformation $\mathbf{AZ}$ (where $\mathbf{A}$ is an appropriately dimensioned constant matrix) is also JG:

$$\mathbf{AZ} \sim \mathcal{N}(\mathbf{A}\mu_{\mathbf{Z}}, \mathbf{A}\Sigma_{\mathbf{Z}}\mathbf{A}^\top)$$

We can derive the mean and covariance of $\mathbf{AZ}$ using the linearity of expectations:

$$\mu_{\mathbf{AZ}} = \mathbb{E}[\mathbf{AZ}] = \mathbf{A}\mathbb{E}[\mathbf{Z}] = \mathbf{A}\mu_{\mathbf{Z}}$$

and

$$\Sigma_{\mathbf{AZ}} = \mathbb{E}[(\mathbf{AZ} - \mathbb{E}[\mathbf{AZ}])(\mathbf{AZ} - \mathbb{E}[\mathbf{AZ}])^\top] = \mathbf{A}\mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top]\mathbf{A}^\top = \mathbf{A}\Sigma_{\mathbf{Z}}\mathbf{A}^\top$$

Note that the statements above did not rely on the fact that $\mathbf{Z}$ is JG, so this reasoning applies to all random vectors. We know that $\mathbf{AZ}$ is JG itself, because it can be expressed as a linear transformation of i.i.d. Gaussians: $\mathbf{AZ} = \mathbf{ARU}$.

Now suppose that we have the partition $\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$ whose distribution is given by $\mathbf{Z} \sim \mathcal{N}(\mu_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and

$$\mu_{\mathbf{Z}} = \begin{bmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{bmatrix}, \Sigma_{\mathbf{Z}} = \begin{bmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{bmatrix}$$

It turns out that the **marginal distribution** of the individual random vector $\mathbf{X}$ (and $\mathbf{Y}$) is JG:

$$\mathbf{X} \sim \mathcal{N}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{XX}})$$
However, the converse is not necessarily true: if $\mathbf{X}$ and $\mathbf{Y}$ are each individually JG, it is not necessarily the case that $\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$ is JG! To see why, let’s suppose that $\mathbf{X}$ and $\mathbf{Y}$ are individually JG. Thus, we can express each as a linear transformation of i.i.d. Gaussian random variables:

$$
\mathbf{X} = \mathbf{R}_X \mathbf{U}_X, \quad \mathbf{Y} = \mathbf{R}_Y \mathbf{U}_Y
$$

we would expect that the expression for the joint distribution would be JG:

$$
\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_X & 0 \\ 0 & \mathbf{R}_Y \end{bmatrix} \begin{bmatrix} \mathbf{U}_X \\ \mathbf{U}_Y \end{bmatrix}
$$

However, since we cannot guarantee that the entries of $\mathbf{U}_X$ are independently distributed from the entries of $\mathbf{U}_Y$, we cannot conclude that the joint distribution is JG. If the entries are independently distributed, then we would be able to conclude that the joint distribution is JG.

Let’s now transition back to our discussion of $\mathbf{Z}$. The conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ (and vice versa) is also JG:

$$
\mathbf{X} | \mathbf{Y} \sim \mathcal{N}(\mu_X + \Sigma_{XY} \Sigma_{YY}^{-1} (\mathbf{Y} - \mu_Y), \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX})
$$

If $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated (that is, if $\Sigma_{XY} = \Sigma_{YX} = 0$), we can say that they are independent. Namely, the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ does not depend on $\mathbf{Y}$:

$$
\mathbf{X} | \mathbf{Y} \sim \mathcal{N}(\mu_X + 0 \Sigma_{YY}^{-1} (\mathbf{Y} - \mu_Y), \Sigma_{XX} - 0 \Sigma_{YY}^{-1} 0) = \mathcal{N}(\mu_X, \Sigma_{XX})
$$

This also follows from the multivariate Gaussian pdf:

$$
f_Z(x, y) = \frac{1}{(\sqrt{2\pi})^{n_z}} \begin{bmatrix} \Sigma_{XX} & 0 \\ 0 & \Sigma_{YY} \end{bmatrix}^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix}^{\top} \begin{bmatrix} \Sigma_{XX} & 0 \\ 0 & \Sigma_{YY} \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \right)
$$

$$
= \frac{1}{(\sqrt{2\pi})^{n_x}} |\Sigma_{XX}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} x^{\top} \Sigma_{XX}^{-1} x \right) \cdot \frac{1}{(\sqrt{2\pi})^{n_y}} |\Sigma_{YY}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} y^{\top} \Sigma_{YY}^{-1} y \right)
$$

$$
= f_X(x) \cdot f_Y(y)
$$

Note the significance of this statement. Given any two general random vectors, we cannot necessarily say “if they are uncorrelated, then they are independent”. However in the case of random vectors from the same JG joint distribution, we can make this claim.

### 2.4 MLE and MAP for Regression (Part II)

The power of probabilistic thinking is that it allows us a way to model situations that arise and adapt our approaches in a reasonably principled way. This is particularly true when it comes to incorporating information about the situation that comes from the physical context of the data gathering process. In this note, we will explore what happens as we vary our assumptions about the noise in our data and the priors for our parameters, as well as the “importance” of certain training points.

So far we have used MLE and MAP to justify the optimization formulation of OLS and ridge regression, respectively. The MLE formulation assumes that the observation $Y_i$ is a noisy version of the true underlying output:

$$
Y_i = f(x_i) + Z_i
$$
where the noise for each datapoint is crucially i.i.d. The MAP formulation assumes that the model parameter $W_j$ is according to an i.i.d. Gaussian prior

$$W_j \sim \mathcal{N}(\mu_j, \sigma_h^2)$$

So far, we have restricted ourselves to the case when the noise/parameters are i.i.d:

$$Z \sim \mathcal{N}(0, \sigma^2 I), \quad W \sim \mathcal{N}(\mu_W, \sigma^2 h I)$$

However, what about the case when $N_i$'s/$W_j$'s are non-identical or dependent on one another? We would like to explore the case when the observation noise and underlying parameters are jointly Gaussian with arbitrary individual covariance matrices, but are independent of each other.

$$Z \sim \mathcal{N}(0, \Sigma_Z), \quad W \sim \mathcal{N}(\mu_W, \Sigma_W)$$

It turns out that via a change of coordinates, we can reduce these non-i.i.d. problems back to the i.i.d. case and solve them using the original techniques we used to solve OLS and Ridge Regression! Changing coordinates is a powerful tool in thinking about machine learning.

**Weighted Least Squares**

The basic idea of **weighted least squares** is the following: we place more emphasis on the loss contributed from certain data points over others - that is, we care more about fitting some data points over others. It turns out that this weighted perspective is very useful as a building block when we go beyond traditional least-squares problems.

**Optimization View**

From an optimization perspective, the problem can be expressed as

$$\hat{w}_{WLS} = \arg \min_{w \in \mathbb{R}^d} \left( \sum_{i=1}^{n} \omega_i (y_i - x_i^T w)^2 \right)$$

This objective is the same as OLS, except that each term in the sum is weighted by a positive coefficient $\omega_i$. As always, we can vectorize this problem:

$$\hat{w}_{WLS} = \arg \min_{w \in \mathbb{R}^d} (y - Xw)^T \Omega (y - Xw)$$

Where the $i$'th row $X$ is $x_i^T$, and $\Omega \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $\Omega_{i,i} = \omega_i$.

We rewrite the WLS objective to an OLS objective:

$$\hat{w}_{WLS} = \arg \min_{w \in \mathbb{R}^d} (y - Xw)^T \Omega (y - Xw)$$

$$= \arg \min_{w \in \mathbb{R}^d} (y - Xw)^T (\Omega^{1/2}\Omega^{1/2}(y - Xw))$$

$$= \arg \min_{w \in \mathbb{R}^d} (\Omega^{1/2}y - \Omega^{1/2}Xw)^T (\Omega^{1/2}y - \Omega^{1/2}Xw)$$

$$= \arg \min_{w \in \mathbb{R}^d} \| \Omega^{1/2}y - \Omega^{1/2}Xw \|^2$$
This formulation is identical to OLS except that we have scaled the data matrix and the observation vector by $\mathbf{\Omega}^{1/2}$, and we conclude that

$$\hat{\mathbf{w}}_{\text{WLS}} = \left( (\mathbf{\Omega}^{1/2}\mathbf{X})^\top(\mathbf{\Omega}^{1/2}\mathbf{X}) \right)^{-1} (\mathbf{\Omega}^{1/2}\mathbf{X})^\top\mathbf{\Omega}^{1/2}\mathbf{y} = (\mathbf{X}^\top\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}\mathbf{y}$$

**Probabilistic View**

As in MLE, we assume that our observations $\mathbf{y}$ are noisy, but now suppose that some of the $y_i$'s are more noisy than others. How can we take this into account in our learning algorithm so we can get a better estimate of the weights? Our probabilistic model looks like

$$Y_i = \mathbf{x}_i^\top \mathbf{w} + Z_i$$

where the $Z_i$'s are still independent Gaussians random variables, but not necessarily identical: $Z_i \sim \mathcal{N}(0, \sigma_i^2)$. Jointly, we have that $\mathbf{Z} \sim \mathcal{N}(\mathbf{\mu}_Z, \mathbf{\Sigma}_Z)$, where

$$\mathbf{\Sigma}_Z = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

We can morph the problem into an MLE one by scaling the data to make sure all the $Z_i$'s are identically distributed, by dividing by $\sigma_i$:

$$\frac{Y_i}{\sigma_i} = \frac{\mathbf{x}_i^\top \mathbf{w}}{\sigma_i} + Z_i$$

Note that the scaled noise entries are now i.i.d:

$$\frac{Z_i}{\sigma_i} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

Jointly, we can express this change of coordinates as

$$\mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{y} \sim \mathcal{N}(\mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{X} \mathbf{w}, \mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{\Sigma}_Z \mathbf{\Sigma}_Z^{-\frac{1}{2}})^{-1} = \mathcal{N}(\mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{X} \mathbf{w}, \mathbf{I})$$

This change of variable is sometimes called the **reparameterization trick**. Now that the noise is i.i.d. using the change of coordinates, we rewrite our original problem as a scaled MLE problem:

$$\hat{\mathbf{w}}_{\text{WLS}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{\sigma_i^2} \right) + n \log \sqrt{2\pi}$$

$$= \arg\min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - \mathbf{x}_i^\top \mathbf{w})^2$$

The MLE estimate of this scaled problem is equivalent to the WLS estimate of the original problem:

$$\hat{\mathbf{w}}_{\text{WLS}} = (\mathbf{X}^\top \mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{y} = (\mathbf{X}^\top \mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Sigma}_Z^{-\frac{1}{2}} \mathbf{y}$$

As long as no $\sigma$ is 0, $\mathbf{\Sigma}_Z$ is invertible. Note that $\omega_i$ from the optimization perspective is directly related to $\sigma_i^2$ from the probabilistic perspective: $\omega_i = \frac{1}{\sigma_i^2}$. Or at the level of matrices, $\mathbf{\Omega} = \mathbf{\Sigma}_Z^{-1}$. As the variance $\sigma_i^2$ of the noise corresponding to data point $i$ decreases, the weight $\omega_i$ increases: we are more concerned about fitting data point $i$ because it is likely to match the true underlying denoised point. Inversely, as the variance $\sigma_i^2$ increases, the weight $\omega_i$ decreases: we are less concerned about fitting data point $i$ because it is noisy and should not be trusted.
CHAPTER 2. REGRESSION II

Generalized Least Squares

Now let’s consider the case when the noise random variables are dependent on one another. We have

\[ Y = Xw + Z \]

where \( Z \) is now a jointly Gaussian random vector. That is,

\[ Z \sim \mathcal{N}(0, \Sigma_Z), \quad Y \sim \mathcal{N}(Xw, \Sigma_Z) \]

This problem is known as generalized least squares. Our goal is to maximize the probability of our data over the set of possible \( w \)’s:

\[
\hat{w}_{\text{gls}} = \arg \max_{w \in \mathbb{R}^d} \frac{1}{\sqrt{\det(\Sigma_Z)}} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}(y-Xw)^\top \Sigma_Z^{-1}(y-Xw)}
\]

\[
= \arg \min_{w \in \mathbb{R}^d} (y - Xw)^\top \Sigma_Z^{-1}(y - Xw)
\]

The optimization problem is therefore given by

\[
\hat{w}_{\text{gls}} = \arg \min_{w \in \mathbb{R}^d} (y - Xw)^\top \Sigma_Z^{-1/2}(y - Xw)
\]

Since \( \Sigma_Z \) is symmetric, we can decompose it into its eigen structure using the spectral theorem:

\[
\Sigma_Z = Q \left[ \begin{array}{cccc}
\sigma_1^2 & 0 & \cdots & 0 \\
0 & \sigma_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & \cdots & \sigma_n^2
\end{array} \right] Q^\top
\]

where \( Q \) is orthonormal. As before with weighted least squares, our goal is to find an appropriate linear transformation so that we can reduce the problem into the i.i.d. case.

Consider

\[
\Sigma_Z^{-\frac{1}{2}} = Q \left[ \begin{array}{cccc}
\frac{1}{\sigma_1} & 0 & \cdots & 0 \\
0 & \frac{1}{\sigma_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & \cdots & \frac{1}{\sigma_n}
\end{array} \right] Q^\top
\]

We can scale the data to morph the problem into an MLE problem with i.i.d. noise variables, by premultiplying the data matrix \( X \) and the observation vector \( y \) by \( \Sigma_Z^{-\frac{1}{2}} \). Jointly, we can express this change of coordinates as

\[
\Sigma_Z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_Z^{-\frac{1}{2}} Xw, \Sigma_Z^{-\frac{1}{2}} \Sigma_Z \Sigma_Z^{-\frac{1}{2}}) = \mathcal{N}(\Sigma_Z^{-\frac{1}{2}} Xw, I).
\]

Consequently, in a very similar fashion to the independent noise problem, the MLE of the scaled dependent noise problem is

\[
\hat{w}_{\text{gls}} = (X^\top \Sigma_Z^{-1} X)^{-1} X^\top \Sigma_Z^{-1} y.
\]
“Ridge Regression” with Dependent Parameters

In the ordinary least squares (OLS) statistical model, we assume that the output $Y$ is a linear function of the input, plus some Gaussian noise. We take this one step further in MAP estimation, where we assume that the weights are a random variable. The new statistical model is

$$Y = XW + Z$$

where $Y$ and $Z$ are $n$-dimensional random vectors, $W$ is a $d$-dimensional random vector, and $X$ is a fixed $n \times d$ matrix. Note that random vectors are not notationally distinguished from matrices here, so keep in mind what each symbol represents.

We have seen that ridge regression can be derived by assuming a prior distribution on $W$ in which $W_i$ are i.i.d. (univariate) Gaussian, or equivalently,

$$W \sim \mathcal{N}(0, I)$$

But more generally, we can allow $W$ to be any multivariate Gaussian:

$$W \sim \mathcal{N}(\mu_W, \Sigma_W)$$

Recall that we can rewrite a multivariate Gaussian variable as an affine transformation of a standard Gaussian variable:

$$W = \Sigma_W^{1/2}V + \mu_W, \quad V \sim \mathcal{N}(0, I)$$

Plugging this parameterization into our previous statistical model gives

$$Y = X(\Sigma_W^{1/2}V + \mu_W) + Z$$

But this can be re-written

$$Y - X\mu_W = X\Sigma_W^{1/2}V + Z$$

which we see has the form of the statistical problem that underlies traditional Ridge Regression with $\lambda = 1$, and therefore

$$\hat{v} = (\Sigma_W^{T/2}X^T\Sigma_W^{1/2} + I)^{-1}\Sigma_W^{T/2}X^T(y - X\mu_W)$$

However $V$ is not what we care about – we need to convert back to the actual weights $W$ in order to make predictions. Since $W$ is completely determined by $V$ (assuming fixed mean and covariance),

$$\hat{w} = \Sigma_W^{1/2}\hat{v} + \mu_W$$

but

$$\hat{w} = \mu_W + \Sigma_W^{1/2}(\Sigma_W^{T/2}X^T\Sigma_W^{1/2} + I)^{-1}\Sigma_W^{T/2}X^T(y - X\mu_W)$$

$$= \mu_W + (X^T\Sigma_W^{-1}X + \Sigma_W^{-1/2}\Sigma_W^{-T/2})^{-1}X^T(y - X\mu_W)$$

Note that there are two terms: the prior mean $\mu_W$, plus another term that depends on both the data and the prior. The positive-definite precision matrix of $W$'s prior ($\Sigma_W^{-1}$) controls how the data fit error affects our estimate. This is called Tikhonov regularization in the literature and generalizes ridge regularization.
To gain intuition, let us consider the simplified case where

$$\Sigma_{\mathbf{W}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}$$

When the prior variance $\sigma_j^2$ for dimension $j$ is large, the prior is telling us that $W_j$ may take on a wide range of values. Thus we do not want to penalize that dimension as much, preferring to let the data fit sort it out. And indeed the corresponding entry in $\Sigma_{\mathbf{W}}^{-1}$ will be small, as desired.

Conversely if $\sigma_j^2$ is small, there is little variance in the value of $W_j$, so $W_j \approx \mu_j$. As such we penalize the magnitude of the data-fit contribution to $\hat{W}_j$ more heavily.

If all the $\sigma_j^2$ are the same, then we have traditional ridge regularization.

**Alternative derivation: directly conditioning jointly Gaussian random variables**

In an explicitly probabilistic perspective, MAP with colored noise (and known $\mathbf{X}$) can be expressed as:

$$\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2.6}$$

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{W} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_Z & \mathbf{XR}_W \\ \mathbf{0} & \mathbf{R}_W \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \tag{2.7}$$

where $\mathbf{R}_Z$ and $\mathbf{R}_W$ are relationships with $\mathbf{W}$ and $\mathbf{Z}$, respectively. Note that the $\mathbf{R}_W$ appears twice because our model assumes $\mathbf{Y} = \mathbf{XW} + \text{noise}$, so if $\mathbf{W} = \mathbf{R}_W \mathbf{V}$, then we must have $\mathbf{Y} = \mathbf{XR}_W \mathbf{V} + \text{noise}$.

We want to find the posterior $\mathbf{W} \mid \mathbf{Y} = \mathbf{y}$. The formulation above makes it relatively easy to find the posterior of $\mathbf{Y}$ conditioned on $\mathbf{W}$ (see below), but not vice-versa. So let’s pretend instead that

$$\begin{bmatrix} \mathbf{U}' \\ \mathbf{V}' \end{bmatrix} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{U}' \\ \mathbf{V}' \end{bmatrix}$$

Now $\mathbf{W} \mid \mathbf{Y} = \mathbf{y}$ is straightforward. Since $\mathbf{V}' = \mathbf{D}^{-1} \mathbf{Y}$, the conditional mean and variance of $\mathbf{W} \mid \mathbf{Y} = \mathbf{y}$ can be computed as follows:

$$\mathbb{E}[\mathbf{W} \mid \mathbf{Y} = \mathbf{y}] = \mathbb{E}[\mathbf{AU}' + \mathbf{BV}' \mid \mathbf{Y} = \mathbf{y}]$$

$$= \mathbb{E}[\mathbf{AU}' \mid \mathbf{Y} = \mathbf{y}] + \mathbb{E}[\mathbf{BD}^{-1} \mathbf{Y} \mid \mathbf{Y} = \mathbf{y}]$$

$$= \mathbf{A} \mathbb{E}[\mathbf{U}'] + \mathbb{E}[\mathbf{BD}^{-1} \mathbf{Y} \mid \mathbf{Y} = \mathbf{y}]$$

$$= \mathbf{B} \mathbf{D}^{-1} \mathbf{y}$$

$$\text{Var}(\mathbf{W} \mid \mathbf{Y} = \mathbf{y}) = \mathbb{E}[(\mathbf{W} - \mathbb{E}[\mathbf{W}])(\mathbf{W} - \mathbb{E}[\mathbf{W}])^\top \mid \mathbf{Y} = \mathbf{y}]$$

$$= \mathbb{E}[(\mathbf{AU}' + \mathbf{BD}^{-1} \mathbf{Y} - \mathbf{BD}^{-1} \mathbf{Y})(\mathbf{AU}' + \mathbf{BD}^{-1} \mathbf{Y} - \mathbf{BD}^{-1} \mathbf{Y})^\top \mid \mathbf{Y} = \mathbf{y}]$$

$$= \mathbb{E}[(\mathbf{AU}')^\top(\mathbf{AU}') \mid \mathbf{Y} = \mathbf{y}]$$

$$= \mathbb{E}[\mathbf{AU}'(\mathbf{U}')^\top \mathbf{A}^\top]$$
\[ = \mathbf{A} \mathbb{E} [\mathbf{U}'(\mathbf{U}')^\top] \mathbf{A}^\top \]
\[= \text{Var}(\mathbf{U}') = \mathbf{I} \]
\[= \mathbf{A} \mathbf{A}^\top \]

In both cases above where we drop the conditioning on \( \mathbf{Y} \), we are using the fact \( \mathbf{U}' \) is independent of \( \mathbf{V}' \) (and thus independent of \( \mathbf{Y} = \mathbf{D} \mathbf{V}' \)). Therefore

\[ \mathbf{W} \mid \mathbf{Y} = \mathbf{y} \sim \mathcal{N}(\mathbf{B} \mathbf{D}^{-1} \mathbf{y}, \mathbf{A} \mathbf{A}^\top) \]

Recall that a Gaussian distribution is completely specified by its mean and covariance matrix. We see that the covariance matrix of the joint distribution is

\[ \mathbb{E} \left[ \begin{bmatrix} \mathbf{W} \\ \mathbf{Y} \end{bmatrix} \begin{bmatrix} \mathbf{W}^\top \\ \mathbf{Y}^\top \end{bmatrix} \right] = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{A}^\top & \mathbf{0} \\ \mathbf{B}^\top & \mathbf{D}^\top \end{bmatrix} \]
\[= \begin{bmatrix} \mathbf{A} \mathbf{A}^\top + \mathbf{B} \mathbf{B}^\top & \mathbf{B} \mathbf{D}^\top \\ \mathbf{D} \mathbf{B}^\top & \mathbf{D} \mathbf{D}^\top \end{bmatrix} \]
\[= \begin{bmatrix} \Sigma_{\mathbf{W}} & \Sigma_{\mathbf{W}, \mathbf{Y}} \\ \Sigma_{\mathbf{Y}, \mathbf{W}} & \Sigma_{\mathbf{Y}} \end{bmatrix} \]

Matching the corresponding terms, we can express the conditional mean and variance of \( \mathbf{W} \mid \mathbf{Y} = \mathbf{y} \) in terms of these (cross-)covariance matrices:

\[ \mathbf{B} \mathbf{D}^{-1} \mathbf{Y} = \mathbf{B} \mathbf{D} D^{-1} \mathbf{Y} = \left( \mathbf{B} \mathbf{D} \right) \left( \mathbf{D} \mathbf{D}^\top \right)^{-1} \mathbf{Y} = \Sigma_{\mathbf{W}, \mathbf{Y}} \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y} \]
\[\mathbf{A} \mathbf{A}^\top = \mathbf{A} \mathbf{A}^\top + \mathbf{B} \mathbf{B}^\top - \mathbf{B} \mathbf{D} D^{-1} \mathbf{D}^\top \mathbf{B}^\top \]
\[= \mathbf{A} \mathbf{A}^\top + \mathbf{B} \mathbf{B}^\top - \left( \mathbf{B} \mathbf{D} D^{-1} \mathbf{D} \right) \mathbf{B}^\top \]
\[= \Sigma_{\mathbf{W}} - \Sigma_{\mathbf{W}, \mathbf{Y}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}, \mathbf{W}} \]

We can then apply the same reasoning to the original setup:

\[ \mathbb{E} \left[ \begin{bmatrix} \mathbf{Y} \\ \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{Y}^\top \\ \mathbf{W}^\top \end{bmatrix} \right] = \begin{bmatrix} \mathbf{R}_\mathbf{Z} \mathbf{R}_\mathbf{Z}^\top + \mathbf{X} \mathbf{R}_\mathbf{W} \mathbf{R}_\mathbf{W}^\top \mathbf{X}^\top & \mathbf{X} \mathbf{R}_\mathbf{W} \mathbf{R}_\mathbf{W}^\top \\ \mathbf{R}_\mathbf{W} \mathbf{R}_\mathbf{W}^\top \mathbf{X}^\top & \mathbf{R}_\mathbf{W} \mathbf{R}_\mathbf{W}^\top \end{bmatrix} \]
\[= \begin{bmatrix} \Sigma_{\mathbf{Y}} & \Sigma_{\mathbf{Y}, \mathbf{W}} \\ \Sigma_{\mathbf{W}, \mathbf{Y}} & \Sigma_{\mathbf{W}} \end{bmatrix} \]

Therefore after defining \( \Sigma_{\mathbf{Z}} = \mathbf{R}_\mathbf{Z} \mathbf{R}_\mathbf{Z}^\top \), we can read off

\[ \Sigma_{\mathbf{W}} = \mathbf{R}_\mathbf{W} \mathbf{R}_\mathbf{W}^\top \]
\[ \Sigma_{\mathbf{Y}} = \Sigma_{\mathbf{Z}} + \mathbf{X} \Sigma_{\mathbf{W}} \mathbf{X}^\top \]
\[ \Sigma_{\mathbf{Y}, \mathbf{W}} = \mathbf{X} \Sigma_{\mathbf{W}} \]
\[ \Sigma_{\mathbf{W}, \mathbf{Y}} = \Sigma_{\mathbf{W}} \mathbf{X}^\top \]

Plugging this into our estimator yields

\[ \hat{\mathbf{w}} = \mathbb{E}[\mathbf{W} \mid \mathbf{Y} = \mathbf{y}] \]
\[= \Sigma_{\mathbf{W}, \mathbf{Y}} \Sigma_{\mathbf{Y}}^{-1} \mathbf{y} \]
This operation involves calculating \( \Phi \). To achieve this goal, recall the solution to ridge regression:

\[
\hat{w} = (X^\top \Sigma_z^{-1} X + \Sigma_w^{-1})^{-1} X^\top \Sigma_z^{-1} y
\]

which looks more familiar. In fact, you can recognize this as the general solution when we have both a generic Gaussian prior on the parameters and colored noise in the observations.

### Summary of Linear Gaussian Statistical Models

We have seen a number of related linear models, with varying assumptions about the randomness in the observations and the weights. We summarize these below:

<table>
<thead>
<tr>
<th>( W )</th>
<th>( Z )</th>
<th>( \mathcal{N}(0, \mathbf{I}) )</th>
<th>( \mathcal{N}(0, \Sigma_z) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>No prior</td>
<td>( \hat{w}_{\text{ols}} = (X^\top X)^{-1} X^\top y )</td>
<td>( \hat{w}_{\text{gls}} = (X^\top \Sigma_z^{-1} X)^{-1} X^\top \Sigma_z^{-1} y )</td>
<td></td>
</tr>
<tr>
<td>( \mathcal{N}(0, \lambda^{-1} \mathbf{I}) )</td>
<td>( \hat{w}_{\text{ridge}} = (X^\top X + \lambda \mathbf{I})^{-1} X^\top y )</td>
<td>( (X^\top \Sigma_z^{-1} X + \lambda \mathbf{I})^{-1} X^\top \Sigma_z^{-1} y )</td>
<td></td>
</tr>
<tr>
<td>( \mathcal{N}(\mu_w, \Sigma_w) )</td>
<td>( \mu_w + (X^\top X + \Sigma_w^{-1})^{-1} X^\top (y - X \mu_w) )</td>
<td>( \mu_w + (X^\top \Sigma_z^{-1} X + \Sigma_w^{-1})^{-1} X^\top \Sigma_z^{-1} (y - X \mu_w) )</td>
<td></td>
</tr>
</tbody>
</table>

#### 2.5 Kernels and Ridge Regression

In ridge regression, we given a vector \( y \in \mathbb{R}^n \) and a matrix \( X \in \mathbb{R}^{n \times \ell} \), where \( n \) is the number of training points and \( \ell \) is the dimension of the raw data points. In most settings we don’t want to work with just the raw feature space, so we augment features to the data points and replace \( X \) with \( \Phi \in \mathbb{R}^{n \times d} \), where \( \Phi \in \Phi : \phi(x_i) \in \mathbb{R}^d \). Then we solve a well-defined optimization problem that involves \( \Phi \) and \( y \), over the parameters \( w \in \mathbb{R}^d \). Note the problem that arises here. If we have polynomial features of degree at most \( p \) in the raw \( \ell \) dimensional space, then there are \( d = \binom{\ell + p}{p} \) terms that we need to optimize, which can be very, very large (much larger than the number of training points \( n \)). Wouldnt’ it be useful, if instead of solving an optimization problem over \( d \) variables, we could solve an equivalent problem over (potentially much smaller) \( n \) variables, and achieve a computational runtime independent of the number of augmented features? As it turns out, the concept of kernels (in addition to a technique called the kernel trick) will allow us to achieve this goal. Recall the solution to ridge regression:

\[
w^* = (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \Phi^\top y
\]

This operation involves calculating \( \Phi^\top \Phi \), which is a \( d \times d \) matrix and takes \( O(d^3 n) \) time to compute. The matrix inversion operation takes an additional \( O(d^3) \) time to compute. What we would really like to do is to have an \( n \times n \) matrix. It takes \( O(n^3) \) time to invert. Here’s a simple observation: if we flip the order of \( \Phi^\top \) and \( \Phi \), we end up with an \( n \times n \) matrix \( \Phi \Phi^\top \). In fact, the matrix \( \Phi \Phi^\top \) has a very intuitive meaning: it is the matrix of inner products between all of the augmented datapoints, which in loose terms measures the “similarity” among the properties of the datapoints and their relationship. Now let’s see if we could somehow express the solution to ridge regression using the matrix \( \Phi \Phi^\top \).

\[\begin{align*}
(A + UCV)^{-1} &= A^{-1} - A^{-1} U(C^{-1} + VA^{-1} U)^{-1} VA^{-1}
\end{align*}\]
2.5. KERNS AND RICE REGRESSION

Derivation

For simplicity of notation, let’s revert back to using $X$ instead of $\Phi$ (pretend that we are only working with raw features, our analysis of kernel ridge regression still holds if we use just the raw features). Rearranging the terms of the original ridge regression solution, we have

$$w = (X^T X + \lambda I)^{-1} X^T y$$

$$(X^T X + \lambda I)w = X^T y$$

$$X^T X w + \lambda w = X^T y$$

$$\lambda w = X^T y - X^T X w$$

$$w = \frac{1}{\lambda} (X^T y - X^T X w)$$

$$w = \frac{X^T y - X^T X w}{\lambda}$$

which says that whatever $w$ is, it is some linear combination of the training points $x_i$ (because anything of the form $X^T v$ is a linear combination of the columns of $X^T$, which are the training points). To find $w$ it suffices to find $v$, where $w = X^T v$.

Recall that the relationship we have to satisfy is $X^T X w - \lambda w = X^T y$. Let’s assume that we had $v$, and just substitute $X^T v$ in for all the $w$’s.

$$X^T X (X^T v) + \lambda (X^T v) = X^T y$$

$$X^T X^T v + X^T (\lambda v) = X^T y$$

$$X^T (X X^T v + \lambda v) = X^T (y)$$

We can’t yet isolate $v$ and have a closed-form solution for it, but we can make the observation that if we found an $v$ such that we had

$$XX^T v + \lambda v = y$$

that would imply that this $v$ also satisfies the above equation. Note that we did not “cancel the $X^T$’s on both sides of the equation.” We saw that having $v$ satisfy one equation implied that it satisfied the other as well. So, indeed, we can isolate $v$ in this new equation:

$$(XX^T + \lambda I) v = y \implies v^* = (XX^T + \lambda I)^{-1} y$$

and have that the $v$ which satisfies this equation will be such that $X^T v$ equals $w$. We conclude that the optimal $w$ is

$$w^* = X^T v^* = X^T (XX^T + \lambda I)^{-1} y$$

Recall that previously, we derived ridge regression and ended up with

$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

In fact, these two are equivalent expressions! The question that now arises is which expression should you pick? Which is more efficient to calculate? We will answer this question after we introduce kernels.
Linear Algebra Derivation

The previous derivation involved using some intuitive manipulations to achieve the desired answer. Let’s formalize our derivation using more principled arguments from linear algebra and optimization.

Before we do so, we must first introduce the **Fundamental Theorem of Linear Algebra (FTLA):** Suppose that there is a matrix (linear map) $X$ that maps $\mathbb{R}^f$ to $\mathbb{R}^n$. Denote $\mathcal{N}(X)$ as the nullspace of $X$, and $\mathcal{R}(X)$ as the range of $X$. Then the following properties hold:

1. $\mathcal{N}(X) \perp \mathcal{R}(X^\top) = \mathbb{R}^f$ and $\mathcal{N}(X^\top) \perp \mathcal{R}(X) = \mathbb{R}^n$ by symmetry

   The symbol $\perp$ indicates that we taking a direct sum of $\mathcal{N}(X)$ and $\mathcal{R}(X^\top)$, which means that $\forall u \in \mathbb{R}^f$ there exist unique elements $u_1 \in \mathcal{N}(X)$ and $u_2 \in \mathcal{R}(X^\top)$ such that $u = u_1 + u_2$.

   Furthermore, the symbol $\perp$ indicates that $\mathcal{N}(X)$ and $\mathcal{R}(X^\top)$ are orthogonal subspaces.

2. $\mathcal{N}(X^\top X) = \mathcal{N}(X)$ and $\mathcal{N}(XX^\top) = \mathcal{N}(X^\top)$ by symmetry

3. $\mathcal{R}(X^\top X) = \mathcal{R}(X^\top)$ and $\mathcal{R}(XX^\top) = \mathcal{R}(X)$ by symmetry.

Here’s where FTLA comes, in the context of kernel ridge regression. We know that we can express any $w \in \mathbb{R}^f$ as a unique combination $w = w_1 + w_2$, where $w_1 \in \mathcal{R}(X^\top)$ and $w_2 \in \mathcal{N}(X)$. Equivalently we can express this as $w = X^\top v + r$, where $v \in \mathbb{R}^n$ and $r \in \mathcal{N}(X)$. Now, instead of optimizing over $w \in \mathbb{R}^f$, we can optimize over $v \in \mathbb{R}^n$ and $r \in \mathbb{R}^f$, which equates to optimizing over $n + \ell$ variables. However, as we shall see, the optimization over $r$ will be trivial so we just have to optimize an $n$ dimensional problem.

We know that $w = X^\top v + r$, where $v \in \mathbb{R}^n$ and $r \in \mathcal{N}(X)$. Let’s now solve ridge regression by optimizing over the variables $v$ and $r$ instead of $w$:

\[
\begin{align*}
v^*, r^* &= \arg\min_{v \in \mathbb{R}^n, r \in \mathcal{N}(X)} \|Xw - y\|^2_2 + \lambda\|w\|^2_2 \\
&= \arg\min_{v \in \mathbb{R}^n, r \in \mathcal{N}(X)} \|X(X^\top v + r) - y\|^2_2 + \lambda\|X^\top v + r\|^2_2 \\
&= \arg\min_{v \in \mathbb{R}^n, r \in \mathcal{N}(X)} \|XX^\top v + Xr - y\|^2_2 + \lambda\|X^\top v + r\|^2_2 \\
&= \arg\min_{v \in \mathbb{R}^n, r \in \mathcal{N}(X)} \Big( v^\top XX^\top XX^\top v - 2v^\top XX^\top y + y^\top y \Big) + \lambda \Big( v^\top XX^\top v + 2v^\top Xr + r^\top r \Big) \\
&= \arg\min_{v \in \mathbb{R}^n, r \in \mathcal{N}(X)} \Big( v^\top XX^\top XX^\top v - 2v^\top XX^\top y \Big) + \lambda \Big( v^\top XX^\top v + r^\top r \Big)
\end{align*}
\]

We crossed out $Xr$ and $2v^\top Xr$ because $r \in \mathcal{N}(X)$ and therefore $Xr = 0$. Now we are optimizing over $L(v, r)$, which is jointly convex in $v$ and $r$, because its Hessian is PSD. Let’s show that this is indeed the case:

\[
\nabla^2_r L(v, r) = 2I \succeq 0 \\
\nabla_r \nabla_v L(v, r) = \nabla_v \nabla_r L(v, r) = 0 \\
\nabla^2_v L(v, r) = 2XX^\top XX^\top + 2\lambda XX^\top \succeq 0
\]

Since the cross terms of the Hessian are 0, it suffices that $\nabla^2_r L(v, r)$ and $\nabla^2_v L(v, r)$ are PSD to establish joint convexity. With joint convexity established, we can set the gradient to 0 w.r.t $r$ and $v$ and obtain the global minimum:

\[
\nabla_r L(v, r^*) = 2r^* = 0 \implies r^* = 0
\]
Note that $r^* = 0$ just so happens in to be in $N(X)$, so it is a feasible point.

$$
\nabla_v L(v^*, r^*) = 2XX^TV^rv^* - 2XX^Tv + 2\lambda XX^Tv^* = 0
\implies XX^T(XX^T + \lambda I)v^* = XX^Tv
\implies v^* = (XX^T + \lambda I)^{-1}y
$$

Note that $XX^T + \lambda I$ is positive definite and therefore invertible, so we can compute $(XX^T + \lambda I)^{-1}y$. Even though $(XX^T + \lambda I)^{-1}y$ is a critical point for which the gradient is 0, it must achieve the global minimum because the objective is jointly convex. We conclude that 

$$
w^* = X^T(XX^T + \lambda I)^{-1}y
$$

and arrive at the same solution as in the previous derivation.

**Non-i.i.d. Case**

So far we have assumed the special i.i.d. case of ridge regression, where

$$Y|W \sim N(XW, \sigma^2I), \quad W \sim N(0, \sigma^2I)$$

In the non-i.i.d case we consider arbitrary covariance matrices:

$$Y|W \sim N(XW, \Sigma_Z), \quad W \sim N(0, \Sigma_W)$$

As we’ve seen already, the solution in this case can be expressed in two forms, either the familiar case

$$w^* = (X^T\Sigma_Z^{-1}X + \Sigma_W^{-1})^{-1}X^T\Sigma_Z^{-1}y$$

or the case that we desire in kernel ridge regression

$$w^* = \Sigma_WX^T(X\Sigma_WX^T + \Sigma_Z)^{-1}y$$

The principal difference in the non-i.i.d case is that we are computing $X\Sigma_WX^T$ as opposed to $XX^T$.

**Kernels**

Having derived the kernel ridge regression formulation for the raw data matrix $X$, we can apply the exact same logic to the augmented data matrix $\Phi$ and replace the optimal expression with

$$w^* = \Phi^T(\Phi\Phi^T + \lambda I)^{-1}y$$

Let’s explore the $\Phi\Phi^T$ term in kernel ridge regression in more detail:

$$
\Phi\Phi^T = \begin{pmatrix}
\phi_1^T & \phi_2^T & \cdots & \phi_n^T \\
\phi_1^T & \phi_2^T & \cdots & \phi_n^T \\
\vdots & \vdots & \ddots & \vdots \\
\phi_1^T & \phi_2^T & \cdots & \phi_n^T
\end{pmatrix} \begin{pmatrix}
\phi_1 & \phi_2 & \cdots & \phi_n \\
\phi_1 & \phi_2 & \cdots & \phi_n \\
\vdots & \vdots & \ddots & \vdots \\
\phi_1 & \phi_2 & \cdots & \phi_n
\end{pmatrix} = 
\begin{pmatrix}
\phi_1^T\phi_1 & \phi_1^T\phi_2 & \cdots \\
\phi_2^T\phi_1 & \phi_2^T\phi_2 & \cdots \\
\vdots & \vdots & \ddots
\end{pmatrix}
$$

Each entry $\Phi\Phi^T_{ij}$ is a dot product between $\phi(x_i)$ and $\phi(x_j)$ and can be interpreted as a similarity measure:

$$\Phi\Phi^T_{ij} = \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$$
where $k(.,.)$ is the kernel function. The kernel function takes raw-feature inputs and outputs their inner product in the augmented feature space. We denote the matrix of $k(x_i, x_j)$ terms as the Gram matrix and denote it as $K$:

$$K = \Phi^\top \Phi = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \ldots \\ k(x_2, x_1) & \ddots \\ \vdots & & k(x_n, x_n) \end{pmatrix}$$

Formally, $k(x_i, x_j)$ is defined to be a valid kernel function if either of the following definitions are met:

- There exists a feature map $\phi(.)$ such that $\forall x_i, x_j, \ k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$
- For all sets $D = \{x_1, x_2, \ldots, x_n\}$, the Gram matrix $K(D)$ is PSD

We will now state some basic properties of kernels.

- Given two valid kernels $k_a$ and $k_b$, their linear combination
  $$k(x_i, x_j) = \alpha k_a(x_i, x_j) + \beta k_b(x_i, x_j)$$
  where $\alpha, \beta \geq 0$ is also a valid kernel. We can show this from the second property:
  $$\forall v \in \mathbb{R}^n, v^\top (\alpha K_a + \beta K_b)v = \alpha v^\top K_a v + \beta v^\top K_b v \geq 0$$

- Given a positive semidefinite matrix $\Sigma$,
  $$k(x_i, x_j) = \phi(x_i)^\top \Sigma \phi(x_j)$$
  is a valid kernel. We can show this from the first property: $\tilde{\phi}(x_i) = \Sigma^{1/2} \phi(x_i)$

- Given a valid kernel $k_a$,
  $$k(x_i, x_j) = f(x_i)f(x_j)k_a(x_i, x_j)$$
  is a valid kernel. We can show this from the first property: $\tilde{\phi}(x_i) = f(x_i)\phi(x_i)$

Computing the each Gram matrix entry $k(x_i, x_j)$ can be done in a straightforward fashion if we apply the feature map to $x_i$ and $x_j$ and then take their dot product in the augmented feature space — this takes $O(d)$ time, where $d$ is the dimensionality of the problem in the augmented feature space. However, if we use the kernel trick, we can perform this operation in $O(\ell + \log p)$ time, where $\ell$ is the dimensionality of the problem in the raw feature space and $p$ is the degree of the polynomials in the augmented feature space.

### Kernel Trick

Suppose that we are computing $k(x, z)$, using a $p$-degree polynomial feature map that maps $\ell$ dimensional inputs to $d = O(\ell^p)$ dimensional outputs. Let’s take $p = 2$ and $\ell = 2$ as an example. Define the polynomial feature map as

$$\phi(x) = \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2}x_1x_2 & \sqrt{2}x_1 & \sqrt{2}x_2 & 1 \end{bmatrix}^\top$$
the kernel function can be expressed as
\[ k(x, z) = \phi(x)\top\phi(z) \]
\[ = \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2}x_1x_2 & \sqrt{2}x_1 & \sqrt{2}x_2 & 1 \end{bmatrix}\top \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2}z_1z_2 & \sqrt{2}z_1 & \sqrt{2}z_2 & 1 \end{bmatrix} \]
\[ = \begin{bmatrix} x_1^2 & x_2^2 & 2x_1z_1x_2z_2 + 2x_1z_1 + 2x_2z_2 + 1 \end{bmatrix} \]
\[ = (x_1z_1 + x_2z_2)^2 + 2(x_1z_1 + x_2z_2) + 1 \]
\[ = (x\top z)^2 + 2x\top z + 1 \]
\[ = (x\top z + 1)^2 \]

We can compute \( k(x, z) \) either by

1. Raising the inputs to the augmented feature space and take their inner product
2. Computing \( (x\top z + 1)^2 \), which involves an inner product of the raw-feature inputs

Clearly, the latter option is much cheaper to calculate, taking \( O(\ell + \log p) \) time, instead of \( O(\ell^p) \) time. In fact, this concept generalizes for any arbitrary \( \ell \) and \( p \), and for \( p \)-degree polynomial features, we have that
\[ k(x, z) = (x\top z + 1)^p \]
The kernel trick makes computations significantly cheaper to perform, making kernelization much more appealing! The takeaway here is that no matter what the degree \( p \) is, the computational complexity is the same — it is only dependent on the dimensionality of the raw feature space!

Note that we can equivalently express the degree-2 polynomial features problem using the more natural mapping
\[ \tilde{\phi}(x) = \begin{bmatrix} x_1^2 & x_2^2 & x_1x_2 & x_1 & x_2 & 1 \end{bmatrix}\top \]
in which case the kernel function would be expressed as
\[ k(x, z) = \tilde{\phi}(x)\top \Sigma \tilde{\phi}(z) = (x\top z + 1)^2, \quad \Sigma = \text{Diag} (1, 1, 2, 2, 2, 1) \]
Thus we can view kernel ridge regression with the kernel trick in two ways:

1. i.i.d. prior \( W \sim \mathcal{N} \left( \mathbf{0}, \text{Diag} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \right) \), using the feature mapping \( \phi(x) \)
2. non-i.i.d prior \( W \sim \mathcal{N} \left( \mathbf{0}, \text{Diag} \begin{bmatrix} 1 & 1 & 2 & 2 & 2 & 1 \end{bmatrix} \right) \), using the feature mapping \( \tilde{\phi}(x) \) (note that the kernel trick is only applicable for this specific setting of \( \Sigma \) — it does not necessarily apply to arbitrary \( \Sigma \).)

**Computational Analysis**

Back to the original question: in ridge regression, should we compute
\[ w^* = \Phi\top(\Phi\Phi\top + \lambda I)^{-1}y \]
or
\[ w^* = (\Phi\top\Phi + \lambda I)^{-1}\Phi\top y \]
Let’s compare their computational complexities. Suppose you are given an arbitrary test point \( z \in \mathbb{R}^\ell \), and you would like to compute its predicted value \( \hat{y} \). Let’s see how these values are calculated in each case:
1. Kernelized

\[ \hat{y} = \langle \phi(z), w^* \rangle = \phi(z)^\top (\Phi \Phi^\top + \lambda I)^{-1} y = [k(x_1, z) \ldots k(x_n, z)] (K + \lambda I)^{-1} y \]

Computing the K term takes \( O(n^2(\ell + \log p)) \), and inverting the matrix takes \( O(n^3) \). These two computations dominate, for a total computation time of \( O(n^3 + n^2(\ell + \log p)) \).

2. Non-kernelized

\[ \hat{y} = \langle \phi(z), w^* \rangle = \phi(z)^\top (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top y \]

Computing the \( \Phi^\top \Phi \) term takes \( O(d^2 n) \), and inverting the matrix takes \( O(d^3) \). These two computations dominate, for a total computation time of \( O(d^3 + d^2 n) \).

Here is the takeaway: if \( d \ll n \), the non-kernelized method is preferable. Otherwise if \( n \ll d \), the kernelized method is preferable.

### 2.6 Sparse Least Squares

Suppose we want to solve the least squares objective, subject to a constraint that \( w \) is sparse. Mathematically this is expressed as

\[
\begin{align*}
\min_w \quad & \|Xw - y\|_2^2 \\
\text{s.t.} \quad & \|w\|_0 \leq k
\end{align*}
\]

where the \( \ell_0 \) norm of \( w \) is simply the number of non-zero elements in \( w \). This quantity is otherwise known as the Hamming Distance between \( w \) and \( 0 \), the vector of zeros.

There are several motivations for designing optimization problems with sparse solutions. One advantage is that sparse weights speed up testing time. In the context of primal problems, if the weight vector \( w \) is sparse, then after we compute \( w \) in training, we can discard features/dimensions with 0 weight, as they will contribute nothing to the evaluation of the hypothesized regression values of test points. A similar reasoning applies to dual problems with dual weight vector \( v \), allowing us to discard the training points corresponding to dual weight 0, ultimately allowing for faster evaluation of our hypothesis function on test points.

Note that the \( \ell_0 \) norm does not actually satisfy the properties of a norm, evident by the fact that the it is not convex, a property that all norms share. Solving this optimization problem is NP-hard, so we instead aim to find a computationally feasible alternative method that can approximate the optimal solution. We will present two such methods: LASSO, a relaxed version of the problem that replaces the \( \ell_0 \) norm with a \( \ell_1 \) norm, and Matching pursuit, a greedy algorithm that iteratively updates one entry of \( w \) at a time until the sparsity constraint can not longer be satisfied.

**LASSO**

The least absolute shrinkage and selection operator (LASSO), introduced in 1996 by Robert Tibshirani, is identical to the sparse least squares objective, except that the \( \ell_0 \) norm penalizing \( w \) is now changed to an \( \ell_1 \) norm:

\[
\begin{align*}
\min_w \quad & \|Xw - y\|_2^2 \\
\text{s.t.} \quad & \|w\|_1 \leq k
\end{align*}
\]
2.6. SPARSE LEAST SQUARES

(The $k$ in the constraint is not necessarily the same $k$ in the sparse least squares objective.) The $\ell_1$ norm of $w$ is the sum of absolute values of its entries:

$$\|w\|_1 = \sum_{i=1}^{d} |w_i|$$

Unlike the $\ell_0$ norm, the $\ell_1$ norm actually satisfies the properties of norms. The relaxation from the $\ell_0$ to $\ell_1$ norm is desirable, because it makes the optimization problem convex, and is no longer NP-hard to solve. But does the $\ell_1$ norm still induce sparsity like the $\ell_0$? As we will see, the answer is yes!

Due to strong duality, we can equivalently express the LASSO problem in the unconstrained form

$$\min_w \|Xw - y\|^2 + \lambda \|w\|_1$$

We make a striking observation here: LASSO is identical to the ridge regression objective, except that the $\ell_2$ norm (squared) penalizing $w$ is now changed to an $\ell_1$ norm (with no squaring term).

Recall that the $\ell_2$ norm squared of $w$, the sum of squared values of its entries:

$$\|w\|^2_2 = \sum_{i=1}^{d} w_i^2$$

As it turns out, the simple change from the $\ell_2$ to $\ell_1$ norm inherently leads to a sparse solution. In fact, the sparsity inducing properties of the $\ell_1$ norm are not just unique to least squares. To illustrate the point, let's take a step away from least squares for a moment and discuss the $\ell_1$ norm in the context of SVMs. Recall the soft-margin SVM problem (constraints omitted for brevity):

$$\min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

The slack $\xi_i$ is constrained to be either positive or zero. Note that if a point $x_i$ has a nonzero slack $\xi_i > 0$, by definition it must lie inside the margin. Due to the heavy penalty factor $C$ for violating the margin there are relatively few such points, and thus the slack vector $\xi$ is sparse — most of its entries are 0. We are interested in explaining why this phenomenon occurs in this specific optimization problem, and identifying the key properties that determine sparse solutions for arbitrary optimization problems.

To reason about the SVM case, let's see how changing some arbitrary slack variable $\xi_i$ affects the loss. A unit decrease in $\xi_i$ results in a “reward” of $C$, and is captured by the partial derivative $\frac{\partial L}{\partial \xi_i}$. Note that no matter what the current value of $\xi_i$ is, the reward for decreasing $\xi_i$ is constant. Of course, decreasing $\xi_i$ may change the boundary and thus the cost attributed to the size of the margin $\|w\|^2$. The overall reward for decreasing $\xi_i$ is either going to be worth the effort (greater than cost incurred from $w$) or not worth the effort (less than cost incurred from $w$). Intuitively, $\xi_i$ will continue to decrease until it hits a lower-bound “equilibrium” — which is often just 0.

Now consider the following formulation (constraints omitted for brevity again):

$$\min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i^2$$

The reward for decreasing $\xi_i$ is no longer constant — at any point, a unit decrease in $\xi_i$ results in a “reward” of $2C\xi_i$. As $\xi_i$ approaches 0, the rewards get smaller and smaller, reaching infinitesimal
values. On the other hand, decreasing $\xi_i$ causes a finite increase in the cost incurred by the $\|w\|_2^2$ — the same increase in cost as in the previous example. Intuitively, we can reason that there will be a threshold value $\xi_i^*$ such that decreasing $\xi_i$ further will no longer outweigh the cost incurred by the size of the margin, and that the $\xi_i$’s will halt their descent before they hit zero.

The same reasoning applies to least squares as well. For any particular component $w_i$ of $w$, the corresponding loss in LASSO is the absolute value $|w_i|$, while the loss in ridge regression is the squared term $w_i^2$. In the case of LASSO the “reward” for decreasing $w_i$ by a unit amount is a constant $\lambda$, while for ridge regression the equivalent “reward” is $2\lambda w_i$, which depends on the value of $w_i$. There is a compelling geometric argument behind this reasoning as well.

Suppose for simplicity that we are only working with 2-dimensional data points and are thus optimizing over two weight variables $w_1$ and $w_2$. In both figures above, the red ellipses represent isocontours in $w$-space of the squared loss $\|Xw - y\|_2^2$. In ridge regression, each isocontour of $\lambda\|w\|_2^2$ is represented by a circle, one of which is shown in the right figure. Note that the optimal $w$ will only occur at points of tangency between the red ellipse and the blue circle. Otherwise we could always move along the isocontour of one of the functions (keeping its overall cost fixed) while improving the value of the the other function, thereby improving the overall value of the loss function. We can’t really infer much about these points of tangency other than the fact that the blue circle centered at the origin draws the optimal point closer to the origin (ridge regression penalizes large weights).

Now, let’s examine the LASSO case. The red ellipses represent the same objective $\|Xw - y\|_2^2$, but now the $\ell_1$ regularization term $\lambda\|w\|_1$ is represented by diamond isocontours. As with ridge regression, note that the optimal point in $w$-space must occur at points of tangency between the ellipse and the diamond. Due to the “pointy” property of the diamonds, tangency is very likely to happen at the corners of the diamond because they are single points from which the rest of the diamond draws away from. And what are the corners of the diamond? Why, they are points at which one component of $w$ is 0!

![Figure 2.2: Comparing contour plots for LASSO (left) vs. ridge regression (right).](image-url)
Solving LASSO

Convinced that LASSO achieves sparsity, now let’s find the optimal solution to LASSO. Unlike ridge regression, it is not exactly clear what the closed form solution is through linear algebra or gradient methods, since the objective function not differentiable (due to the “pointiness” of the $\ell_1$ norm). Specifically, LASSO zeros out features, and once these weights are set to 0 the objective function becomes non-differentiable. Note however, that the objective is still convex, and we could use an iterative method such as subgradient descent or line search to solve the problem. Here, we will use a line search method called coordinate descent.

While SGD focuses on iteratively optimizing the value of the objective $L(w)$ for each sample in the training set, coordinate descent iteratively optimizes the value of the objective for each feature.

Algorithm 1: Coordinate Descent

```
while w has not converged do
  pick a feature index $i$
  update $w_i$ to arg min $w_i L(w)$
```

Coordinate descent is guaranteed to find the global minimum if $L$ is jointly convex. No such guarantees can be made however if $L$ is only elementwise convex, since it may have local minima. To understand why, let’s start by understanding elementwise vs joint convexity. Suppose we are trying to minimize $f(x,y)$, a function of two scalar variables $x$ and $y$. For simplicity, assume that $f$ is twice differentiable, so we can take its Hessian. $f(x,y)$ is element-wise convex in $x$ if its Hessian is psd when $y$ is fixed:

$$\frac{\partial^2}{\partial x \partial x} f(x,y) \geq 0$$

Same goes for element-wise convexity in $y$.

$f(x,y)$ is jointly convex in $x$ and $y$ if its Hessian $\nabla^2 f(x,y)$ is psd. Note that being element-wise convex in both $x$ and $y$ does not imply joint convexity in $x$ and $y$ (consider $f(x,y) = x^2 + y^2 - 4xy$ as an example). However, being joint convexity in $x$ and $y$ does imply being element-wise convex in both $x$ and $y$.

Now, if $f(x,y)$ was jointly convex, then we could find the gradient wrt. $x$ and $y$ individually, set them to 0, and be guaranteed that would be the global minimum. Can we do this if $f(x,y)$ is element-wise convex in both $x$ and $y$? Even though it is true that $\min_{x,y} f(x,y) = \min_x \min_y f(x,y)$, we can’t always just set gradients to 0 if $f(x,y)$ is not jointly convex. While the inner optimization problem over $y$ is convex, the outer optimization problem over $x$ may no longer be convex. In the case when joint convexity is not reached, there is no clean strategy to find global minimum and we must analyze all of the critical points to find the minimum.

In the case of LASSO, the objective function is jointly convex, so we can use coordinate descent. There are a few details to be filled in, namely the choice of which feature to update and how $w_i$ is updated. One simple way is to just pick a random feature $i$ each iteration. After choosing the feature, we have to update $w_i \leftarrow \arg \min_{w_i} L(w)$. For LASSO, it turns out there is a closed-form solution (note that we are only minimizing with respect to one feature instead of all the features).

Let’s solve the line search problem $\min_{w_i} L(w)$. For convenience, let’s separate the terms that depend on $w_i$ from those that don’t. Denoting $x_j$ as the $j$-th column of $X$, we have

$$L(w) = \|Xw - y\|^2 + \lambda \|w\|_1$$

CHAPTER 2. REGRESSION II

\[ \sum_{j=1}^{d} w_j x_j - y \]
\[ = \left\| w_i x_i + r \right\|_2^2 + \lambda |w_i| + C \]

where \( r = \sum_{j \neq i} w_j x_j - y \) and \( C = \lambda \sum_{j \neq i} |w_j| \). The objective can in turn be written as

\[ L(w) = \lambda |w_i| + C + \sum_{j=1}^{n} (w_i x_{ji} + r_j)^2 \]

where \( x_{ji} \) is the i'th feature of the j'th datapoint.

Suppose that the optimal \( w_i^* \) is strictly positive: \( w_i^* > 0 \). Setting the partial derivative of the objective wrt. \( w_i \) (the partial derivative in this case is defined since \( w_i^* \neq 0 \)) to 0, we obtain

\[ \frac{\partial L}{\partial w_i} = \lambda + \sum_{j=1}^{n} 2x_{ji}(w_i x_{ji} + r_j) = 0 \implies w_i^* = \frac{-\lambda - \sum_{j=1}^{n} 2x_{ji}r_j}{\sum_{j=1}^{n} 2x_{ji}^2} \]

Denoting \( a = -\sum_{j=1}^{n} 2x_{ji}r_j \) and \( b = \sum_{j=1}^{n} 2x_{ji}^2 \), we have

\[ w_i^* = \frac{-\lambda + a}{b} \]

But this only holds if the right hand side, \( \frac{-\lambda + a}{b} \), is actually positive. If it is negative or 0, then this means there is no optimum in \((0, \infty)\).

When \( w_i^* < 0 \), then similar calculations will lead to

\[ w_i^* = \frac{\lambda + a}{b} \]

Again, this only holds if \( \frac{\lambda + a}{b} \) is actually negative. If it is positive or 0, then there is no optimum in \((-\infty, 0)\).

If neither the conditions \( \frac{-\lambda + a}{b} > 0 \) or \( \frac{\lambda + a}{b} < 0 \) hold, then there is no optimum in \((-\infty, 0) \) or \( (0, \infty) \).

But the LASSO objective is convex in \( w_i \) and has an optimum somewhere, thus in this case \( w_i^* = 0 \). In order for this case to hold, we must have that \( \frac{-\lambda + a}{b} \leq 0 \) and \( \frac{\lambda + a}{b} \geq 0 \). Rearranging, we can see this is equivalent to \( |a| \leq \lambda \).

Examine each of the following cases:

- \( -\frac{\lambda + a}{b} \leq 0 \) and \( \frac{\lambda + a}{b} \geq 0 \): \( w_i^* = 0 \)
- \( -\frac{\lambda + a}{b} \leq 0 \) and \( \frac{\lambda + a}{b} < 0 \): \( w_i^* < 0 \)
- \( -\frac{\lambda + a}{b} > 0 \) and \( \frac{\lambda + a}{b} \geq 0 \): \( w_i^* > 0 \)
- \( -\frac{\lambda + a}{b} > 0 \) and \( \frac{\lambda + a}{b} < 0 \): impossible since this implies \( \frac{-\lambda + a}{b} > \frac{\lambda + a}{b} \) and \( \lambda \) and \( b \) are non-negative

The cases above imply the optimal solution \( w_i^* \):

\[ w_i^* = \begin{cases} 0 & \text{if } |a| \leq \lambda \\ \frac{-\lambda + a}{b} & \text{if } -\frac{\lambda + a}{b} > 0 \\ \frac{\lambda + a}{b} & \text{if } -\frac{\lambda + a}{b} < 0 \end{cases} \]
2.6. SPARSE LEAST SQUARES

where
\[ a = -\sum_{j=1}^{n} 2x_{ji}r_j, \quad b = \sum_{j=1}^{n} 2x_{ji}^2 \]

This is not a gradient-descent update — we have a closed-form solution for the optimum \( w_i \), given that all of the other weights are fixed constants. We can see explicitly how the LASSO objective induces sparsity — \( a \) is some function of the data and the other weights, and when \( |a| \leq \lambda \), we set \( w_i = 0 \) in this iteration of coordinate descent. By increasing \( \lambda \), we increase the threshold of \( |a| \) for \( w_i \) to be set to 0, and our solution becomes more sparse. Also note that the term \( \frac{a}{b} \) is the least squares solution (without regularization), so we can see that the regularization term tries to pull the least squares update towards 0.

One subtle point: during coordinate descent, weights can be “reactivated” after having been set to 0 in a previous iteration, since \( a \) is affected by factors other than \( w_i \).

Matching Pursuit

Rather than relaxing the \( \ell_0 \) constraint (as seen in LASSO), the matching pursuit algorithm keeps the constraint, and instead finds an approximate solution to the sparse least squares problem in a greedy fashion. The algorithm starts with with a completely sparse solution \( (w^0 = 0) \), and iteratively updates \( w \) until the sparsity constraint \( \|w\|_0 \leq k \) can no longer be met. At iteration \( t \), the algorithm can only update one entry of \( w^{t-1} \), and it chooses the feature that minimizes the (squared) norm of the resulting residual \( \|r^t\|^2 = \|y - Xw^t\|^2 \).

**Algorithm 2: Matching Pursuit**

initialize the weights \( w^0 = 0 \) and the residual \( r^0 = y - Xw^0 = y \)

while \( \|w\|_0 < k \) do

  find the feature \( i \) for which the length of the projected residual onto \( x_i \) is minimized:
  \[ i = \arg \min_j \left( \min_{\nu} \|r^{t-1} - \nu x_j\| \right) = \arg \max_j \frac{\langle r^{t-1}, x_j \rangle}{\|x_j\|} \]

  update the \( i \)'th feature entry of the weight vector:
  \[ w_i^t = w_i^{t-1} - \frac{\langle r^{t-1}, x_i \rangle}{\|x_i\|^2} \]

  update the residual vector: \( r^t = y - Xw^t \)

At iteration \( t \), we pick the coordinate \( i \) such that the projection of the residual \( r^{t-1} \) onto \( x_i \) (the \( i \)'th column of \( X \) corresponding to feature \( i \), not datapoint \( i \)) is minimized:

\[ i = \arg \min_j \left( \min_{\nu} \|r^{t-1} - \nu x_j\| \right) \]

This equates to finding the index \( i \) for which the length of the projection onto \( x_i \) is maximized:

\[ i = \arg \max_j \frac{\langle r^{t-1}, x_j \rangle}{\|x_j\|} \]

Let’s see why this is true. The inner optimization problem \( \min_{\nu} \|r^{t-1} - \nu x_i\| \) is simply a projection problem, and its solution is

\[ \nu^* = \frac{\langle r^{t-1}, x_j \rangle}{\langle x_j, x_j \rangle} \]
which gives a value of
\[ \left\| r^{t-1} - \frac{\langle r^{t-1}, x_j \rangle}{\langle x_j, x_j \rangle} x_i \right\| \]

The outer optimization problem selects the feature that minimizes this quantity:
\[
i = \arg \min_j \left\| r^{t-1} - \frac{\langle r^{t-1}, x_j \rangle}{\langle x_j, x_j \rangle} x_i \right\|
\]
\[
= \arg \min_j \left\| r^{t-1} - \frac{\langle r^{t-1}, x_j \rangle}{\langle x_j, x_j \rangle} x_i \right\|^2
\]
\[
= \arg \min_j \| r^{t-1} \|^2 + \frac{(r^{t-1}, x_j)^2}{\langle x_j, x_j \rangle} - 2 \frac{(r^{t-1}, x_j)^2}{\langle x_j, x_j \rangle}
\]
\[
= \arg \min_j \| r^{t-1} \|^2 - \frac{(r^{t-1}, x_j)^2}{\langle x_j, x_j \rangle}
\]
\[
= \arg \max_j \frac{(r^{t-1}, x_j)^2}{\langle x_j, x_j \rangle}
\]
\[
= \arg \max_j \frac{|(r^{t-1}, x_j)|}{\|x_j\|}
\]

Now that we have found the feature \( i \) that minimizes the projection distance, we must update corresponding weight \( i \) and the residual vector. The updated residual \( r^t \) is the result of projecting \( r^{t-1} \) onto feature \( x_i \):

\[
Figure 2.3: The updated residual \( r^t \), current residual \( r^{t-1} \), and scaled feature \( x_i \) form a right triangle.
\]

The updated residual is given by
\[
r^t = r^{t-1} - \frac{(r^{t-1}, x_i)}{\|x_i\|^2} x_i
\]
which corresponds to subtracting \( \frac{(r^{t-1}, x_i)}{\|x_i\|^2} \) from the corresponding weight \( w_i^t \):
\[
w_i^t = w_i^{t-1} - \frac{(r^{t-1}, x_i)}{\|x_i\|^2}
\]
We update \( w_i \) to the optimum projection value and repeat greedily at each iteration. At each iteration, the (squared) length of the residual \( \|Xw^t - y\|^2 \) monotonically decreases since \( \|r^{t-1}\|^2 > \|r^t\|^2 \). While matching pursuit is not guaranteed to find the optimal \( w^* \), in practice it works well for most applications.
Orthogonal Matching Pursuit

The Orthogonal Matching Pursuit (OMP) algorithm is an extension to the standard Matching Pursuit algorithm with the following difference: at iteration $t$, we maintain a set $I^t$ of all features selected by the algorithm so far, and instead of updating just the one weight corresponding to feature $i$ found at iteration $t$, we update all weights corresponding to the features in $I^t$ using Least Squares.

Algorithm 3: Orthogonal Matching Pursuit

1. Initialize the weights $w^0 = 0$ and the residual $r^0 = y - Xw^0 = y$
2. Initialize a set of features $I^0 = \emptyset$
3. While $\|w\|_0 < k$
   a. Find the feature $i$ for which the length of the projected residual onto $x_i$ is minimized:
   $$i = \arg\min_j (\min \nu \|r^{t-1} - \nu x_j\|) = \arg\max_j \frac{|\langle r^{t-1}, x_j \rangle|}{\|x_j\|}$$
   b. Add feature $i$ to the set of features:
   $$I^t = I^{t-1} \cup \{i\}$$
   c. Estimate the best linear fit of the target $y$ using the features obtained so far. Given that we have found $t$ good features, we now find the best linear fit for the target $y$ using these $t$-features. Define $X_t = [x_i, \ldots, x_t]$ made up of these $t$-features. Then we determine $w^t$ as the solution for the following least-squares problem:
   $$w^t = \arg\min_{w \in \mathbb{R}^t} \|y - X_t w\|_2^2$$
   d. Update the residual vector:
   $$r^t = y - X w^t$$

The motivation for OMP is as follows: if we do not refit on all features updated so far after choosing a new feature, the new residual will not necessarily be orthogonal to the span of the canonical basis vectors corresponding to those chosen coordinates, and is therefore not optimal. The OMP algorithm ensures that $w^t$ corresponds to an optimal Least Squares solution if we restricted our features to just those in $I^t$.

2.7 Total Least Squares

Previously, we have covered Ordinary Least Squares (OLS) which assumes that the dependent variable $y$ is noisy but the independent variables $x$ are noise-free. We now discuss Total Least Squares (TLS), where we assume that our independent variables are also corrupted by noise. For this reason, TLS is considered an errors-in-variables model.

A probabilistic motivation?

We might begin with a probabilistic formulation and fit the parameters via maximum likelihood estimation, as before. Consider for simplicity a one-dimensional linear model

$$y_{true} = wx_{true}$$
where the observations we receive are corrupted by Gaussian noise

\[(x, y) = (x_{\text{true}} + z_x, y_{\text{true}} + z_y) \quad z_x, z_y \sim \mathcal{N}(0, 1)\]

Combining the previous two relations, we obtain

\[y = y_{\text{true}} + z_y = wx_{\text{true}} + z_y = w(x - z_x) + z_y = wx - wz_x + z_y \sim \mathcal{N}(0, w^2 + 1)\]

The likelihood for a single point is then given by

\[P(x, y; w) = \frac{1}{\sqrt{2\pi(w^2 + 1)}} \exp \left( -\frac{1}{2} \frac{(y - wx)^2}{w^2 + 1} \right)\]

Thus the log likelihood is

\[\log P(x, y; a) = \text{constant} - \frac{1}{2} \log \left( w^2 + 1 \right) - \frac{1}{2} \frac{(y - wx)^2}{w^2 + 1}\]

Observe that the parameter \(w\) shows up in three places, unlike the form that we are familiar with, where it only appears in the quadratic term. Our usual strategy of setting the derivative equal to zero to find a maximizer will not yield a nice system of linear equations in this case, so we'll try a different approach.

**Low-rank formulation**

To solve the TLS problem, we develop another formulation that can be solved using the singular value decomposition. To motivate this formulation, recall that in OLS we attempt to minimize \(\|Xw - y\|_2^2\), which is equivalent to

\[\min_{w, \epsilon} \|\epsilon\|_2^2 \quad \text{subject to} \quad y = Xw + \epsilon\]

This only accounts for errors in the dependent variable, so for TLS we introduce a second residual \(\epsilon_X \in \mathbb{R}^{n \times d}\) to account for independent variable error:

\[\min_{w, \epsilon_X, \epsilon_y} \left\| \begin{bmatrix} \epsilon_X & \epsilon_y \end{bmatrix} \right\|_F^2 \quad \text{subject to} \quad (X + \epsilon_X)w = y + \epsilon_y\]

For comparison to the OLS case, note that the Frobenius norm is essentially the same as the 2-norm, just applied to the elements of a matrix rather than a vector.

From a probabilistic perspective, finding the most likely value of a Gaussian corresponds to minimizing the squared distance from the mean. Since we assume the noise is 0-centered, we want to minimize the sum of squares of each entry in the error matrix, which corresponds exactly to minimizing the Frobenius norm.

In order to separate out the terms being minimized, we rearrange the constraint equation as

\[\left[ \begin{bmatrix} X + \epsilon_X & y + \epsilon_y \end{bmatrix} \right] w = 0\]
This expression tells us that the vector $\begin{bmatrix} w^T & -1 \end{bmatrix}^T$ lies in the nullspace of the matrix on the left. However, if the matrix is full rank, its nullspace contains only $0$, and thus the equation cannot be satisfied (since the last component, $-1$, is always nonzero). Therefore we must choose the perturbations $\epsilon_X$ and $\epsilon_y$ in such a way that the matrix is not full rank.

It turns out that there is a mathematical result, the Eckart-Young theorem, that can help us pick these perturbations. This theorem essentially says that the best low-rank approximation (in terms of the Frobenius norm$^2$) is obtained by throwing away the smallest singular values.

**Theorem.** Suppose $A \in \mathbb{R}^{m \times n}$ has rank $r \leq \min(m, n)$, and let $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ be its singular value decomposition. Then

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T = U \begin{bmatrix} \sigma_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \sigma_k & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} V^T$$

where $k \leq r$, is the best rank-$k$ approximation to $A$ in the sense that

$$\|A - A_k\|_F \leq \|A - \tilde{A}\|_F$$

for any $\tilde{A}$ such that $\text{rank}(\tilde{A}) \leq k$.

Let us assume that the data matrix $\begin{bmatrix} X & y \end{bmatrix}$ is full rank.$^3$ Write its singular value decomposition:

$$\begin{bmatrix} X & y \end{bmatrix} = \sum_{i=1}^{d+1} \sigma_i u_i v_i^T$$

Then the Eckart-Young theorem tells us that the best rank-$d$ approximation to this matrix is

$$\begin{bmatrix} X + \epsilon_X & y + \epsilon_y \end{bmatrix} = \sum_{i=1}^{d} \sigma_i u_i v_i^T$$

which is achieved by setting

$$\begin{bmatrix} \epsilon_X & \epsilon_y \end{bmatrix} = -\sigma_{d+1} u_{d+1} v_{d+1}^T$$

The nullspace of our resulting matrix is then

$$\text{null} \left( \begin{bmatrix} X + \epsilon_X & y + \epsilon_y \end{bmatrix} \right) = \text{null} \left( \sum_{i=1}^{d} \sigma_i u_i v_i^T \right) = \text{span} \{ v_{d+1} \}$$

where the last equality holds because $\{ v_1, \ldots, v_{d+1} \}$ form an orthogonal basis for $\mathbb{R}^{d+1}$. To get the weight $w$, we find a scaling $\alpha$ such that $\begin{bmatrix} w^T & -1 \end{bmatrix}^T$ is in the nullspace, i.e.

$$\begin{bmatrix} w \\ -1 \end{bmatrix} = \alpha v_{d+1}$$

Note that this requires the $(d + 1)$st component of $v_{d+1}$ to be nonzero. (See Section for details.)

$^2$ There is a more general version that holds for any unitary invariant norm.

$^3$ This should be the case in practice because the noise will cause $y$ not to lie in the columnspace of $X$. 
Noise, regularization, and reverse-regularization

In a sense, above we have solved the problem of total least squares by reducing it to computing an appropriate SVD. Once we have \( v_{d+1} \), or any scalar multiple of it, we simply rescale it so that the last component is \(-1\), and then the first \( d \) components give us \( w \). However, we can look at this more closely to uncover the relationship between TLS and the ideas of regularization that we have seen earlier in the course.

Since \( v_{d+1} \) is a right-singular vector of \( [X \ y] \), it is an eigenvector of the matrix

\[
[X \ y] \begin{bmatrix} X & y \\ y^\top & y^\top \end{bmatrix} = \begin{bmatrix} X^\top X & X^\top y \\ y^\top X & y^\top y \end{bmatrix}
\]

So to find it we solve

\[
\begin{bmatrix} X^\top X & X^\top y \\ y^\top X & y^\top y \end{bmatrix} \begin{bmatrix} w \\ -1 \end{bmatrix} = \sigma_{d+1}^2 \begin{bmatrix} w \\ -1 \end{bmatrix}
\]

From the top line we see that \( w \) satisfies

\[
X^\top Xw - X^\top y = \sigma_{d+1}^2 w
\]

which can be rewritten as

\[
(X^\top X - \sigma_{d+1}^2 I)w = X^\top y
\]

Thus, assuming \( X^\top X - \sigma_{d+1}^2 I \) is invertible (see the next section), we can solve for the weights as

\[
\hat{w}_{TLS} = (X^\top X - \sigma_{d+1}^2 I)^{-1} X^\top y
\]

This result is like ridge regression, but with a negative regularization constant!

Why does this make sense? One of the original motivations of ridge regression was to ensure that the matrix being inverted is in fact nonsingular, and subtracting a scalar multiple of the identity seems like a step in the opposite direction. We can make sense of this by recalling our original model:

\[
X = X_{true} + Z
\]

where \( X_{true} \) are the actual values before noise corruption, and \( Z \) is a zero-mean noise term with i.i.d. entries. Then

\[
E[X^\top X] = E[(X_{true} + Z)^\top (X_{true} + Z)]
\]

\[
= E[X_{true}^\top X_{true}] + E[X_{true}^\top Z] + E[Z^\top X_{true}] + E[Z^\top Z]
\]

\[
= X_{true}^\top X_{true} + X_{true}^\top E[Z] + E[Z^\top] X_{true} + E[Z^\top Z]
\]

\[
= X_{true}^\top X_{true} + E[Z^\top Z]
\]

Observe that the off-diagonal terms of \( E[Z^\top Z] \) terms are zero because the \( i \)th and \( j \)th rows of \( Z \) are independent for \( i \neq j \), and the on-diagonal terms are essentially variances. Thus the \(-\sigma_{d+1}^2 I\) term is there to compensate for the extra noise introduced by our assumptions regarding the independent variables.

For another perspective, note that

\[
E[X^\top] = E[(X_{true} + Z)^\top] = E[X_{true}^\top + Z^\top] = E[X_{true}^\top] + E[Z^\top] = X_{true}^\top
\]
If we plug this into the OLS solution (where we have assumed no noise in the independent variables), we see

\[ \hat{w}_{\text{OLS}} = (X_{\text{true}}^\top X_{\text{true}})^{-1}X_{\text{true}}^\top y = (E[X^\top X] - E[Z^\top Z])^{-1}E[X^\top y] \]

which strongly resembles the TLS solution, but expressed in terms of expectations over the noise \( Z \).

So, is this all just a mathematical trick or is there a practical sense in which ridge regularization itself is related to adding noise? The math above suggests that we can take the original training data set and instead of working with that data set, just sample lots of points (say \( r \) times each) with i.i.d. zero-mean Gaussian noise with variance \( \lambda \) added to each of their features. Call this the \( X \) and have the corresponding \( y \) just keep the original \( y \) values. Then, doing ordinary least squares on this noisily degraded data set will end up behaving like ridge regression since the laws of large numbers will make \( \frac{1}{r}X^\top X \) concentrate around \( X_{\text{true}}^\top X_{\text{true}} + \lambda I \). Meanwhile \( X^\top y \) will concentrate to \( rX_{\text{true}}^\top y_{\text{orig}} \) with \( O(\sqrt{r}) \) noise on top of this by the Central Limit Theorem (if we used other-than-Gaussian noise to noisily resample), and straight variance-\( O(r) \) Gaussian noise if we indeed used Gaussian noise. Putting them together means that the result of OLS with noisily augmented training data will result in approximately the same solution as ridge-regression, with the solutions approaching each other as the number of noisy copies \( r \) goes to infinity.

Why does this make intuitive sense? How can adding noise make learning more reliable? The intuitive reason is that this added noise destroys inadvertent conspiracies. Overfitting happens because the learning algorithm sees some degree of conspiracies between the observed training labels \( y \) and the input features. By adding lots of copies of the training data with additional noise added into them, many of these conspiracies will be masked by the added noise because they are fundamentally sensitive to small details — this is why they manifest as large weights \( w \). We know from our studies of the bias/variance tradeoff that having more training samples reduces this variance. Adding our own noisy samples exploits this variance reduction.

In many practical machine learning situations, appropriately adding noise to your training data can be an important tool in helping generalization performance.

Existence of the solution

In the discussion above, we have in some places made assumptions to move the derivation forward. These do not always hold, but we can provide sufficient conditions for the existence of a solution.

**Proposition.** Let \( \sigma_1, \ldots, \sigma_{d+1} \) denote the singular values of \([X \ y]\), and \( \tilde{\sigma}_1, \ldots, \tilde{\sigma}_d \) denote the singular values of \( X \). If \( \sigma_{d+1} < \tilde{\sigma}_d \), then the total least squares problem has a solution, given by

\[ \hat{w}_{\text{TLS}} = (X^\top X - \sigma^2_{d+1} I)^{-1}X^\top y \]

**Proof.** Let \( \sum_{i=1}^{d+1} \sigma_i u_i v_i^\top \) be the SVD of \([X \ y]\), and suppose \( \sigma_{d+1} < \tilde{\sigma}_d \). We first show that the \((d+1)\)st component of \( v_{d+1} \) is nonzero. To this end, suppose towards a contradiction that \( v_{d+1} = [a^\top \ 0]^\top \) for some \( a \neq 0 \). Since \( v_{d+1} \) is a right-singular vector of \([X \ y]\), i.e. an eigenvector of \([X \ y]^\top [X \ y]\), we have

\[ [X \ y]^\top [X \ y] \begin{bmatrix} a \\ 0 \end{bmatrix} = \begin{bmatrix} X^\top X & X^\top y \\ y^\top X & y^\top y \end{bmatrix} \begin{bmatrix} a \\ 0 \end{bmatrix} = \sigma^2_{d+1} \begin{bmatrix} a \\ 0 \end{bmatrix} \]

Then

\[ X^\top X a = \sigma^2_{d+1} a \]
i.e. \( \mathbf{a} \) is an eigenvector of \( \mathbf{X}^\top \mathbf{X} \) with eigenvalue \( \sigma_{d+1}^2 \). However, this contradicts the fact that

\[
\tilde{\sigma}_d^2 = \lambda_{\text{min}}(\mathbf{X}^\top \mathbf{X})
\]

since we have assumed \( \sigma_{d+1} < \tilde{\sigma}_d \). Therefore the \((d + 1)\)st component of \( \mathbf{v}_{d+1} \) is nonzero, which guarantees the existence of a solution.

We have already derived the given expression for \( \hat{\mathbf{w}}_{\text{TLS}} \), but it remains to show that the matrix \( \mathbf{X}^\top \mathbf{X} - \sigma_{d+1}^2 \mathbf{I} \) is invertible. This is fairly immediate from the assumption that \( \sigma_{d+1} < \tilde{\sigma}_d \), since this implies

\[
\sigma_{d+1}^2 < \tilde{\sigma}_d^2 = \lambda_{\text{min}}(\mathbf{X}^\top \mathbf{X})
\]

giving

\[
\lambda_{\text{min}}(\mathbf{X}^\top \mathbf{X} - \sigma_{d+1}^2 \mathbf{I}) = \lambda_{\text{min}}(\mathbf{X}^\top \mathbf{X}) - \sigma_{d+1}^2 > 0
\]

which guarantees that the matrix is invertible.

This gives us a nice mathematical characterization of the existence of a solution, showing that the two technical requirements we raised earlier (the last entry of \( \mathbf{v}_{d+1} \) being nonzero, and the matrix \( \mathbf{X}^\top \mathbf{X} - \sigma_{d+1}^2 \mathbf{I} \) being invertible) happen together. However, is the assumption of the proof likely to hold in practice? We give an intuitive argument that it is.

Consider that in solving the TLS problem, we have determined the error term \( \mathbf{e}_X \). In principle, we could use this to denoise \( \mathbf{X} \), as in \( \hat{\mathbf{X}}_{\text{true}} = \mathbf{X} - \mathbf{e}_X \), and then perform OLS as normal. This process is essentially the same as TLS if we compare the original formulations. Assuming the error is drawn from a continuous distribution, the probability that the denoised matrix \( \hat{\mathbf{X}}_{\text{true}} \) has collinear columns is zero.

**TLS minimizes perpendicular distance**

Recall that OLS tries to minimize the vertical distance between the fitted line and data points. TLS, on the other hand, tries to minimize the perpendicular distance. For this reason, TLS may sometimes be referred to as **orthogonal regression**.

The red lines represent vertical distance, which OLS aims to minimize. The blue lines represent perpendicular distance, which TLS aims to minimize. Note that all blue lines are perpendicular to the black line (hypothesis model), while all red lines are perpendicular to the \( x \) axis.
Chapter 3

Dimensionality Reduction

3.1 Dimensionality reduction and Principal Component Analysis

In machine learning, the data we have are often very high-dimensional. In fact, when we introduced the idea of features (like polynomial features), these made the dimensionality of the data even higher. The kernel trick was something that let us partially deal with this by working with vectors only as long as there are training samples.

However, there are a number of reasons why we might want to work with a lower-dimensional representation:

- Visualization (if we can get it down to 2 or 3 dimensions), e.g. for exploratory data analysis
- Reduce computational load
- Reduce variance in estimation — regularize the problem

So, how can we reduce the dimensionality of data? There are obvious ways — just keeping a subset of features. But which features? In general, that presumably depends on what you are trying to do. What could you do if you didn’t know what you were trying to predict with those features? This corresponds to unsupervised dimensionality reduction. There are a couple of intuitive choices. First, just pick some features at random to keep. This is appealing for its symmetry, but it makes you wonder if we could do better by actually looking at the data before deciding which features to keep.

Consequently, another thing that you could do is to just keep the few features that have the most variability — which you could measure by the variance of that feature. But what if two of the most variable features were actually very correlated to each other? Should we really be including both of them? Maybe we should focus on “fresh” variability somehow. To do this, maybe it would be helpful to allow ourselves to synthesize linear combinations of features and keep some of these synthesized features.

Principal Component Analysis (PCA) is exactly such an unsupervised dimensionality reduction technique. Given a matrix of data points, it finds one or more orthogonal directions that capture the largest amount of variance in the data. Intuitively, the directions with less variance contain less information and may be discarded without introducing too much error. One of the practical motivations for taking this kind of unsupervised approach to dimensionality reduction is that labeled training data might be hard or expensive to get, but unlabeled training data (i.e. no
CHAPTER 3. DIMENSIONALITY REDUCTION

just \( x \) might be more easily available. PCA is able to extract meaningful directions from such unlabeled data.

Not coincidentally, PCA turns out to be intimately connected to the ideas of Total Least Squares.

**Projection**

Let us first review the meaning of scalar projection of one vector onto another. If \( v \in \mathbb{R}^d \) is a unit vector, i.e. \( \|v\| = 1 \), then the scalar projection of another vector \( x \in \mathbb{R}^d \) onto \( v \) is given by \( x^\top v \). This quantity tells us roughly how much of the projected vector \( x \) lies along the direction given direction \( v \). Why does this expression make sense? Recall the slightly more general formula which holds for vectors of any length:

\[
x^\top v = \|x\| \|v\| \cos \theta
\]

where \( \theta \) is the angle between the vectors. In this case, since \( \|v\| = 1 \), the expression simplifies to \( x^\top v = \|x\| \cos \theta \). But since cosine gives the ratio of the adjacent side (the projection we want to find) to the hypotenuse (\( \|x\| \)), this is exactly what we want:

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i^\top v)^2 = \frac{1}{n} ||Xv||^2 = \frac{1}{n} v^\top X^\top X v
\]

One approach to dimensionality reduction by using projections is to choose projections at random — sample from an iid Gaussian and then normalize the vector to get our \( v \). This creates a degree of fairness across the individual features since an iid Gaussian is uniform over directions in \( d \)-dimensional space. As you will see in homework, this approach to dimensionality reduction actually has many interesting properties. By construction, however, it does not look at any data itself and thus is unable to prioritize important vs unimportant feature directions.

**The first principal component**

Let \( X \in \mathbb{R}^{n \times d} \) be our matrix of data, where each row is a \( d \)-dimensional datapoint. These are to be thought of as i.i.d. samples from some random vector \( x \).

We will assume that the data points have mean zero; if this is not the case, we can make it so by subtracting the average of all the rows, \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \), from each row. The motivation for this is that we want to find directions of high variance within the data, and variance is defined relative to the mean of the data. If we did not zero-center the data, the directions found would be heavily influenced by where the data lie relative to the origin, rather than where they lie relative to the other data, which is more useful.

Since \( X \) is zero-mean, the sample variance of the datapoints’ projections onto a unit vector \( v \) is given by

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i^\top v)^2 = \frac{1}{n} ||Xv||^2 = \frac{1}{n} v^\top X^\top X v
\]
3.1. DIMENSIONALITY REDUCTION AND PRINCIPAL COMPONENT ANALYSIS

where $v$ is constrained to have unit norm.\(^1\)

With this motivation, we define the **first loading vector** $v_1$ as the solution to the constrained optimization problem

$$
\max_v v^\top X^\top X v \quad \text{subject to} \quad v^\top v = 1
$$

Note that we have discarded the positive constant factor $1/n$ which does not affect the optimal value of $v$.

To reduce this constrained optimization problem to an unconstrained one, we write down its Lagrangian:

$$
\mathcal{L}(v) = v^\top X^\top X v - \lambda(v^\top v - 1)
$$

First-order necessary conditions for optima imply that

$$
0 = \nabla \mathcal{L}(v_1) = 2X^\top X v_1 - 2\lambda v_1
$$

Hence $X^\top X v_1 = \lambda v_1$, i.e. $v_1$ is an eigenvector of $X^\top X$ with eigenvalue $\lambda$. Since we constrain $v_1^\top v_1 = 1$, the value of the objective is precisely

$$
v_1^\top X^\top X v_1 = v_1^\top (\lambda v_1) = \lambda v_1^\top v_1 = \lambda
$$

so the optimal value is $\lambda = \lambda_{\text{max}}(X^\top X)$, which is achieved when $v_1$ is a unit eigenvector of $X^\top X$ corresponding to its largest eigenvalue.

**Finding more principal components**

We have seen how to find the first loading vector, which is the unit vector that maximizes the variance of the projected data points. However, in most applications, we want to find more than one direction. We want the subsequent directions found to also be directions of high variance, but they ought to be orthogonal to the existing directions in order to minimize redundancy in the information captured. Thus we define the $k$th loading vector $v_k$ as the solution to the constrained optimization problem

$$
\max_v v^\top X^\top X v \quad \text{subject to} \quad v^\top v = 1
$$

$$
v^\top v_i = 0, \quad i = 1, \ldots, k - 1
$$

We claim that $v_k$ is a unit eigenvector of $X^\top X$ corresponding to its $k$th largest eigenvalue.

**Proof.** By induction on $k$. We have already shown that the claim is true for the base case $k = 1$ (where there are no orthogonality constraints). Now assume that it is true for the first $k$ loading vectors $v_1, \ldots, v_k$, and consider the problem of finding $v_{k+1}$.

\(^1\) To make sense of the sample variance, recall that for any random variable $Z$,

$$
\text{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}[Z])^2]
$$

so if $\mathbb{E}[Z] = 0$ then $\text{Var}(Z) = \mathbb{E}[Z^2]$. In practice we will not have the true random variable $Z$, but rather i.i.d. observations $z_1, \ldots, z_n$ of $Z$. The expected value can then be approximated by a sample average, i.e.

$$
\mathbb{E}[Z^2] \approx \frac{1}{n} \sum_{i=1}^{n} z_i^2
$$

which is justified by the law of large numbers, which states that (under mild conditions) the sample average converges to the expected value as $n \to \infty$. In our case the random variable $Z$ is the principal component $v^\top x$, and the i.i.d. observations are the projections of our datapoints, i.e. $z_i = v^\top x_i$. 
CHAPTER 3. DIMENSIONALITY REDUCTION

By the inductive hypothesis, we know that $v_1, \ldots, v_k$ are orthonormal eigenvectors of $X^\top X$. Denote the $i$th largest eigenvalue of $X^\top X$ by $\lambda_i$, noting that $X^\top X v_i = \lambda_i v_i$.

The Lagrangian of the objective function is

$$L(v) = v^\top X^\top X v - \lambda(v^\top v - 1) + \sum_{i=1}^k \eta_i v^\top v_i$$

First-order necessary conditions for optima imply that

$$0 = \nabla L(v_{k+1}) = 2X^\top X v_{k+1} - 2\lambda v_{k+1} + \sum_{i=1}^k \eta_i v_i$$

This implies that, if $v_{k+1}$ is orthogonal to $v_1, \ldots, v_k$ (as we constrain it to be), then

$$0 = v_j^\top 0 = 2v_j^\top X^\top X v_{k+1} - 2\lambda v_j^\top v_{k+1} + \sum_{i=1}^k \eta_i v_j^\top v_i$$

$$= 2(X^\top X v_j)^\top v_{k+1} + \eta_j$$

$$= 2(\lambda_j v_j)^\top v_{k+1} + \eta_j$$

$$= 2\lambda_j v_j^\top v_{k+1} + \eta_j$$

$$= \eta_j$$

for all $j = 1, \ldots, k$.

Plugging these values back into the optimality equation above, we see that $v_{k+1}$ must satisfy $X^\top X v_{k+1} = \lambda v_{k+1}$, i.e. $v_{k+1}$ is an eigenvector of $X^\top X$ with eigenvalue $\lambda$. As before, the value of the objective function is then $\lambda$. To maximize, we want the largest eigenvalue, but we must respect the constraints that $v_{k+1}$ is orthogonal to $v_1, \ldots, v_k$. Clearly if $v_{k+1}$ is equal to any of these eigenvectors (up to sign), then one of these constraints will not be satisfied. Thus to maximize the expression, $v_{k+1}$ should be a unit eigenvector of $X^\top X$ corresponding to its $(k+1)$st largest eigenvalue. By the spectral theorem, we can always choose this vector in such a way that it is orthogonal to $v_1, \ldots, v_k$, so we are done.

We have shown that the loading vectors are orthonormal eigenvectors of $X^\top X$. In other words, they are right-singular vectors of $X$, so they can all be found simultaneously by computing the SVD of $X$.

**Projecting onto the PCA coordinate system**

Once we have computed the loading vectors, we can use them as a new coordinate system. The $k$th principal component of a datapoint $x_i \in \mathbb{R}^d$ is defined as the scalar projection of $x_i$ onto the $k$th loading vector $v_k$, i.e. $x_i^\top v_k$. We can compute all the principal components of all the datapoints at once using a matrix-matrix multiplication:

$$Z_k = X V_k$$

where $V_k \in \mathbb{R}^{d \times k}$ is a matrix whose columns are the first $k$ loading vectors $v_1, \ldots, v_k$. 
Below we plot the result of such a projection in the case \( d = k = 2 \):

![Figure 3.1: Left: data points; Right: PCA projection of data points](image)

Observe that the data are uncorrelated in the projected space. Also note that this example does not show the power of PCA since we have not reduced the dimensionality of the data at all – the plot is merely to show the PCA coordinate transformation.

Once we’ve computed the principal components, we can approximately reconstruct the original points by

\[
\tilde{X}_k = Z_k V_k^\top = X V_k V_k^\top
\]

The rows of \( \tilde{X}_k \) are the projections of the original rows of \( X \) onto the subspace spanned by the loading vectors.

### Other derivations of PCA

We have given the most common derivation of PCA above, but it turns out that there are other ways to solve the optimization problem, or to arrive at the same formulation. These give us helpful additional perspectives on what PCA is doing.

### Changing coordinates

In PCA we want to find the unit length \( v \) that maximizes \( v^\top X^\top X v \). It turns out that there is a result, sometimes referred to as the variational characterization of eigenvalues, that tells us which vectors \( v \) achieve this. The key idea in the proof is a length-preserving change of coordinates.

**Theorem.** Let \( A \in \mathbb{R}^{d \times d} \) be symmetric. Then for any \( v \in \mathbb{R}^d \) satisfying \( \|v\|_2 = 1 \),

\[
\lambda_{\min}(A) \leq v^\top A v \leq \lambda_{\max}(A)
\]

where for both bounds, equality holds if and only if \( v \) is a corresponding eigenvector.

**Proof.** We show only the max case because the argument for the min case is entirely analogous.

Since \( A \) is symmetric, we can decompose it as \( A = Q \Lambda Q^\top \), where \( Q \in \mathbb{R}^{d \times d} \) is orthogonal and \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d) \) contains the eigenvalues of \( A \). For any \( v \) satisfying \( \|v\|_2 = 1 \), define \( z = Q^\top v \), noting that the relationship between \( v \) and \( z \) is one-to-one because \( Q \) is invertible and that \( \|z\|_2 = 1 \) because \( Q \) is orthogonal. Hence

\[
\max_{\|v\|_2 = 1} v^\top A v = \max_{\|z\|_2 = 1} z^\top \Lambda z = \max_{\|z\|_2 = 1} \sum_{i=1}^d \lambda_i z_i^2
\]
CHAPTER 3. DIMENSIONALITY REDUCTION

We note that
\[ \sum_{i=1}^{d} \lambda_i z_i^2 \leq \sum_{i=1}^{d} \lambda_{\text{max}}(A) z_i^2 = \lambda_{\text{max}}(A) \sum_{i=1}^{d} z_i^2 \]
so the constraint \( \|z\|_2^2 = \sum_{i=1}^{d} z_i^2 = 1 \) implies
\[ \sum_{i=1}^{d} \lambda_i z_i^2 \leq \lambda_{\text{max}}(A) \]

Defining \( I = \{ i : \lambda_i = \lambda_{\text{max}}(A) \} \), the index set of the largest eigenvalue, we see that the bound is achieved with equality if and only if \( \sum_{i \in I} z_i^2 = 1 \) and \( z_j = 0 \) for \( j \notin I \). Suppose \( z^* \) satisfies this condition. Then writing \( q_1, \ldots, q_d \) for the columns of \( Q \), we have
\[ v^* = Q z^* = \sum_{i=1}^{d} z_i^* q_i = \sum_{i \in I} z_i^* q_i \]

Recall that \( q_1, \ldots, q_d \) are eigenvectors of \( A \) and form an orthonormal basis for \( \mathbb{R}^d \). Therefore by construction, the set \( \{ q_i : i \in I \} \) forms an orthonormal basis for the eigenspace of \( \lambda_{\text{max}}(A) \). Hence \( v^* \), which is a linear combination of these, lies in that eigenspace and thus is an eigenvector of \( A \) corresponding to \( \lambda_{\text{max}}(A) \).

Conversely, suppose \( v \in \mathbb{R}^d \) is unit-length but not an eigenvector corresponding to \( \lambda_{\text{max}}(A) \). The vectors \( q_1, \ldots, q_d \) are still a basis for \( \mathbb{R}^d \), so we have a unique expansion
\[ v = z_1 q_1 + \cdots + z_d q_d \]
Since \( v \) does not lie in the eigenspace of \( \lambda_{\text{max}}(A) \), one of the components \( z_j \) must be nonzero for an index \( j \notin I \), so equality does not hold in the bound above.

With this result established, we see that the vector we seek (which maximizes \( v^\top X^\top X v \)) must be an eigenvector corresponding to \( \lambda_{\text{max}}(X^\top X) \). This is the same solution we derived via the Lagrangian formulation above.

**Minimizing reconstruction error**

Recall that ordinary least squares minimizes the vertical distance between the fitted line and the data points. We show that PCA can be interpreted as minimizing the perpendicular distance between the data points and the subspace onto which we are projecting them.

The orthogonal projection of a vector \( x \) onto the subspace spanned by a unit vector \( v \) equals \( v \) scaled by the scalar projection of \( x \) onto \( v \):
\[ P_v x = (x^\top v) v \]
Suppose we want to minimize the total reconstruction error:
\[ \sum_{i=1}^{n} \|x_i - P_v x_i\|^2 \]
For any \( x \in \mathbb{R}^d \), we know \( x - P_v x \perp P_v x \), so the Pythagorean Theorem tells us that
\[ \|x - P_v x\|^2 + \|P_v x\|^2 = \|x\|^2 \]
Thus
\[ \sum_{i=1}^{n} \| x_i - P_v x_i \|^2 = \sum_{i=1}^{n} \left( \| x_i \|^2 - \| P_v x_i \|^2 \right) \]
\[ = \sum_{i=1}^{n} \| x_i \|^2 - \sum_{i=1}^{n} \| (x_i^\top v) v \|^2 \]
\[ = \sum_{i=1}^{n} \| x_i \|^2 - \sum_{i=1}^{n} (x_i^\top v)^2 \]

Then since the first term \( \sum_{i=1}^{n} \| x_i \|^2 \) is constant with respect to \( v \), minimizing reconstruction error is equivalent to maximizing \( \sum_{i=1}^{n} (x_i^\top v)^2 \), which is (up to an irrelevant positive constant factor \( 1/n \)) the projected variance.

Another way to write this interpretation is that the reconstructed matrix \( \tilde{X}_k \) is the best rank-\( k \) approximation to \( X \) in the Frobenius norm. To see this, first note that (writing \( X = \sum_{i=1}^{d} \sigma_i u_i v_i^\top \))
\[ \tilde{X}_k = X V_k V_k^\top = \sum_{i=1}^{d} \sigma_i u_i v_i^\top V_k V_k^\top \]

By orthonormality, the product \( v_i^\top V_k \) results in a \( k \)-dimensional row vector with 1 in the \( i \)th place and 0 everywhere else, i.e. \( e_i^\top \), as long as \( i \leq k \). In this case,
\[ v_i^\top V_k V_k^\top = e_i^\top (V_k e_i)^\top = v_i^\top \]
If \( i > k \), \( v_i^\top V_k = 0^\top \), so the term disappears. Therefore we see that
\[ \tilde{X}_k = \sum_{i=1}^{d} \sigma_i u_i v_i^\top V_k V_k^\top = \sum_{i=1}^{k} \sigma_i u_i v_i^\top \]
which is the best rank-\( k \) approximation to \( X \) by the Eckart-Young theorem.

### Probabilistic PCA

We have seen probabilistic motivations or derivations of many of the methods discussed so far in this class. In a similar vein, **probabilistic PCA** (PPCA) is a **generative** model for PCA. Here we make the following assumptions about how the data were generated: for each datapoint \( i \), there is a \( k \)-dimensional **latent variable**
\[ z_i \sim \mathcal{N}(0, I) \]
which we cannot observe, and the actual \( d \)-dimensional observation is distributed conditionally on this latent variable as
\[ x_i | z_i \sim \mathcal{N}(\Lambda z_i + \mu, \Psi) \]
Here \( \Lambda \in \mathbb{R}^{d \times k} \) and \( \mu \in \mathbb{R}^d \) are parameters to be estimated. Since \( z_i \) is Gaussian and \( x_i | z_i \) is Gaussian, so its marginal \( x_i \) is Gaussian. In particular, by integrating out the latent variable
\[ p(x_i) = \int_z p(x_i, z) dz = \int_z p(x_i | z)p(z) dz \]
one can show that
\[ x_i \sim \mathcal{N}(\mu, \Lambda \Lambda^\top + \Psi) \]
It is common to assume $\Psi = \sigma^2 \mathbf{I}$. In this case, if we let $\sigma^2 \to 0$, we recover the original PCA solution in the sense that the columnspace of $\mathbf{A}_{\text{MLE}}$ approaches the PCA subspace (i.e. the columnspace of $\mathbf{V}_k$). \(^2\)

### 3.2 Canonical Correlation Analysis

PCA provided us with a dimensionality-reduction approach that didn’t use the labels $y$ in any way. In that way, it was fundamentally unsupervised by nature. However, we can imagine that there can be situations in which the most relevant directions in $\mathbf{x}$ for understanding $y$ are not necessarily the directions of greatest variation in $\mathbf{x}$. For example, what if the $\mathbf{x}$ data by nature was contaminated with a strong correlated noise signal? PCA would find the noise dimensions to be those that have the greatest variation and keep them, throwing away those dimensions where we could actually hope to get information relevant for predicting $y$!

The other potentially troublesome aspect of PCA is that it is not invariant to a change of units or scaling. If we changed the units of some feature from meters to millimeters, then all the values for that feature would increase by a factor of a thousand, and suddenly, this direction might be favored by PCA. This is unavoidable because there is no natural reference point that would allow us to treat units as arbitrary.

Consequently, it is important to have an approach to dimensionality reduction and the discovery of linear structure from data that does take advantage of paired $(\mathbf{x}, y)$ data, preferably in a way that is robust to linear transformations of both $\mathbf{x}$ and $y$ individually.

**A latent space view with Gaussian random variables**

What does it mean to extract the linear structure establishing the underlying relationship between $\mathbf{X}$ and $\mathbf{Y}$, two vector-valued quantities of which we have many paired samples. To understand what this should mean, we need to construct a model. The first thing that we do is assume we have a joint distribution for $X$ and $Y$ as random variables. In practice, we won’t have the random variables in distribution, just paired samples of them. But it is easier to start understanding what we want by assuming that we have the entire distribution. This corresponds to how well we think we can do given infinite amounts of training data. The next we do is assume a particular form for the random variables. Since we are interested in linear structure, jointly Gaussian random variables are a useful model.

Our goal is to extract the underlying relationship or commonality between $\mathbf{X}$ and $\mathbf{Y}$. To do this, we assume that we have three underlying iid standard Gaussian random vectors $\mathbf{Z}_j$ (representing the common/joint part), $\mathbf{Z}_x$ (representing the randomness that is purely in $\mathbf{x}$ and not shared by $\mathbf{Y}$), and $\mathbf{Z}_y$ (representing the randomness that is purely in $\mathbf{y}$ and not shared by $\mathbf{X}$). Then we can assume that they are related by an underlying linear relationship:

$$
\begin{bmatrix}
\mathbf{X} \\
\mathbf{Y}
\end{bmatrix} =
\begin{bmatrix}
A & B & 0 \\
0 & C & D
\end{bmatrix}
\begin{bmatrix}
\mathbf{Z}_x \\
\mathbf{Z}_j \\
\mathbf{Z}_y
\end{bmatrix}
$$

As is typical in these situations, there is going to be some ambiguity in choosing the $A, B, C, D$ matrices. But the important thing is that somehow the $B$ and $C$ matrices together capture the joint relationship between $\mathbf{X}$ and $\mathbf{Y}$.

\(^2\) See Tipping and Bishop’s original paper for derivations and more information.
How will such a joint relationship manifest in the joint distributions for $X$ and $Y$? To understand that, we should first consider the scalar case.

**Correlation and Scalar Gaussians**

For the scalar case, $A, B, C, D$ are just real numbers. So, the joint distribution of $X, Y$ is $\mathcal{N}(0, \Sigma)$ where $\Sigma = \begin{bmatrix} A^2 + B^2 & BC \\ BC & C^2 + D^2 \end{bmatrix}$. The first thing that we notice is that we cannot disentangle $B$ and $C$. The second is that the information about the joint relationship (which we know is encoded by $B$ and $C$) is all in the cross-covariance term, not in the individual variance term. Recall that we want to pull out the relationship in a way that does not depend on any individual scaling or linear transformation that we apply to $X$ and $Y$.

Here’s a neat fact: if $X$ and $Y$ are jointly Gaussian, i.e.

$$
\begin{bmatrix} X \\
Y \end{bmatrix} \sim \mathcal{N}(0, \Sigma)
$$

then we can define a distribution on *individually normalized* $X$ and $Y$ and have their joint inter-relationship entirely captured by $\rho(X,Y)$. First write

$$
\rho(X,Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}
$$

Then

$$
\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}
= \begin{bmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{bmatrix}
$$

so

$$
\begin{bmatrix} \sigma_x^{-1} & 0 \\ 0 & \sigma_y^{-1} \end{bmatrix} \begin{bmatrix} X \\
Y \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_x^{-1} & 0 \\ 0 & \sigma_y^{-1} \end{bmatrix} \Sigma \begin{bmatrix} \sigma_x^{-1} & 0 \\ 0 & \sigma_y^{-1} \end{bmatrix}^\top\right)
\sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)
$$

This $\rho$ quantity is the signature of the joint inter-relationship of the $X$ and $Y$ random variables. To make things explicit, once we have the $\rho = \frac{BC}{\sqrt{(A^2 + B^2)(C^2 + D^2)}}$, we can come up with many possible backstories for the latent picture behind the observed random variables. Here is one that splits the influence of the latent space proportionately.

$$
A = \sigma_x \sqrt{1 - |\rho|} \tag{3.2}
$$

$$
B = \sigma_x \sqrt{|\rho|} \tag{3.3}
$$

$$
C = \sigma_y \text{sign}(\rho) \sqrt{|\rho|} \tag{3.4}
$$

$$
D = \sigma_y \sqrt{1 - |\rho|} \tag{3.5}
$$

Because $A^2 + B^2 = \sigma_x^2$, $C^2 + D^2 = \sigma_y^2$, and $\rho = \frac{BC}{\sqrt{(A^2 + B^2)(C^2 + D^2)}}$, this works.
Pearson Correlation

Although we defined this $\rho$ above for a pair of jointly Gaussian random variables, it is really about linear structure. The Pearson Correlation Coefficient $\rho(X, Y)$ is effectively a way to measure how linearly related (in other words, how well a linear model captures the relationship between) random variables $X$ and $Y$.

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

Here are some important facts about it:

- It is commutative: $\rho(X, Y) = \rho(Y, X)$
- It always lies between -1 and 1: $-1 \leq \rho(X, Y) \leq 1$
- It is completely invariant to affine transformations: for any $a, b, c, d \in \mathbb{R}$,

$$\rho(aX + b, cY + d) = \frac{\text{Cov}(aX + b, cY + d)}{\sqrt{\text{Var}(aX + b) \text{Var}(cY + d)}}$$
$$= \frac{\text{Cov}(aX, cY)}{\sqrt{a^2 \text{Var}(X) \cdot c^2 \text{Var}(Y)}}$$
$$= \frac{a \cdot c \cdot \text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$
$$= \rho(X, Y)$$

The correlation is defined in terms of random variables rather than observed data. Assume now that $x, y \in \mathbb{R}^n$ are vectors containing $n$ independent observations of $X$ and $Y$, respectively. Recall the law of large numbers, which states that for i.i.d. $X_i$ with mean $\mu$,

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu \quad \text{as } n \to \infty$$

We can use this law to justify a sample-based approximation to the mean:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \approx \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where the bar indicates the sample average, i.e. $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Then as a special case we have

$$\text{Var}(X) = \text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \approx \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$\text{Var}(Y) = \text{Cov}(Y, Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] \approx \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Plugging these estimates into the definition for correlation and canceling the factor of $1/n$ leads us to the Sample Pearson Correlation Coefficient $\hat{\rho}$:

$$\hat{\rho}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$
3.2. CANONICAL CORRELATION ANALYSIS

\[ \rho = \frac{\hat{x}^\top \hat{y}}{\sqrt{\hat{x}^\top \hat{x} \cdot \hat{y}^\top \hat{y}}} \quad \text{where} \quad \hat{x} = x - \bar{x}, \hat{y} = y - \bar{y} \]

Here are some 2-D scatterplots and their corresponding correlation coefficients:

You should notice that:

- The magnitude of \( \hat{\rho} \) increases as \( X \) and \( Y \) become more linearly correlated.
- The sign of \( \hat{\rho} \) tells whether \( X \) and \( Y \) have a positive or negative relationship.
- The correlation coefficient is undefined if either \( X \) or \( Y \) has 0 variance (horizontal line).

Canonical Correlation Analysis

**Canonical Correlation Analysis (CCA)** is a method of modeling the relationship between two point sets by making use of the correlation coefficients.

As in PCA, it is useful to start with trying to find the directions that represent the most correlation. You can think of this as finding the parts of \( X_{rv} \) and \( Y_{rv} \) that depend on the first coordinate of \( Z_{ij} \), where we choose the convention that the first coordinate represents the most shared dimension. We will then see how to move on to get the rest.

Formally, given zero-mean random vectors \( X_{rv} \in \mathbb{R}^p \) and \( Y_{rv} \in \mathbb{R}^q \), we want to find projection vectors \( u \in \mathbb{R}^p \) and \( v \in \mathbb{R}^q \) that maximizes the correlation between \( X_{rv}^\top u \) and \( Y_{rv}^\top v \):

\[
\max_{u,v} \rho(X_{rv}^\top u, Y_{rv}^\top v) = \max_{u,v} \frac{\text{Cov}(X_{rv}^\top u, Y_{rv}^\top v)}{\sqrt{\text{Var}(X_{rv}^\top u) \cdot \text{Var}(Y_{rv}^\top v)}}
\]

Observe that

\[
\text{Cov}(X_{rv}^\top u, Y_{rv}^\top v) = \mathbb{E}[(X_{rv}^\top u - \mathbb{E}[X_{rv}^\top u])(Y_{rv}^\top v - \mathbb{E}[Y_{rv}^\top v])] \\
= \mathbb{E}[u^\top (X_{rv} - \mathbb{E}[X_{rv}])(Y_{rv} - \mathbb{E}[Y_{rv}])^\top v] \\
= u^\top \mathbb{E}[(X_{rv} - \mathbb{E}[X_{rv}])(Y_{rv} - \mathbb{E}[Y_{rv}])^\top] v \\
= u^\top \text{Cov}(X_{rv}, Y_{rv}) v
\]

which also implies (since \( \text{Var}(Z) = \text{Cov}(Z, Z) \) for any random variable \( Z \)) that

\[
\text{Var}(X_{rv}^\top u) = u^\top \text{Cov}(X_{rv}, X_{rv}) u \\
\text{Var}(Y_{rv}^\top v) = v^\top \text{Cov}(Y_{rv}, Y_{rv}) v
\]
so the correlation can be written
\[
\rho(X_{rv}^\top u, Y_{rv}^\top v) = \frac{u^\top \text{Cov}(X_{rv}, Y_{rv}) v}{\sqrt{u^\top \text{Cov}(X_{rv}, X_{rv}) u \cdot v^\top \text{Cov}(Y_{rv}, Y_{rv}) v}}
\]

Unfortunately, we do not have access to the true distributions of $X_{rv}$ and $Y_{rv}$, so we cannot compute these covariances matrices. However, we can estimate them from data. Assume now that we are given zero-mean data matrices $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$, where the rows of the matrix $X$ are i.i.d. samples $x_i \in \mathbb{R}^p$ from the random variable $X_{rv}$, and correspondingly for $Y_{rv}$. Then
\[
\text{Cov}(X_{rv}, Y_{rv}) = \mathbb{E}[(X_{rv} - \mathbb{E}[X_{rv}])(Y_{rv} - \mathbb{E}[Y_{rv}])^\top] = \mathbb{E}[X_{rv} Y_{rv}^\top] \approx \frac{1}{n} \sum_{i=1}^{n} x_i y_i^\top = \frac{1}{n} X^\top Y
\]
where again the sample-based approximation is justified by the law of large numbers. Similarly,
\[
\text{Cov}(X_{rv}, X_{rv}) = \mathbb{E}[X_{rv} X_{rv}^\top] \approx \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top = \frac{1}{n} X^\top X
\]
\[
\text{Cov}(Y_{rv}, Y_{rv}) = \mathbb{E}[Y_{rv} Y_{rv}^\top] \approx \frac{1}{n} \sum_{i=1}^{n} y_i y_i^\top = \frac{1}{n} Y^\top Y
\]
Plugging these estimates in for the true covariance matrices, we arrive at the problem
\[
\max_{u,v} \frac{u^\top \left( \frac{1}{n} X^\top Y \right) u}{\sqrt{u^\top \left( \frac{1}{n} X^\top X \right) u \cdot v^\top \left( \frac{1}{n} Y^\top Y \right) v}} = \max_{u,v} \frac{u^\top X^\top Y v}{\sqrt{u^\top X^\top X u \cdot v^\top Y^\top Y v \rho(X_{rv}, Y_{rv})}}
\]
Let’s try to massage the maximization problem into a form that we can reason with more easily. Our strategy is to choose matrices to transform $X$ and $Y$ such that the maximization problem is equivalent but easier to understand.

1. First, let’s choose matrices $W_x, W_y$ to whiten $X$ and $Y$. This will make the (co)variance matrices $(XW_x)^\top (XW_x)$ and $(YW_y)^\top (YW_y)$ become identity matrices and simplify our expression. To do this, note that $X^\top X$ is positive definite (and hence symmetric), so we can employ the eigendecomposition
\[
X^\top X = U_x S_x U_x^\top
\]
Since
\[
S_x = \text{diag}(\lambda_1(X^\top X), \ldots, \lambda_d(X^\top X))
\]
where all the eigenvalues are positive, we can define the “square root” of this matrix by taking the square root of every diagonal entry:
\[
S_x^{1/2} = \text{diag}\left(\sqrt{\lambda_1(X^\top X)}, \ldots, \sqrt{\lambda_d(X^\top X)}\right)
\]
Then, defining $W_x = U_x S_x^{-1/2} U_x^\top$, we have
\[
(XW_x)^\top (XW_x) = W_x^\top X^\top X W_x
= U_x S_x^{-1/2} U_x^\top U_x S_x U_x^\top U_x S_x^{-1/2} U_x^\top
= U_x S_x^{-1/2} S_x S_x^{-1/2} U_x^\top
\]
Let's denote the decorrelated data as reducing the problem to a sequence of scalar problems. The whitening was a normalizing change of coordinates. The decorrelation is there to produce a whitening matrix \( W_x = U_x \Sigma_x^{-1/2} U_x^\top \) for \( X \).

Let’s denote the whitened data \( X_w = X W_x \) and \( Y_w = Y W_y \). Then by the change of variables \( u_w = W_x^{-1} u, v_w = W_y^{-1} v, \)

\[
\max_{u,v} \hat{\rho}(Xu, Yv) = \max_{u,v} \frac{(Xu)^\top Yv}{\sqrt{(Xu)^\top Xu(Yv)^\top Yv}}
\]

\[
= \max_{u,v} \frac{(Xu)^\top Yv}{\sqrt{(XW_x W_x^{-1} u)^\top XW_x W_x^{-1} u}}
\]

\[
= \max_{u,v} \frac{(Xu)^\top Yv}{\sqrt{(XW_x W_x^{-1} u)^\top XW_x W_x^{-1} u(YW_y W_y^{-1} v)^\top YW_y W_y^{-1} v}}
\]

\[
= \max_{u,v} \frac{(Xu)^\top Yv}{\sqrt{(Xu)^\top Xu(Yv)^\top Yv}}
\]

\[
= \max_{u,v} \frac{(Xu)^\top Yv}{\sqrt{(Xu)^\top Xu(Yv)^\top Yv}}
\]

\[
= \max_{u,v} \frac{(Xu)^\top Yv}{\sqrt{(Xu)^\top Xu(Yv)^\top Yv}}
\]

Note we have used the fact that \( X_w^\top X_w \) and \( Y_w^\top Y_w \) are identity matrices by construction.

2. Second, let’s choose matrices \( D_x, D_y \) to decorrelate \( X_w \) and \( Y_w \). This will let us simplify the covariance matrix \( (X_w D_x)^\top (Y_w D_y) \) into a diagonal matrix.

Recall that our ultimate goal is to understand the underlying latent structure behind \( X_{tv} \) and \( Y_{rv} \). The whitening was a normalizing change of coordinates. The decorrelation is there so that we can pick out independent underlying components of \( Z_j \). (Since jointly Gaussian random variables are independent if they are uncorrelated.) Alternatively, you can consider decorrelation as reducing the problem to a sequence of scalar problems.

To do this, we’ll make use of the SVD:

\[
X_w^\top Y_w = USV^\top
\]

The choice of \( U \) for \( D_x \) and \( V \) for \( D_y \) accomplishes our goal, since

\[
(X_w U)^\top (Y_w V) = U^\top X_w^\top Y_w V = U^\top (USV^\top) V = S
\]

Let’s denote the decorrelated data \( X_d = X_w D_y \) and \( Y_d = Y_w W_y \). Then by the change of variables \( u_d = D_x^{-1} u_w, v_d = D_y^{-1} v_w, \)

\[
\max_{u_v, v_w} \hat{\rho}(X_w u_w, Y_w v_w) = \max_{u_w, v_w} \frac{(X_w u_w)^\top Y_w v_w}{\sqrt{(X_w u_w)^\top X_w u_w \cdot v_w^\top v_w}}
\]

\[
= \max_{u_w, v_w} \frac{(X_w D_x D_x^{-1} u_w)^\top Y_w D_y D_y^{-1} v_w}{\sqrt{(D_x u_w)^\top D_x u_w \cdot (D_y v_w)^\top D_y v_w}}
\]

\[
= \max_{u_w, v_w} \frac{(X_d u_d)^\top Y_d v_d}{\sqrt{u_d^\top u_d \cdot v_d^\top v_d}}
\]
Without loss of generality, suppose \( \mathbf{u}_d \) and \( \mathbf{v}_d \) are unit vectors\(^3\) so that the denominator becomes 1, and we can ignore it:

\[
\max_{\mathbf{u}_d, \mathbf{v}_d} \frac{\mathbf{u}_d^\top \mathbf{S} \mathbf{v}_d}{\sqrt{\mathbf{u}_d^\top \mathbf{u}_d \cdot \mathbf{v}_d^\top \mathbf{v}_d}} = \max_{\mathbf{u}_d, \mathbf{v}_d} \frac{\mathbf{u}_d^\top \mathbf{S} \mathbf{v}_d}{\mathbf{u}_d^\top \mathbf{S} \mathbf{v}_d} = \max_{\mathbf{u}_d, \mathbf{v}_d} \mathbf{u}_d^\top \mathbf{S} \mathbf{v}_d
\]

The diagonal nature of \( \mathbf{S} \) implies \( S_{ij} = 0 \) for \( i \neq j \), so our simplified objective expands as

\[
\mathbf{u}_d^\top \mathbf{S} \mathbf{v}_d = \sum_i \sum_j (\mathbf{u}_d)_i S_{ij} (\mathbf{v}_d)_j = \sum_i S_{ii} (\mathbf{u}_d)_i (\mathbf{v}_d)_i
\]

where \( S_{ii} \), the singular values of \( \mathbf{X}_w^\top \mathbf{Y}_w \), are arranged in descending order. Thus we have a weighted sum of these singular values, where the weights are given by the entries of \( \mathbf{u}_d \) and \( \mathbf{v}_d \), which are constrained to have unit norm. To maximize the sum, we “put all our eggs in one basket” and extract \( S_{11} \) by setting the first components of \( \mathbf{u}_d \) and \( \mathbf{v}_d \) to 1, and the rest to 0:

\[
\mathbf{u}_d = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^p \\
\mathbf{v}_d = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^q
\]

Any other arrangement would put weight on \( S_{ii} \) at the expense of taking that weight away from \( S_{11} \), which is the largest, thus reducing the value of the sum.

Finally we have an analytical solution, but it is in a different coordinate system than our original problem! In particular, \( \mathbf{u}_d \) and \( \mathbf{v}_d \) are the best weights in a coordinate system where the data has been whitened and decorrelated. To bring it back to our original coordinate system and find the vectors we actually care about (\( \mathbf{u} \) and \( \mathbf{v} \)), we must invert the changes of variables we made:

\[
\mathbf{u} = \mathbf{W}_x \mathbf{u}_w = \mathbf{W}_x \mathbf{D}_x \mathbf{u}_d \\
\mathbf{v} = \mathbf{W}_y \mathbf{v}_w = \mathbf{W}_y \mathbf{D}_y \mathbf{v}_d
\]

More generally, to get the best \( k \) directions, we choose

\[
\mathbf{U}_d = \begin{bmatrix} \mathbf{I}_k \\ 0_{p-k,k} \end{bmatrix} \in \mathbb{R}^{p \times k} \\
\mathbf{V}_d = \begin{bmatrix} \mathbf{I}_k \\ 0_{q-k,k} \end{bmatrix} \in \mathbb{R}^{q \times k}
\]

where \( \mathbf{I}_k \) denotes the \( k \)-dimensional identity matrix. Then

\[
\mathbf{U} = \mathbf{W}_x \mathbf{D}_x \mathbf{U}_d \\
\mathbf{V} = \mathbf{W}_y \mathbf{D}_y \mathbf{V}_d
\]

Note that \( \mathbf{U}_d \) and \( \mathbf{V}_d \) have orthogonal columns. The columns of \( \mathbf{U} \) and \( \mathbf{V} \), which are the projection directions we seek, will in general not be orthogonal, but they will be linearly independent (since they come from the application of invertible matrices to the columns of \( \mathbf{U}_d, \mathbf{V}_d \)).

\(^3\) Why can we assume this? Observe that the value of the objective does not change if we replace \( \mathbf{u}_d \) by \( \alpha \mathbf{u}_d \) and \( \mathbf{v}_d \) by \( \beta \mathbf{v}_d \), where \( \alpha \) and \( \beta \) are any positive constants. Thus if there are maximizers \( \mathbf{u}_d, \mathbf{v}_d \) which are not unit vectors, then \( \mathbf{u}_d/\|\mathbf{u}_d\| \) and \( \mathbf{v}_d/\|\mathbf{v}_d\| \) (which are unit vectors) are also maximizers.
Following (3.2), (3.3), (3.4), and (3.5), it is also possible to use what we have calculated to give an explicit learned latent-space realization for the $X_{rv}, Y_{rv}$ in terms of standard Gaussian random variables $Z_X, Z_J, Z_Y$. In particular, matrices $A, B, C, D$ of the appropriate sizes. This is left as an exercise to the reader once you realize that after whitening and decorrelating (both invertible transformations), we are left with a collection of scalar problems that would represent independent random variables if all the variables were indeed jointly Gaussian.

CCA thus illustrates how it is possible to learn a latent representation for common (linear) structure given paired data. This is a powerful idea not limited to the specific case of CCA. In effect, CCA shows how we can discover (synthesize) features that distill what aspects of input data is relevant for understanding output data.

This is subtly different from what happens in ordinary least squares because in ordinary least squares, each individual element of $y$ is predicted independently. In OLS, the different output variables are not used collectively to distill the most relevant dimensions of the input. By contrast, in CCA, the different output variables do vote collectively to determine relevant dimensions in the input.

**Comparison with PCA**

An advantage of CCA over PCA is that it is invariant to scalings and affine transformations of $X$ and $Y$. Consider a simplified scenario in which two matrix-valued random variables $X, Y$ satisfy $Y = X + \epsilon$ where the noise $\epsilon$ has huge variance. What happens when we run PCA on $Y$? Since PCA maximizes variance, it will actually project $Y$ (largely) into the column space of $\epsilon$! However, we’re interested in $Y$’s relationship to $X$, not its dependence on noise. How can we fix this? As it turns out, CCA solves this issue. Instead of maximizing variance of $Y$, we maximize correlation between $X$ and $Y$. In some sense, we want the maximize “predictive power” of information we have.

**CCA regression**

Once we’ve computed the CCA coefficients, one application is to use them for regression tasks, predicting $Y$ from $X$ (or vice-versa). Recall that the correlation coefficient attains a greater value when the two sets of data are more linearly correlated. Thus, it makes sense to find the $k \times k$ weight matrix $A$ that linearly relates $X_U$ and $Y_V$. We can accomplish this with ordinary least squares.

Denote the projected data matrices by $X_c = X_U$ and $Y_c = Y_V$. Observe that $X_c$ and $Y_c$ are zero-mean because they are linear transformations of $X$ and $Y$, which are zero-mean. Thus we can fit a linear model relating the two:

\[ Y_c \approx X_cA \]

The least-squares solution is given by

\[
A = (X_c^\top X_c)^{-1}X_c^\top Y_c \\
= (U^\top X^\top XU)^{-1}U^\top X^\top YV
\]

However, since what we really want is an estimate of $Y$ given new (zero-mean) observations $\tilde{X}$ (or vice-versa), it’s useful to have the entire series of transformations that relates the two. The predicted canonical variables are given by

\[
\hat{Y}_c = \tilde{X}cA = \tilde{X}U(U^\top X^\top XU)^{-1}U^\top X^\top YV
\]
Then we use the canonical variables to compute the actual values:

\[
\hat{Y} = \hat{Y}_c (V^T V)^{-1} V^T \\
= \tilde{X} U (U^T X^T X U)^{-1} (U^T X^T Y V) (V^T V)^{-1} V^T
\]

We can collapse all these terms into a single matrix \( A_{eq} \) that gives the prediction \( \hat{Y} \) from \( \tilde{X} \):

\[
A_{eq} = \underbrace{U}_{\text{projection}} (U^T X^T X U)^{-1} (U^T X^T Y V) (V^T V)^{-1} V^T
\]
4.1 Beyond Least Squares

All the models we’ve seen so far are linear in the parameters we’re trying to learn. That is, our prediction \( \hat{y} = f(x; \theta) \) is some linear function of the parameters \( \theta \). For example, in OLS, \( \theta = w \) and the residuals \( r_i \) are computed by \( y_i - w^\top x_i \), which is linear in the components of \( w \). In the case of least-squares polynomial regression, the predicted value is not a linear function of the input \( x \), but it is still a linear function of the augmented input \( \phi(x_i) \).

However, we may have models which are nonlinear functions of their parameters. Let’s consider a motivating example. Suppose we want to estimate the 2D position \( \theta = (\theta_1, \theta_2) \) of some entity, for example a robot. The information we have to work with are noisy distance estimates \( Y_i \in \mathbb{R} \) from \( n \) sensors whose positions \( x_i \in \mathbb{R}^2 \) are fixed and known. If we assume i.i.d. Gaussian noise as usual, our statistical model has the form

\[
Y_i = \|x_i - \theta\|_2 + Z_i, \quad Z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n
\]

where

\[
\|x_i - \theta\|_2 = \sqrt{(x_{i1} - \theta_1)^2 + (x_{i2} - \theta_2)^2}
\]

Our prediction is

\[
\hat{y} = f(x; \theta) = \|x - \theta\|_2
\]

which is clearly not linear in \( \theta \).

**MLE Formulation**

More generally, let us assume a model where \( f \) is now some arbitrary differentiable function and \( \theta \in \mathbb{R}^d \):

\[
Y_i = f(x_i; \theta) + Z_i, \quad Z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n
\]

Note that this implies \( Y_i \mid x_i \sim \mathcal{N}(f(x_i; \theta), \sigma^2) \). The maximum likelihood estimator is given by

\[
\hat{\theta}_{MLE} = \arg \max_{\theta} \ell(\theta; X, y)
\]

\[
= \arg \max_{\theta} \sum_{i=1}^n \log p(y_i \mid x_i, \theta)
\]
\[
\theta^* = \arg \max_{\theta} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i; \theta))^2}{2\sigma^2}\right)
\]
\[
= \arg \max_{\theta} \sum_{i=1}^{n} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_i - f(x_i; \theta))^2\right]
\]
\[
= \arg \min_{\theta} \sum_{i=1}^{n} (y_i - f(x_i; \theta))^2
\]

Observe that the objective function is a sum of squared residuals as we’ve seen before, even though the function \( f \) is nonlinear in general. For this reason this method is called **nonlinear least squares**.

Motivated by the MLE formulation above, we consider the following optimization problem:

\[
\min_{\theta} \epsilon(\theta) = \min_{\theta} \sum_{i=1}^{n} (y_i - f(x_i; \theta))^2
\]

One way to solve this optimization problem is to find all of its critical points and choose the point that minimizes the objective. From **first-order optimality conditions**, the gradient of the objective function at any minimum must be zero:

\[
\nabla_{\theta} \epsilon = 2 \sum_{i=1}^{n} (y_i - f(x_i; \theta)) \nabla_{\theta} f(x_i; \theta) = 0
\]

In compact matrix notation:

\[
\nabla_{\theta} \epsilon = J(\theta)\top (y - F(\theta)) = 0
\]

where

\[
F(\theta) = \begin{bmatrix} f(x_1; \theta) \\ \vdots \\ f(x_n; \theta) \end{bmatrix}, \quad J(\theta) = \begin{bmatrix} \nabla_{\theta} f(x_1; \theta) \top \\ \vdots \\ \nabla_{\theta} f(x_n; \theta) \top \end{bmatrix}
\]

\( J \) is also referred to as the **Jacobian** of \( F \). Observe that in the special case when \( f \) is linear in \( \theta \) (i.e. \( f(x_i; \theta) = \theta \top x_i \)), the gradient \( \nabla_{\theta} \epsilon \) will only have \( \theta \) in one place because the term \( \nabla_{\theta} f(x_i; \theta) \) will only depend on \( x_i \):

\[
\nabla_{\theta} \epsilon = 2 \sum_{i=1}^{n} (y_i - \theta \top x_i) \nabla_{\theta} (\theta \top x_i) = 2 \sum_{i=1}^{n} (y_i - \theta \top x_i) x_i
\]

and we can derive a closed-form solution for \( \theta \), arriving at the OLS solution:

\[
2X\top (y - X\theta) = 0
\]
\[
2X\top y - 2X\top X\theta = 0
\]
\[
X\top y = X\top X\theta
\]
\[
\theta = (X\top X)^{-1}X\top y
\]

However, in the general case where \( f \) is nonlinear in \( \theta \), it would not be possible to write out a closed-form solution for \( \theta \).

**Remark:** Without more assumptions on \( f \), the NLS objective is not convex in general. This means that the first-order optimality condition is a **necessary** but not **sufficient** condition for a local minimum. That is, it is possible that the derivative is zero for some value of \( \theta \), but that value is not a local minimum. It could be a saddle point, or worse, a local maximum! Even if it is a minimum, it may not be the global minimum.
4.1. BEYOND LEAST SQUARES

Gauss-Newton Algorithm

Since there is no closed-form solution to the nonlinear least squares optimization problem, we must resort an iterative algorithm instead. One such algorithm is the Gauss-Newton algorithm. At each iteration, this method linearly approximates the function \( F \) about the current iterate and solves a least-squares problem involving the linearization in order to compute the next iterate.

Let’s say that we have a “guess” for \( \theta \) at iteration \( k \), which we denote \( \theta^{(k)} \). We consider the first-order approximation of \( F(\theta) \) about \( \theta^{(k)} \):

\[
F(\theta) \approx \tilde{F}(\theta) = F(\theta^{(k)}) + \nabla_{\theta} F(\theta^{(k)})(\theta - \theta^{(k)})
\]

where \( \Delta \theta := \theta - \theta^{(k)} \).

Now since \( \tilde{F} \) is linear in \( \Delta \theta \) (the Jacobian and \( F \) are just constants: functions evaluated at \( \theta^{(k)} \)), we can use the closed form solution for \( \Delta \theta \) from the optimality condition to update our current guess \( \theta^{(k)} \). Applying the first-order optimality condition to the objective \( \tilde{F} \) yields the following equation:

\[
0 = J_{\tilde{F}}(\theta)^{\top}(y - \tilde{F}(\theta)) = J(\theta^{(k)})^{\top}\left(y - \left(F(\theta^{(k)}) + J(\theta^{(k)})\Delta \theta\right)\right)
\]

Note that the Jacobian of the linearized function \( \tilde{F} \), evaluated at any \( \theta \), is precisely \( J(\theta^{(k)}) \). Writing \( J = J(\theta^{(k)}) \) for brevity, we have

\[
J^{\top}y = J^{\top}(F(\theta^{(k)}) + J\Delta \theta)
\]

\[
J^{\top}(y - F(\theta^{(k)})) = J^{\top}\Delta \theta
\]

\[
\Delta \theta = (J^{\top}J)^{-1}J^{\top}(y - F(\theta^{(k)})) = (J^{\top}J)^{-1}J^{\top}\Delta y
\]

where \( \Delta y := y - F(\theta^{(k)}) \). By comparing this solution to OLS, we see that it is effectively solving

\[
\Delta \theta = \arg \min_{\delta \theta} ||J\delta \theta - \Delta y||^2
\]

Since \( \delta F \approx J\delta \theta \) is close to \( \theta^{(k)} \), this is saying that we choose a change to the weights that corrects for the current error in the function values, but it bases this calculation on the linearization of \( F \). Recalling that \( \Delta \theta = \theta - \theta^{(k)} \), we can improve upon our current guess \( \theta^{(k)} \) with the update

\[
\theta^{(k+1)} = \theta^{(k)} + \Delta \theta = \theta^{(k)} + (J^{\top}J)^{-1}J^{\top}\Delta y
\]

Note that the solution will depend on the initial value \( \theta^{(0)} \) in general. Here is the entire process laid out in steps: The choice for measuring convergence is up to the practitioner. Some common choices include testing changes in the objective value:

\[
\left| \frac{\epsilon^{(k+1)} - \epsilon^{(k)}}{\epsilon^{(k)}} \right| \leq \text{threshold}
\]

or in the iterates themselves:

\[
\max_{j} \left| \frac{\Delta \theta_j}{\theta_j^{(k)}} \right| \leq \text{threshold}
\]
4.2 Optimization

As we move into the realm of neural networks and beyond, we will be solving arbitrary problems of the form

$$\min_{x \in \mathcal{X}} f(x)$$

over an arbitrary objective function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and arbitrary domain $\mathcal{X}$. Solving such problems is the focus of optimization, an extensive field that has applications in control theory, finance, and machine learning. Most optimization problems do not necessarily have a closed form solution, and therefore an iterative algorithm is needed to solve them. As we will see, there is no one universal algorithm that is suited for solving all problems — rather, some algorithms are more suitable over others depending on the specific underlying assumptions about the problem, such as convexity and smoothness.

Gradients

Gradients form the basis for many of the optimization algorithms that we will study. Given that $f$ is continuously differentiable, the gradient is defined as the vector of partial derivatives of $f$, denoted by

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}$$

Why do we care about gradients? Among many reasons, the gradient being zero is a necessary condition for local minima. Let’s understand why this is true. We define critical points as points where the gradient is zero. These points can be classified into three categories:

(i) **local/global minimum**: a point $x \in \mathcal{X}$ such that there exists a neighborhood around $x$ where $f(x)$ attains the minimum value

(ii) **local/global minima**: a point $x \in \mathcal{X}$ such that there exists a neighborhood around $x$ where $f(x)$ attains the maximum value

(iii) **saddle point**: a point $x \in \mathcal{X}$ such that for all neighborhoods around $x$ there exists $y, z$ such that $f(y) \leq f(x) \leq f(z)$

In optimization we are interested in finding the global minimum of a function, but in many circumstances we may settle for local minima instead.

**Proposition 1.** If $x^*$ is a local minimum of $f$ and $f$ is continuously differentiable in a neighborhood of $x^*$, then $\nabla f(x^*) = 0$. 

---

**Algorithm 4: Gauss-Newton**

Initialize $\theta^{(0)}$ with some guess

while $\theta^{(k)}$ has not converged do

- Compute Jacobian with respect to the current iterate, $J = J(\theta^{(k)})$
- Compute $\Delta y = y - F(\theta^{(k)})$
- Update: $\theta^{(k+1)} = \theta^{(k)} + (J^T J)^{-1} J^T \Delta y$
4.3. CONVEXITY

Proof. See math4ml.

Note however, that while setting the gradient to zero is a necessary condition for local minima, it is not a sufficient condition. In many circumstances, the function we are optimizing may not have a local minima, and generally setting the gradient to zero could yield us local maxima or saddle points.

Hessian

Given that $f$ is twice continuously differentiable, we define the Hessian as the matrix of second partial derivatives of $f$, denoted by

$$\nabla^2 f = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2}
\end{bmatrix}$$

Just as with gradients being zero, the Hessian being PSD is another necessary condition for local minima.

**Proposition 2.** If $x^*$ is a local minimum of $f$ and $f$ is twice continuously differentiable in a neighborhood of $x^*$, then $\nabla^2 f(x^*)$ is positive semi-definite.

Proof. See math4ml.

Unfortunately, the gradient being zero and the Hessian being PSD together are still not sufficient conditions local minima. Slightly stronger conditions are needed to establish that a point is a strict local minimum.

**Proposition 3.** Suppose $f$ is twice continuously differentiable with $\nabla^2 f$ positive semi-definite in a neighborhood of $x^*$, and that $\nabla f(x^*) = 0$. Then $x^*$ is a local minimum of $f$. Furthermore if $\nabla^2 f(x^*)$ is positive definite, then $x^*$ is a strict local minimum.

Proof. See math4ml.

4.3 Convexity

**Convex optimization** is a subfield of optimization which deals with convex problems — problems in which the objective function is convex and the domain is a convex set. Convex functions are convenient due to their “bowl shape,” which induces many useful properties.
Given that \( f : \mathbb{R}^n \to \mathbb{R} \) is twice continuously differentiable, the following are equivalent conditions of convexity:

\[
\begin{align*}
(i) \quad f(tx + (1-t)y) &\leq tf(x) + (1-t)f(y), \quad \forall x, y, t \in [0,1] \\
(ii) \quad f(y) &\geq f(x) + \nabla f(x)^\top (y-x), \quad \forall x, y \\
(iii) \quad (\nabla f(y) - \nabla f(x))^\top (y-x) &\geq 0, \quad \forall x, y \\
(iv) \quad \nabla^2 f(x) &\succeq 0, \quad \forall x
\end{align*}
\]

Let’s understand the equivalent conditions of convexity. The first condition says that any two points \( x, y \), the function lies below the line segment connecting \( x \) and \( y \).

The second condition, also known as the first-order condition, says that any tangent line to \( f \) must lie below the entire function. The third condition intuitively says that if \( y \) is greater than \( x \), then the derivative of \( y \) is also greater than the derivative of \( x \). Finally, the last condition says that the second derivative of \( f \) is always non-negative.

Why are we interested in convex problems? One reason is that convex functions do not have saddle points or local maxima, so any critical point must be a local minimum. Furthermore, any local
minimum is also a global minimum, so any point that has a zero gradient must be the global minimum.

**Proposition 4.** Let $X$ be a convex set. If $f$ is convex, then any local minimum of $f$ in $X$ is also a global minimum.

**Proof.** See math4ml.

Therefore we can find any point for which the gradient is zero and guarantee that it is the global minimum (this is exactly the case in OLS and Ridge Regression since the objective function is PSD and therefore convex).

## Strong Convexity

For a strictly positive $m \in \mathbb{R}$, a function is **$m$-strongly convex** if the following equivalent conditions hold:

1. $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{t(1-t)m}{2} ||y-x||^2$, $\forall x, y, t \in [0, 1]$  
2. $g(x) = f(x) - \frac{m}{2} ||x||^2$ is convex  
3. $f(y) \geq f(x) + \nabla f(x)^\top(y-x) + \frac{m}{2} ||y-x||^2$, $\forall x, y$  
4. $(\nabla f(y) - \nabla f(x))^\top(y-x) \geq m||y-x||^2$, $\forall x, y$  
5. $\nabla^2 f(x) \succeq mI$, $\forall x$

These conditions are the same as those for convex functions, with the exception of an extra term involving $m$. Strongly convex functions provide several advantages over general convex functions. Intuitively, strongly convex functions can be lower bounded by a quadratic function, which establishes the uniqueness of a global minimum.

**Proposition 5.** Let $X$ be a convex set. If $f$ is strictly convex, then there exists at most one local minimum of $f$ in $X$. Consequently, if it exists it is the unique global minimum of $f$ in $X$.

**Proof.** See math4ml.

If the Hessian of $\nabla^2 f$ has eigenvalues that are all strictly positive at all points, $f$ is an $m$-strongly convex with $m$ equal to the the smallest eigenvalue of $\nabla^2 f$ (over all points $x$). Recall from our discussion of OLS vs. Ridge Regression that while OLS may have several solutions, Ridge Regression has a unique solution. This is because the Ridge Regression formulation is positive definite and thus strongly convex, while OLS is positive semi-definite and thus not necessarily strongly convex.

## Smoothness

While strongly convex functions are lower bounded by a quadratic function, smooth functions are upper bounded by a quadratic function. An $M$-smooth or more formally **Lipschitz continuous gradient** function is one for which there exists a strictly positive $M \in \mathbb{R}$ such that

$$||\nabla f(y) - \nabla f(x)|| \leq M||y-x||, \quad \forall x, y$$ (4.1)
Note that this definition does not assume that \( f \) is convex. 4.1 implies all of the following equivalent conditions:

(i) \( f(tx + (1-t)y) \geq tf(x) + (1-t)f(y) - \frac{t(1-t)M}{2} \|y - x\|^2, \quad \forall x, y, t \in [0, 1] \)

(ii) \( f(y) \leq f(x) + \nabla f(x)^\top(y - x) + \frac{M}{2} \|y - x\|^2, \quad \forall x, y \)

(iii) \( (\nabla f(y) - \nabla f(x))^\top(y - x) \leq M \|y - x\|^2, \quad \forall x, y \)

(iv) \( \nabla^2 f(x) \preceq M I, \quad \forall x \)

When \( f \) is convex, then the above conditions also imply 4.1, establishing equivalence among all of the conditions. Roughly speaking, smoothness is the same as strong convexity, except with the inequality signs flipped. If the Hessian of \( \nabla^2 f \) has eigenvalues that are bounded from above, \( f \) is an \( M \)-smooth with \( M \) equal to the the maximum eigenvalue of \( \nabla^2 f \) (over all points \( x \)). Together, strong convexity and smoothness will give us lower and upper bounds, allowing us to achieve a significantly faster convergence rate for many optimization algorithms.

### 4.4 Gradient Descent Methods

**Gradient Descent** is an iterative algorithm uses gradient information to determine a descent direction and takes arbitrary small steps in that direction. For the purposes of our discussion, let’s assume our optimization problem is unconstrained over \( w \in \mathbb{R}^d \):

\[
\min_{w \in \mathbb{R}^d} f(w)
\]

**Gradient Descent**

In its simplest form, gradient descent takes steps in the direction of *steepest descent*. Suppose that we are currently at a point \( w_t \) in the domain of the function. The direction of steepest descent at \( w_t \) is defined as the negative of the gradient at that point, \( -\nabla f(w_t) \). To see why, recall that the directional derivative in a unit direction \( u \) at \( w_t \) is defined as the inner product of the gradient and the direction:

\[
D_u f(w_t) = \langle \nabla f(w_t), u \rangle = \|\nabla f(w_t)\| \cdot \|u\| \cdot \cos(\theta)
\]

where \( \theta \) is the angle between \( \nabla f(w_t) \) and \( u \). We can minimize the directional derivative by setting \( \theta = -\pi \), which will mean that the direction \( u \) and \( \nabla f(w_t) \) are opposite to each other, and thus the optimal direction \( u \) is in the direction opposite that of the gradient. Hence, the direction of steepest descent is \( -\nabla f(w_t) \).

**Algorithm 5:** Gradient Descent

<table>
<thead>
<tr>
<th>Initialize ( w_0 ) to a random point</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textbf{while} Not converged \textbf{do}</td>
</tr>
<tr>
<td>\textbf{_} ( w_{t+1} \leftarrow w_t - \alpha_t \nabla f(w_t) )</td>
</tr>
</tbody>
</table>

The gradient descent algorithm will take an arbitrary step in this direction, scaling the gradient by a scalar \( \alpha_t \). Determining this scaling \( \alpha_t \) is dependent on the attributes of the function \( f \). Sometimes we can set the scaling to a constant value and converge to the optimum value, whereas in other
instances we need to determine an *adaptive stepsize*. A scaling that is too high may cause the algorithm to diverge from the optimal solution, whereas a scaling that is too low may cause the algorithm to converge too slowly. In practice, the stepsize is a hyperparameter and can be tuned by either playing around with values or even using cross validation. For certain classes of functions, there are theoretical guarantees that establish convergence. Assuming that the distance from the initial point $w_0$ and the optimal point $w^*$ is $R$, we have the following:

<table>
<thead>
<tr>
<th>properties of $f$</th>
<th>stepsize $\alpha_t$</th>
<th>convergence rate to $f(w^*)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>convex, $L$-Lipschitz</td>
<td>$\frac{R}{L\sqrt{t}}$</td>
<td>$O\left(\frac{1}{\sqrt{t}}\right)$</td>
</tr>
<tr>
<td>$m$-strongly convex, $L$-Lipschitz</td>
<td>$\frac{2}{m(t+1)}$</td>
<td>$O\left(\frac{1}{t}\right)$</td>
</tr>
<tr>
<td>convex, $M$-smooth</td>
<td>$\frac{M}{t}$</td>
<td>$O\left(\frac{1}{t}\right)$</td>
</tr>
<tr>
<td>$m$-strongly convex, $M$-smooth</td>
<td>$\frac{M}{m}$</td>
<td>$O\left(\exp\left(-\frac{m}{M}t\right)\right)$</td>
</tr>
</tbody>
</table>

For detailed proofs of rates above, refer to the EE 227C lecture notes. Individually, strong convexity and smoothness will allow us to accelerate the rate of convergence from $O\left(\frac{1}{\sqrt{t}}\right)$ to $O\left(\frac{1}{t}\right)$. Put together, they allow us to achieve an exponential convergence rate — a significant acceleration! The quantity $\kappa = \frac{M}{m}$ is known as the **condition number** — the ratio of the largest over smallest singular value of the Hessian of $f$. Recall from our discussion of OLS vs. Ridge Regression that Ridge Regression adds a small penalty term $\lambda||w||^2$ to the objective, effectively making the problem strongly convex. Since the OLS is already smooth as well, then gradient descent can achieve an exponential rate of convergence to the optimal value. The higher the value of $\lambda$, the lower the value of the condition number $\kappa$, which leads to an even faster convergence rate.

**Gradient Descent with Momentum**

**Algorithm 6**: Gradient Descent with Momentum

Initialize $w_0$ to a random point

while Not converged do

\[ w_{t+1} \leftarrow w_t - \alpha_t \nabla f(w_t) + \beta_t (f(w_t) - f(w_{t-1})) \]

**Stochastic Gradient Descent**

**Algorithm 7**: Stochastic Gradient Descent

Initialize $w_0$ to a random point

while Not converged do

Sample a random index $i_t$

\[ w_{t+1} \leftarrow w_t - \alpha_t \nabla f_{i_t}(w_t) \]
4.5 Line Search Methods

Exact Line Search

Backtracking Line Search

Coordinate Descent

Random Search

4.6 Newton’s Method

4.7 Neural Networks

Neural networks are a class of compositional function approximators. They come in a variety of shapes and sizes. In this class, we will only discuss feedforward neural networks, those networks whose computations can be modeled by a directed acyclic graph.\(^1\) The most basic (but still commonly used) class of feedforward neural networks is the multilayer perceptron. Such a network might be drawn as follows:

Computation flows left-to-right. The circles represent nodes, a.k.a. units or neurons, which are loosely based on the behavior of actual neurons in the brain. Observe that the nodes are organized into layers. The first (left-most) layer is called the input layer, the last (right-most) layer is called the output layer, and any other layers (there is only one here, but there could be multiple) are referred to as hidden layers. The dimensionality of the input and output layers is determined by the function we want the network to compute. For a function from \(\mathbb{R}^d\) to \(\mathbb{R}^k\), we should have \(d\) input nodes and \(k\) output nodes. The number and sizes of the hidden layers are hyperparameters to be chosen by the network designer.

\(^1\) There are also recurrent neural networks whose computation graphs have cycles.
Note that in the diagram above, each non-input layer has the following property: every node in that layer is connected to every node in the previous layer. Layers that have this property are described as fully connected. Each edge in the graph has an associated weight, which is the strength of the connection from the input node in one layer to the node in the next layer. Each node computes a weighted sum of its inputs, with these connection strengths being the weights, and then applies a nonlinear function which is variously referred to as the activation function or the nonlinearity. Concretely, if \( w_i \) denotes the weights and \( \sigma_i \) denotes the activation function of node \( i \), it computes the function

\[
x \mapsto \sigma_i(w_i^T x)
\]

Let us denote the number of (non-input) layers by \( L \), the number of units in layer \( \ell \in \{0, \ldots, L\} \) by \( n_\ell \) (here \( n_0 \) is the size of the input layer), and the nonlinearity for layer \( \ell \in \{1, \ldots, L\} \) by \( \sigma_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_\ell} \). The weights for every node in layer \( \ell \) can be stacked (as rows) into a matrix of weights \( W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}} \). Then layer \( \ell \) performs the computation

\[
x \mapsto \sigma_\ell(W_\ell x)
\]

Since the output of each layer is passed as input to the next layer, the function represented by the entire network can be written

\[
x \mapsto \sigma_L(W_L \sigma_{L-1}(\cdots \sigma_2(W_2 \sigma_1(W_1 x)) \cdots))
\]

This is what we mean when we describe neural networks as compositional.

Note that in most layers, the nonlinearity will be the same for each node within that layer, so it makes sense to refer to a scalar function \( \sigma_\ell : \mathbb{R} \to \mathbb{R} \) as “the” nonlinearity for layer \( \ell \), and apply it element-wise:

\[
\sigma_\ell(x) = \begin{bmatrix}
\sigma_\ell(x_1) \\
\vdots \\
\sigma_\ell(x_{n_\ell})
\end{bmatrix}
\]

The principle exception here is the softmax function \( \sigma : \mathbb{R}^k \to \mathbb{R}^k \) defined by

\[
\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^k e^{x_j}}
\]

which is often used to produce a discrete probability distribution over \( k \) classes. Note that every entry of the softmax output depends on every entry of the input. Also, softmax preserves ordering, in the sense that sorting the indices \( i = 1, \ldots, k \) by the resulting value \( \sigma(x)_i \) yields the same ordering as sorting by the input value \( x_i \). In other words, more positive \( x_i \) leads to larger \( \sigma(x)_i \).

This nonlinearity is used most commonly (but not always) at the output layer of the network.

**Expressive power**

It is the repeated combination of nonlinearities that gives deep neural networks their remarkable expressive power. Consider what happens when we remove the activation functions (or equivalently, set them to the identity function): the function computed by the network is

\[
x \mapsto \underbrace{W_L W_{L-1} \cdots W_2 W_1}_{= : \hat{W}} x
\]
which is linear in its input! Moreover, the size of the smallest layer restricts the rank of $\tilde{W}$, as

$$\text{rank}(\tilde{W}) \leq \min_{\ell \in \{1, \ldots, L\}} \text{rank}(W_\ell) \leq \min_{\ell \in \{0, \ldots, L\}} n_\ell$$

Despite having many layers of computation, this class of networks is not very expressive; it can only represent linear functions.

We would like to produce a class of networks that are \textbf{universal function approximators}. This essentially means that given any continuous function, we can choose a network in this class such that the output of the circuit can be made arbitrarily close to the output of the given function for all given inputs. We make a more precise statement later.

A key observation is that piecewise-constant functions are universal function approximators:

![Graph of piecewise-constant function]

The nonlinearity we use, then, is the step function:

$$\sigma(x) = \begin{cases} 
1 & x \geq 0 \\
0 & x < 0 
\end{cases}$$

We can build very complicated functions from this simple step function by combining translated and scaled versions of it. Observe that

- If $a, b \in \mathbb{R}$, the function $x \mapsto \sigma(a + bx)$ is a translated (and, depending on the sign of $b$, possibly flipped) version of the step function:
• If \( c \neq 0 \), the function \( x \mapsto c\sigma(x) \) is a vertically scaled version of the step function.

It turns out that only one hidden layer is needed for universal approximation, and for simplicity we assume a one-dimensional input. Thus from here on we consider networks with the following structure:

The input \( x \) is one-dimensional, and the weight on \( x \) to node \( j \) is \( b_j \). We also introduce a constant 1, whose weight into node \( j \) is \( a_j \). (This is referred to as the bias, but it has nothing to do with bias in the sense of the bias-variance tradeoff. It’s just there to provide the node with the ability to shift its input.) The function implemented by the network is

\[
h(x) = \sum_{j=1}^{k} c_j \sigma(a_j + b_j x)
\]

where \( k \) is the number of hidden units.

**Choosing weights**

With a proper choice of \( a_j, b_j, \) and \( c_j \), this function can approximate any continuous function we want. But the question remains: given some target function, how do we choose these parameters in the appropriate way?
Let’s try a familiar technique: least squares. Assume we have training data \( \{(x_i, y_i)\}_{i=1}^{n} \). We aim to solve the optimization problem

\[
\min_{a, b, c} \sum_{i=1}^{n} (y_i - h(x_i))^2
\]

To run gradient descent, we need derivatives of the loss with respect to our optimization variables. We compute via the chain rule

\[
\frac{\partial (y_i - h(x_i))^2}{\partial c_j} = -2(y_i - h(x_i)) \frac{\partial h(x_i)}{\partial c_j} = 2(y_i - h(x_i)) \sigma(a_j + b_j x_i)
\]

We see that if this particular step is “off”, as in \( \sigma(a_j + b_j x_i) = 0 \), then

\[
\frac{\partial (y_i - h(x_i))^2}{\partial c_j} = 2(y_i - h(x_i)) \sigma(a_j + b_j x_i) = 0
\]

so no update will be made for that example. More egregiously, consider the derivatives with respect to \( a_j \):\(^3\)

\[
\frac{\partial f}{\partial a_j} = \sum_{i=1}^{n} -2(y_i - h(x_i)) \frac{\partial h(x_i)}{\partial a_j} = 0
\]

and \( b_j \):

\[
\frac{\partial f}{\partial b_j} = \sum_{i=1}^{n} -2(y_i - h(x_i)) \frac{\partial h(x_i)}{\partial b_j} = 0
\]

Since gradient descent changes weights in proportion to their gradient, it will never modify \( a \) or \( b \)!

Even though the step function is useful for the purpose of showing the approximation capabilities of neural networks, it is seldom used in practice because it cannot be trained by conventional gradient-based methods.

The next simplest universal approximator is the class of **piecewise-linear functions**. Just as piecewise-constant functions can be achieved by combinations of the step function as a nonlinearity, piecewise-linear functions can be achieved by combinations of the **rectified linear unit** (ReLU) function

\[
\sigma(x) = \max\{0, x\}
\]

\(^3\) Technically, the derivative of \( \sigma \) is not defined at zero, where there is a discontinuity. However it is defined (and zero) everywhere else. In practice, we will almost never hit the point of discontinuity because it is a set of measure zero.
4.7. NEURAL NETWORKS

Depending on the weights $a$ and $b$, our ReLUs can move to the left or right, increase or decrease their slope, and flip direction.

Let us calculate the gradients again, assuming we replace the step functions by ReLUs:

\[
\frac{\partial f}{\partial c_j} = \sum_{i=1}^{n} -2(y_i - h(x_i)) \max\{0, a_j + b_j x_i\}
\]

\[
\frac{\partial f}{\partial a_j} = \sum_{i=1}^{n} -2(y_i - h(x_i)) c_j \frac{\partial}{\partial a_j} \max\{0, a_j + b_j x_i\} = \sum_{k=1}^{n} -2(y_i - h(x_i)) c_j \begin{cases} 
0 & \text{if } a_j + b_j x_i < 0 \\
1 & \text{if } a_j + b_j x_i > 0
\end{cases}
\]

\[
\frac{\partial f}{\partial b_j} = \sum_{i=1}^{n} -2(y_i - h(x_i)) c_j \frac{\partial}{\partial b_j} \max\{0, a_j + b_j x_i\} = \sum_{i=1}^{n} -2(y_i - h(x_i)) c_j \begin{cases} 
0 & \text{if } a_j + b_j x_i < 0 \\
x_i & \text{if } a_j + b_j x_i > 0
\end{cases}
\]

Crucially, we see that the gradient with respect to $a$ and $b$ is not uniformly zero, unlike with the step function.

Later we will discuss \textbf{backpropagation}, a dynamic programming algorithm for efficiently computing gradients with respect to a neural network’s parameters.

\textbf{Neural networks are universal function approximators}

The celebrated neural network universal approximation theorem, due to Kurt Hornik\footnote{See Approximation Capabilities of Multilayer Feedforward Networks.}, tells us that neural networks are universal function approximators in the following sense.

\textbf{Theorem.} Suppose $\sigma : \mathbb{R} \to \mathbb{R}$ is nonconstant, bounded, nondecreasing, and continuous\footnote{Both ReLU and sigmoid satisfy these requirements.}, and let $S \subseteq \mathbb{R}^d$ be closed and bounded. Then for any continuous function $f : S \to \mathbb{R}$ and any $\epsilon > 0$, there exists a neural network with one hidden layer containing finitely many nodes, which we can write

\[
h(x) = \sum_{j=1}^{k} c_j \sigma(a_j + b_j^\top x)
\]

such that

\[
|h(x) - f(x)| < \epsilon
\]
for all $x \in S$.

There’s some subtlety in the theorem that’s worth noting. It says that for any given continuous function, there exists a neural network of finite size that uniformly approximates the given function. However, it says nothing about how well any particular architecture you’re considering will approximate the function. It also doesn’t tell us how to compute the weights.

It’s also worth pointing out that in the theorem, the network consists of just one hidden layer. In practice, people find that using more layers works better.

### 4.8 Training Neural Networks

**ReLUs as Universal Function Approximators**

Last time we saw that the second-most simple universal function approximator was the piecewise linear function. Specifically, we talked about a specific component of piecewise linear functions called the ReLU, which is defined as $f(x) = \max(0, x)$.

In our discussion of neural nets, we saw that we would have the ReLUs act on linear combinations of neural net units to introduce nonlinearity to the hypothesis function encoded by the neural net. For example, when acting on one input (and a bias term) our ReLUs will take in arguments of the form $a + bx$. Let’s see an example of how expressive they can be. Suppose we wanted to build this function from ReLUs:

\[
\begin{align*}
  f(x) &= x - 2 \\
  f(x) &= 6 - x \\
  f(x) &= 2x - 12 \\
  f(x) &= -1 + \max(0, x - 2) - \max(0, x - 3) + \max(0, 6 - x) - \max(0, 5 - x) + \max(0, 2x - 12)
\end{align*}
\]

Here’s the plot on Wolfram Alpha.

In higher dimensions, i.e. when a ReLU takes in an arbitrarily long dot-product as input: $f(x) =$
4.8. TRAINING NEURAL NETWORKS

Figure 4.3: $f(x_1, x_2) = \max(0, 2x_1 + 3x_2)$

$\max(0, w^\top x)$, the unit can be viewed as representative of a “ramp” in the higher-dimensional space. Here’s a plot of a 3-D ramp:

**Derivatives in Neural Nets**

We have a very powerful tool at our disposal in neural nets, but it does us no good if we can’t train them. Let’s talk about how this happens. The output units of a neural net can be thought of as akin to a regression or hypothesis function determined by the parameters of some model. For example, in ordinary least-squares regression, we learned a hypothesis function $f(x) = w^\top x$ determined by the parameter $w$. It is just the same with neural nets, except that our hypothesis function can be arbitrarily complex. Consider the following neural net:

The hypothesis function that this neural net encodes is represented by the two outputs, $O = [O_1, O_2]$.

Since neural net outputs are not linear functions of their inputs, there is no closed-form solution for the minimum to any reasonable loss function defined on them. Thus, we resort to using gradient
To run gradient descent, we must calculate the derivative of the loss function with respect to the parameters of the neural net. In the network depicted above, the parameters are \( W = [w_1, w_2, w_3], \ a, \) and \( V = [v_1, v_2]. \) These are the values of the model which we are allowed to tweak. **Backpropagation** is an efficient manner of computing these gradients that exploits the nested structure of neural nets.

**The Chain Rule**

This section is an aside meant to recall your knowledge about the chain rule in multivariable calculus. Let’s take a look at two slices of our neural net from above.

If you want to compute the derivatives of upstream stuff with respect to the weights on these connections, you need only consider the input of a single connection at a time. That is to say:

\[
\frac{\partial L}{\partial w_1} = \text{upstream terms} \cdot \frac{\partial H_1}{\partial w_1}
\]

completely independent of \( x_2 \) and \( x_3. \)

If you want to compute the derivatives of upstream stuff with respect to the weights downstream of these connections, you’ll need to sum over the contributions of the inputs to these connections.
That is to say:

\[
\frac{\partial L}{\partial a} = \sum_i \text{upstream terms}_i \cdot \frac{\partial O_i}{\partial a}
\]

**Backpropagation**

A naive way of calculating the gradients might be to differentiate the loss with respect to each parameter we’re interested in updating, one at a time. However, because the later layers of a neural net are just functions of the earlier layers, doing this would be wasteful. We can see this by taking a look at the derivatives of \(L\) with respect to \(v_1\), \(a\), and \(w_1\) in our example:

\[
\frac{\partial L}{\partial v_1} = \frac{\partial L}{\partial O_1} \frac{\partial O_1}{\partial v_1} \\
\frac{\partial L}{\partial a} = \frac{\partial L}{\partial O_1} \frac{\partial O_1}{\partial a} + \frac{\partial L}{\partial O_2} \frac{\partial O_2}{\partial a} \\
\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial O_1} \frac{\partial O_1}{\partial w_1} + \frac{\partial L}{\partial O_2} \frac{\partial O_2}{\partial w_1}
\]

You should notice that by invoking the chain rule, we see that the term \(\frac{\partial L}{\partial O_1}\) is common to all three derivatives, and the term \(\frac{\partial L}{\partial O_2}\) is common to the second two. This suggests that we should consider caching the results of the derivatives of weights in the later layers and reuse them in our computation of the derivatives of the weights of earlier layers: a dynamic programming approach.

The following is a general outline of how backpropagation might be implemented. Keep in mind that the specifics, especially those pertaining to the structure of each successive layer, will depend heavily on the architecture of the neural net in question.

1. The **forward pass**: populate each unit of the neural net with the value it’s supposed to have (i.e., invoke all your dot-products and ReLUs).

2. Start at the upstream end, i.e. the outputs. Compute the gradient of the loss function with respect to the outputs (these are just numbers), and memoize/cache them.

3. Go back one layer. Now, treat your outputs \(O_i\) as endpoints of your neural net, and compute the gradients w.r.t. the previous layer, caching these as well. The contributions to the final loss should be summed appropriately over the paths through which they influence the loss.

4. Repeat until you hit the last layer (the inputs). You should now have all the necessary components to compute the derivative of the loss function with respect to any parameter of the neural net.

### Speeding up Gradient Descent

Backpropagation efficiently computes gradients, so we can run gradient descent to optimize neural nets. However, computing the full gradient may be expensive, particularly if we have a lot of data. Consider that the loss function in machine learning typically is an average (or sum, but this is the same up to a constant factor) of errors over the training points:

\[
L(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i, w), y_i)
\]
By the linearity of differentiation, we easily see that
\[
\nabla L(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla w \ell(h(x_i, w), y_i)
\]
So computing the gradient takes time in linear in \(n\). In the “big data” regime, \(n\) is very large, so this cost may be unacceptable.

For this reason, we have reason to try to approximate the gradient of the loss function of a neural net by taking a representative random sample of the inputs to calculate the gradient over. Since gradient descent is an iterative process, we might consider forfeiting the exactness of the update of each individual iteration in exchange for better speed of computation. If we take the gradient over a random sample of these training samples at each step, we will get a noisy but unbiased estimate of the true gradient. The noise in each step is often an acceptable tradeoff in exchange for the opportunity to take many more steps.

In the regular gradient descent update, we have the policy:
\[
w^{(k+1)} \leftarrow w^{(k)} - \alpha_k \nabla w L(w^{(k)})
\]
In stochastic gradient descent, we have the update rule:
\[
w^{(k+1)} \leftarrow w^{(k)} - \alpha_k G_k
\]
where \(G_k\) is a random variable which satisfies \(\mathbb{E}[G_k] = \nabla L(w^{(k)})\). Typically \(G_k\) is constructed by sampling \(m < n\) training points uniformly at random (say the index is \(I_k \subset \{1, \ldots, n\}\)) and computing the gradient on these points only:
\[
G_k = \frac{1}{m} \sum_{i \in I_k} \nabla w \ell(h(x_i), y_i)
\]
We denote the true optimum by \(w^*\). Our goal is to show the following:
\[
\lim_{k \to \infty} \mathbb{E}[[w^{(k)} - w^*]^2] = 0
\]
We make the following assumptions in the analysis below:

1. The loss function \(f\) is \(\ell\)-strongly convex; that is, there exists a constant \(\ell > 0\) such that \((\nabla f(x) - \nabla f(y))^\top(x - y) \geq \ell \|x - y\|^2\) for all \(x, y\).
2. The expected squared norm of the stochastic gradient is bounded as \(\mathbb{E}[\|G_k\|^2] \leq M^2 < \infty\).

We begin by expanding the desired term:
\[
\mathbb{E}[[w^{(k+1)} - w^*]^2]
\]
\[
= \mathbb{E}[(w^{(k+1)} - w^*)^\top(w^{(k+1)} - w^*)]
\]
\[
= \mathbb{E}[(w^{(k)} - w^* - \alpha_k G_k)^\top(w^{(k)} - w^* - \alpha_k G_k)]
\]
\[
= \mathbb{E}[[w^{(k)} - w^*]^2] - 2\alpha_k \mathbb{E}[(w^{(k)} - w^*)^\top G_k] + \alpha_k^2 \mathbb{E}[\|G_k\|^2]
\]
For brevity, define
\[
d_k = \mathbb{E}[[w^{(k)} - w^*]^2]
\]
Our assumption on the expected squared norm of \(G_k\) implies
\[
d_{k+1} \leq d_k - 2\alpha_k \mathbb{E}[(w^{(k)} - w^*)^\top G_k] + \alpha_k^2 M^2
\]
To analyze this recurrence, we expand the first few terms:

\begin{align*}
E[(w^{(k)} - w^*)^\top G_k] &= E_{\text{past}} \left[ E[(w^{(k)} - w^*)^\top G_k \mid \text{past}] \right]
\end{align*}

Here the inner expectation is taken over the choice of training points used to compute the stochastic gradient. But given the past (which includes \(w^{(k)}\)), we already know \(w^{(k)}\), so

\begin{align*}
E[(w^{(k)} - w^*)^\top G_k \mid \text{past}] &= (w^{(k)} - w^*)^\top E[G_k \mid \text{past}]
\end{align*}

The current gradient \(G_k\) depends on our new choice of evaluation point, which is presumed independent of the past and thus \(E[G_k \mid \text{past}] = E[G_k] = \nabla f(w^{(k)})\), where the second equality holds because \(G_k\) is an unbiased estimator for the true gradient at \(w^{(k)}\). Putting all this together, we have

\begin{align*}
E[(w^{(k)} - w^*)^\top G_k] &= E_{\text{past}}[(w^{(k)} - w^*)^\top \nabla f(w^{(k)})]
\end{align*}

First order necessary conditions for optimality imply that \(\nabla f(w^*) = 0\). Then by the assumption of \(\ell\)-strong convexity, we have

\begin{align*}
(w^{(k)} - w^*)^\top \nabla f(w^{(k)}) &= (w^{(k)} - w^*)^\top (\nabla f(w^{(k)}) - \nabla f(w^*)) \\
&\geq \ell \|w^{(k)} - w^*\|_2^2
\end{align*}

Taking expectations yields

\begin{align*}
E[(w^{(k)} - w^*)^\top \nabla f(w^{(k)})] \geq \ell E[\|w^{(k)} - w^*\|_2^2] = \ell d_k
\end{align*}

Putting this back into our inequality for \(d_{k+1}\), we get

\begin{align*}
d_{k+1} \leq d_k - 2\alpha_k \ell d_k + \alpha_k^2 M^2 = (1 - 2\alpha_k \ell)d_k + \alpha_k^2 M^2
\end{align*}

The \(\alpha_k^2 M^2\) term was incurred by the randomness in our updates. We can try to send this term to 0 by diminishing the step size \(\alpha_k\) over time – but decreasing \(\alpha_k\) will also decrease the effect of the \((1 - 2\alpha_k \ell)d_k\) term, so we need to choose our step size carefully.

Setting \(\alpha_k = \frac{1}{2\ell k}\) gives

\begin{align*}
d_{k+1} \leq \left(1 - \frac{1}{k}\right) d_k + \frac{1}{(2\ell k)^2} M^2
\end{align*}

Let \(S = \frac{M^2}{(2\ell)^2}\) so that this inequality becomes

\begin{align*}
d_{k+1} \leq \left(1 - \frac{1}{k}\right) d_k + \frac{1}{k^2} S
\end{align*}

To analyze this recurrence, we expand the first few terms:

\begin{align*}
d_2 &\leq S \\
d_3 &\leq \left(1 - \frac{1}{2}\right) d_2 + \frac{1}{2^2} S = \frac{1}{1 \cdot 2} S + \frac{1}{2^2} S \\
d_4 &\leq \left(1 - \frac{1}{3}\right) d_3 + \frac{1}{3^2} S = \frac{1}{1 \cdot 3} S + \frac{1}{2 \cdot 3} S + \frac{1}{3^2} S \\
d_5 &\leq \left(1 - \frac{1}{4}\right) d_4 + \frac{1}{4^2} S = \frac{1}{1 \cdot 4} S + \frac{1}{2 \cdot 4} S + \frac{1}{3 \cdot 4} S + \frac{1}{4^2} S
\end{align*}
Inductively, we find that

\[ d_n \leq \frac{S}{n-1} \sum_{j=1}^{n-1} \frac{1}{j} \]

Because the harmonic sum \( \sum_{j=1}^{n-1} \frac{1}{j} \) behaves as \( \ln(n) \), we see that \( d_n \) is upper bounded in the limit by \( S \frac{\ln(n)}{n} \to 0 \). Hence \( \lim_{n \to \infty} d_n = 0 \) as desired.
Chapter 5

Classification

5.1 Generative vs. Discriminative Classification

The task of classification differs from regression in that we are now interested in assigning a \( d \)-dimensional data point one of a discrete number of classes, instead of assigning it a continuous value. Thus, the task is simpler in that there are fewer choices of labels per data point but more complicated in that we now need to somehow factor in information about each class to obtain the classifier that we want.

Given a training set \( \mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \) of \( n \) points, where each data point \( x_i \in \mathbb{R}^d \) is paired with a known discrete class label \( y_i \in \{1, 2, ..., K\} \), our goal is to train a classifier which, when fed any arbitrary \( d \)-dimensional data point, classifies that data point as one of the \( K \) discrete classes.

There are two main types of classification models: generative models and discriminative models. **Generative models** have strong roots in probabilistic modeling. The idea is that we form a joint probability distribution \( p(X, Y) \) over the input \( X \) (which we treat as a random vector) and label \( Y \) (which we treat as a random variable), and we classify an arbitrary datapoint \( x \) with the class label that maximizes the joint probability:

\[
\hat{y} = \arg \max_k p(x, Y = k)
\]

Generative models typically form the joint distribution by explicitly forming the following:

- A prior probability distribution over all classes:

\[
P(k) = P(\text{class} = k)
\]

- A conditional probability distribution for each class \( k \in \{1, 2, ..., K\} \):

\[
p_k(X) = p(X|\text{class } k)
\]

Using the prior and the conditional distributions in conjunction, we have (from Bayes’ rule) that maximizing the joint probability over the class labels is equivalent to maximizing the posterior probability of the class label:

\[
\hat{y} = \arg \max_k p(x, Y = k) = \arg \max_k P(k) \ p_k(x) = \arg \max_k P(Y = k|x)
\]
Maximizing the posterior will induce regions in the feature space in which one class has the highest posterior probability, and **decision boundaries** in between classes where the posterior probability of two classes are equal.

![Figure 5.1: A collection (in dark black) of linear (left) vs quadratic (right) level set boundaries in a 2D feature space](image)

Generative classifiers are flexible, quick to train, and can generate new samples (in order to augment the training dataset). However, they are also inefficient, because they require estimation of a large number of parameters (i.e. the covariance matrices of the conditional distributions, which have \( \frac{d(d+1)}{2} \) parameters). Typically, the decision boundary only requires \( O(d) \) parameters, but generative models typically estimate \( O(d^2) \) parameters in order to determine the class-conditional probability distributions. As \( d \) increases, generative models tend to lose their effectiveness, as the number of parameters starts to dominate in comparison to the number of datapoints, and as a result the variance of the model increases.

This leads us to the concept of **discriminative models**, where we bypass learning a generative model altogether and directly learn a decision boundary. Discriminative models are parameterized by weights that either (1) form a posterior distribution \( P(Y|X) \) without considering the prior or conditional distributions, or (2) directly form a hard decision boundary without considering any probabilities in the first place. In the former case, discriminative models choose the class that maximizes the posterior probability distribution:

\[
\hat{y} = \arg \max_k P(Y = k|X)
\]

Generative models also choose the class that maximizes the posterior probability distribution. The only difference is in the way generative and discriminative models form the posterior.

**Bayes’ Decision Rule**

While both generative and discriminative models by default maximize the posterior probability over classes, this strategy may not necessarily be desirable at all times. Rather than maximizing the posterior probability, we would really like to minimize the **risk** of our model. Recall that the risk for a given classifier \( h \) is defined as the expected loss over \( X \) and \( Y \):

\[
R(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(h(x), y)]
\]

where \( \ell(h(x), y) \) measures the loss between the predicted label \( h(x) \) and the true label \( y \). In the context of regression, the loss function was the squared error \( \ell(h(x), y) = (h(x) - y)^2 \). In the
context of classification, the loss function can take many forms, but the simplest is the standard step function

\[ \ell(h(x), y) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases} \]

Our goal is to find a classifier that minimizes the risk, given the loss function. We can equivalently express the risk as

\[ R(h) = \int \left( \sum_{k=1}^{K} L(r(x), k)P(Y = k|x) \right) p(x)dx \]

The Bayes’ classifier \( h^* \) will minimize the risk. Given an arbitrary \( x \), the Bayes’ classifier will pick

\[ h^*(x) = \arg \min_j \sum_{k=1}^{K} L(j, k)P(Y = k|x) \]

Effectively, the Bayes’ classifier will pick the class that minimizes the expected loss for the given \( x \). In the special case where the loss function is the standard step function (as described above),

\[ h^*(x) = \arg \min_j \sum_{k\neq j} P(Y = k|x) = \arg \min_j 1 - P(Y = j|x) = \arg \max_j P(Y = j|x) \]

This is equivalent to selecting the class that maximizes the posterior distribution!

Depending on which loss function we are using, the optimal classifier may or may not maximize the posterior probability. For example, consider the case of cancer diagnosis, where a patient’s diagnosis for cancer can come up as positive or negative. There are four possible cases:

1. Classify the patient cancer +, and in reality the patient is cancer + (Correct Classification)
2. Classify the patient cancer −, and in reality the patient is cancer − (Correct Classification)
3. Classify the patient cancer +, but in reality the patient is cancer − (False positive)
4. Classify the patient cancer −, but in reality the patient is cancer + (False negative)

Classifying the patient’s condition correctly is ideal, so we can reasonably set the loss for those cases to 0. The false positive and false negative cases are bad, and there should be a loss for these cases. But should these cases have the same loss value or should we weigh them differently? A false negative diagnosis would be significantly worse than a false positive, because a false negative diagnosis would go undiagnosed and would probably be fatal. Therefore, the associated loss for the false negative case should be higher than the associated loss for the false positive case. In this case, the goal is no longer to maximize the posterior probability, because otherwise we would be treating the false negative and false positive cases the same.

5.2 Least Squares Support Vector Machine

As a first example of a simple, non-probabilistic discriminative model, we discuss the Least Squares Support Vector Machine (LS-SVM). Consider the binary classification problem where the classes are represented by \(-1\) and \(+1\). One way to classify a data point \( x \) is to estimate parameters \( w \), compute \( w^T x \), and classify \( x \) to be \( \text{sign}(w^T x) \). Geometrically, the decision boundary this produces is a hyperplane, \( w^T x = 0 \).
We need to figure out how to optimize the parameter \( w \). One simple procedure we can try is to fit a least squares objective:

\[
\arg \min_{w} \sum_{i=1}^{n} \| y_i - \text{sign}(w^T x_i) \|^2 + \lambda \| w \|^2
\]

Where \( x_i, w \in \mathbb{R}^{d+1} \). Note that we have not forgotten about the bias term! Even though we are dealing with \( d \) dimensional data, \( x_i \) and \( w \) are \( d + 1 \) dimensional: we add an extra “feature” of 1 to \( x \), and a corresponding bias term of \( k \) in \( w \). Note that in practice, we do not want to penalize the bias term in the regularization term, because the we should be able to work with any affine transformation of the data and still end up with the same decision boundary. Therefore, rather than taking the norm of \( w \), we often take the norm of \( w' \), which is every term of \( w \) excluding the corresponding bias term. For simplicity of notation however, let’s just take the norm of \( w \).

Without the regularization term, this would be equivalent to minimizing the number of misclassified training points. Unfortunately, the “sign” term makes this optimization problem non-convex, and in fact this optimization problem is NP-hard (computationally intractable). Instead we can solve a relaxed version of this problem:

\[
\arg \min_{w} \sum_{i=1}^{n} \| y_i - w^T x_i \|^2 + \lambda \| w \|^2
\]

This method is called the binary least squares support vector machine (LS-SVM). Note that in this relaxed version, we care about the magnitude of \( w^T x \) and not just the sign.

One drawback of LS-SVM is that the hyperplane decision boundary it computes does not necessarily make sense for the sake of classification. For example, consider the following set of data points, color-coded according to the class:

![Figure 5.2: Reasonable fit LS-SVM](image)

LS-SVM will classify every data point correctly, since all the +1 points lie on one side of the decision boundary and all the -1 points lie on the other side. Now if we add another cluster of points as follows:
5.2. LEAST SQUARES SUPPORT VECTOR MACHINE

The original LS-SVM fit would still have classified every point correctly, but now the LS-SVM gets confused and decides that the points at the bottom right are contributing too much to the loss (perhaps for these points, $\mathbf{w}^T \mathbf{x}_i = -5$ for the original choice of $\mathbf{w}$ so even though they are on the correct side of the original separating hyperplane, they incur a high squared loss and thus the hyperplane is shifted to accommodate). This problem will be solved when we introduce **general Support Vector Machines (SVM's)**.

**Feature Extension**

Working with linear classifiers in the raw feature space may be extremely limiting, so we may consider adding features that allow us to come up with nonlinear classifiers (note that we are still working with linear classifiers in the augmented feature space). For example, adding quadratic features allows us to find a linear decision boundary in the augmented quadratic space that corresponds to a nonlinear “circle” decision boundary projected down into the raw feature space.

Figure 5.4: Augmenting Features, image courtesy of Prof. Shewchuk
In order to implement this idea, we re-express our objective as
\[
\arg\min_w \sum_{i=1}^n \| y_i - w^\top \phi(x_i) \|^2 + \lambda \| w \|^2
\]
Note that \( \phi \) is a function that takes as input the data in raw feature space, and outputs the data in augmented feature space.

**Neural Network Extension**

Instead of using the linear function \( w^\top x \) or augmenting features to the data, we can also directly use a non-linear function of our choice in the original feature space, such as a neural network. One can imagine a whole family of discriminative binary classifiers that minimize
\[
\arg\min_w \sum_{i=1}^n \| y_i - g_w(x_i) \|^2 + \lambda \| w \|^2
\]
where \( g_w(x_i) \) can be any function that is easy to optimize. Then we can classify using the rule
\[
\hat{y}_i = \begin{cases} 
1 & g_w(x_i) > \theta \\
-1 & g_w(x_i) \leq \theta
\end{cases}
\]
Where \( \theta \) is some threshold. In LS-SVM, \( g_w(x_i) = x^\top w_i \) and \( \theta = 0 \). Using a neural network with non-linearities as \( g_w \) can produce complex, non-linear decision boundaries.

**Multiclass Extension**

We can also adapt this approach to the case where we have multiple classes. Suppose there are \( K \) classes, labeled 1, 2, ..., \( K \). One possible way to extend the approach from binary classification is to compute \( g_w(x_i) \) and round it to the nearest number from 1 to \( K \). However, this approach gives an “ordering” to the classes, even if the classes themselves have no natural ordering. This is clearly a problem. For example, in fruit classification, suppose 1 is used to represent “peach,” 2 is used to represent “banana,” and 3 is used to represent “apple.” In our numerical representation, it would appear that peaches are less than bananas, which are less than apples. As a result, if we have an image that looks like some cross between an apple and a peach, we may simply end up classifying it as a banana.

The typical way to get around this issue is as follows: if the \( i \)’th observation has class \( k \), instead of using the representation \( y_i = k \), we can use the representation \( y_i = e_k \), the \( k \)'th canonical basis vector. Now there is no relative ordering in the representations of the classes. This method is called one-hot vector encoding.

When we have multiple classes, each \( y_i \) is a \( K \)-dimensional one-hot vector, so for LS-SVM, we instead have a \( K \times (d + 1) \) weight matrix to optimize over:
\[
\arg\min_W \sum_{i=1}^n \| y_i - Wx_i \|^2 + \lambda \| W \|^2
\]
To classify an arbitrary input \( x \), we compute \( Wx \) and see which component \( k \) is the largest:
\[
\hat{y} = \max_k w_k^\top x
\]
5.3 Logistic Regression

**Logistic regression** is a discriminative classification technique that has a direct probabilistic interpretation. We will first present the binary class case, and then we can easily extend the logic to the multiclass case.

**Binary Logistic Regression**

Suppose that we have the binary classification problem where classes are represented by 0 and 1. Note that we instead of using $-1/1$ labels (as in LS-SVM), in binary logistic regression we use 0/1 labels. Logistic regression makes more sense this way because it directly outputs a probability, which belongs in the range of values between 0 and 1.

In binary logistic regression, we would like our model to output the probability that a data point is in class 0 or 1. We can start with the raw linear “score” $w^\top x$ and convert it to a probability between 0 and 1 by applying a sigmoid transformation $s(w^\top x)$, where $s(z) = \frac{1}{1+e^{-z}}$. To classify an arbitrary point $x$, we use the sigmoid function to output a probability distribution $P(\hat{Y})$ over the classes 0 and 1:

$$P(\hat{Y} = 1 \mid x, w) = s(w^\top x), \quad P(\hat{Y} = 0 \mid x, w) = 1 - s(w^\top x)$$

we classify $x$ as the class with the maximum probability:

$$\hat{y} = \max_k P(\hat{Y} = k \mid x, w) = \begin{cases} 1 & \text{if } s(w^\top x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

![Logistic function](image)

Figure 5.5: Logistic function. For our purposes, the horizontal axis is the output of the linear function $w^\top x_i$ and the vertical axis is the output of the logistic function, which can be interpreted as a probability between 0 and 1.

Equivalently, we classify $x$ as

$$\hat{y} = \begin{cases} 1 & \text{if } w^\top x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$
Loss Function

Suppose we are given a training dataset \( D = \{(x_i, y_i)\}_{i=1}^n \). In order to train our model, we need a loss function to optimize. One possibility is least squares:

\[
\arg \min_w \sum_{i=1}^n \|y_i - s(w^\top x_i)\|^2 + \lambda\|w\|^2
\]

However, this may not be the best choice. Ordinary least squares regression has theoretical justifications such as being the maximum likelihood objective under Gaussian noise. Least squares for this classification problem does not have a similar justification.

Instead, the loss function we use for logistic regression is called the log-loss, or cross entropy:

\[
L(w) = \sum_{i=1}^n y_i \ln \left( \frac{1}{s(w^\top x_i)} \right) + (1 - y_i) \ln \left( \frac{1}{1 - s(w^\top x_i)} \right)
\]

If we define \( p_i = s(w^\top x_i) \), then using the properties of logs we can express this as

\[
L(w) = -\sum_{i=1}^n y_i \ln p_i + (1 - y_i) \ln(1 - p_i)
\]

For each \( x_i \), \( p_i \) represents our predicted probability that its corresponding class is 1. Because \( y_i \in \{0, 1\} \), the loss corresponding to the \( i \)'th data point is

\[
L_i(w) = \begin{cases} 
-\ln p_i & \text{when } y_i = 1 \\
-\ln(1 - p_i) & \text{when } y_i = 0 
\end{cases}
\]

Intuitively, if \( p_i = y_i \), then we incur 0 loss. However, this is never actually the case. The logistic function can never actually output a value of exactly 0 or 1, and we will therefore always incur some loss. If the actual label is \( y_i = 1 \), then as we lower \( p_i \) towards 0, the loss for this data point approaches infinity.

The loss function can be derived from a maximum likelihood perspective or an information-theoretic perspective. First let’s present the maximum likelihood perspective. We view each observations \( y_i \) as an independent sample from a Bernoulli distribution \( \hat{Y}_i \sim \text{Bern}(p_i) \) (technically we mean \( \hat{Y}_i \mid x_i, w \), but we remove the conditioning terms for notational brevity), where \( p_i \) is a function of \( x_i \). Thus our observation \( y_i \), which we can view as a “sample,” has probability

\[
P(\hat{Y}_i = y_i) = \begin{cases} 
p_i & \text{if } y_i = 1 \\
1 - p_i & \text{if } y_i = 0 
\end{cases}
\]

One convenient way to write the likelihood of a single data point is

\[
P(\hat{Y}_i = y_i) = p_i^{y_i}(1 - p_i)^{(1 - y_i)}
\]

which holds no matter what \( y_i \) is.

We need a model for the dependency of \( p_i \) on \( x_i \). We have to enforce that \( p_i \) is a transformation of \( x_i \) that results in a number from 0 to 1 (ie. a valid probability). Hence \( p_i \) cannot be, say, linear in \( x_i \). One way to do achieve the 0-1 normalization is by using the sigmoid function

\[
p_i = s(w^\top x_i) = \frac{1}{1 + e^{-w^\top x_i}}
\]
5.3. LOGISTIC REGRESSION

Now we can estimate the parameters $\mathbf{w}$ via maximum likelihood. We have the problem

$$
\hat{w}_{LR} = \arg\max_{\mathbf{w}} P(\hat{Y}_1 = y_1, \ldots, \hat{Y}_n = y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{w})
$$

$$
= \arg\max_{\mathbf{w}} \prod_{i=1}^{n} P(\hat{Y}_i = y_i \mid \mathbf{x}_i, \mathbf{w})
$$

$$
= \arg\max_{\mathbf{w}} \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)}
$$

$$
= \arg\max_{\mathbf{w}} \ln \left[ \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)} \right]
$$

$$
= \arg\max_{\mathbf{w}} \sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i)
$$

$$
= \arg\min_{\mathbf{w}} - \sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i)
$$

which exactly matches the cross-entropy formulation from earlier. The logistic regression loss function can also be justified from an information-theoretic perspective. To motivate this approach, we introduce Kullback-Leibler (KL) divergence (also called relative entropy), which measures the amount that one distribution diverges from another. Given any two discrete random variables $P$ and $Q$, the KL divergence from $Q$ to $P$ is defined as

$$
D_{KL}(P \parallel Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}
$$

Note that $D_{KL}$ is not a true distance metric, because it is not symmetric, ie. $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$ in general. It also does not satisfy the triangle inequality. However, it is always positive, ie. $D_{KL}(P \parallel Q) \geq 0$, with equality iff $P = Q$.

In the context of classification, if the class label $y_i$ is interpreted as the probability of being class 1, then logistic regression provides an estimate $p_i$ of the probability that the data is in class 1. The true class label can be viewed as a sampled value from the “true” distribution $Y_i \sim \text{Bern}(y_i)$. $P(Y_i)$ is not a particularly interesting distribution because all values sampled from it will yield $y_i$: $P(Y_i = y_i) = 1$. Logistic regression yields a distribution $\hat{Y}_i \sim \text{Bern}(p_i)$, which is the posterior probability that estimates the true distribution $P(Y_i)$.

The KL divergence from $P(\hat{Y}_i)$ to $P(Y_i)$ provides a measure of how closely logistic regression can match the true label. We would like to minimize this KL divergence, and ideally we would try to choose our parameters so that $D_{KL}(P(\hat{Y}_i) \parallel P(Y_i)) = 0$. Again, this is impossible for two reasons. First, if we want $D_{KL}(P(\hat{Y}_i) \parallel P(Y_i)) = 0$, then we would need $p_i = y_i$, which is impossible because $p_i$ is the output of a logistic function that can never actually reach 0 or 1. Second, even if we tried tuning the parameters so that $D_{KL}(P(Y_i) \parallel P(\hat{Y}_i)) = 0$, that’s only optimizing one of the data points – we need to tune the parameters so that we can collectively minimize the totality of all of the KL divergences contributed by all data points.

Therefore, our goal is to tune the parameters $\mathbf{w}$ (which indirectly affect the $p_i$ values and therefore the estimated distribution $P(\hat{Y}_i)$), in order to minimize the total sum of KL divergences contributed by all data points:

$$
\hat{w}_{LR} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} D_{KL}(P(Y_i) \parallel P(\hat{Y}_i))
$$
= \arg \min_w \sum_{i=1}^n y_i \ln \frac{y_i}{p_i} + (1 - y_i) \ln \frac{1 - y_i}{1 - p_i}

= \arg \min_w \sum_{i=1}^n y_i (\ln y_i - \ln p_i) + (1 - y_i)(\ln(1 - y_i) - \ln(1 - p_i))

= \arg \min_w \sum_{i=1}^n (-y_i \ln p_i - (1 - y_i) \ln(1 - p_i)) + (y_i \ln y_i + (1 - y_i) \ln(1 - y_i))

= \arg \min_w - \sum_{i=1}^n y_i \ln p_i + (1 - y_i) \ln(1 - p_i)

= \arg \min_w \sum_{i=1}^n H(P(Y_i), P(\hat{Y}_i))

Note that the $y_i \ln y_i + (1 - y_i) \ln(1 - y_i)$ component of the KL divergence is a constant, independent of our changes to $p_i$. Therefore, we are effectively minimizing the sum of the cross entropies $H(P(Y_i), P(\hat{Y}_i))$. We conclude our discussion of KL Divergence by noting the relation between KL divergence and cross entropy:

$$D_{KL}(P(Y_i) \parallel P(\hat{Y}_i)) = H(P(Y_i), P(\hat{Y}_i)) - H(P(Y_i))$$

where:

1. $D_{KL}(P(Y_i) \parallel P(\hat{Y}_i))$ is the KL divergence from $P(\hat{Y}_i)$ to $P(Y_i)$:

$$D_{KL}(P(Y_i) \parallel P(\hat{Y}_i)) = y_i \ln \frac{y_i}{p_i} + (1 - y_i) \ln \frac{1 - y_i}{1 - p_i}$$

2. $H(P(Y_i), P(\hat{Y}_i))$ is the cross entropy between $Y_i$ and $\hat{Y}_i$:

$$H(P(Y_i), P(\hat{Y}_i)) = -y_i \ln p_i - (1 - y_i) \ln(1 - p_i)$$

3. $H(Y_i)$ is the entropy of $Y_i$:

$$H(Y_i) = -y_i \ln y_i - (1 - y_i) \ln(1 - y_i)$$

Since the parameters $W$ do not affect the entropy, we can optimize the cross entropy instead of the KL divergence.

**Multiclass Logistic Regression**

Let’s generalize logistic regression to the case where there are $K$ classes. Similarly to our discussion of the multi-class LS-SVM, it is important to note that there is no inherent ordering to the classes, and predicting a class in the continuous range from 1 to $K$ would be a poor choice. To see why, recall our fruit classification example. Suppose 1 is used to represent “peach,” 2 is used to represent “banana,” and 3 is used to represent “apple.” In our numerical representation, it would appear that peaches are less than bananas, which are less than apples. As a result, if we have an image that looks like some cross between an apple and a peach, we may simply end up classifying it as a banana.
The solution is to use a **one-hot vector encoding** to represent all of our labels. If the \(i\)’th observation has class \(k\), instead of using the representation \(y_i = k\), we can use the representation \(y_i = e_k\), the \(k\)’th canonical basis vector. For example, in our fruit example, if the \(i\)’th image is classified as “banana”, its label representation would be

\[ y_i = [0 \ 1 \ 0] \]

(Be careful to distinguish between the class label \(y_i \in \mathbb{R}\) and its one-hot encoding \(y_i \in \mathbb{R}^K\)). Now there is no relative ordering in the representations of the classes. We must modify our parameter representation accordingly to the one-hot vector encoding. Now, there are a set of \(d + 1\) parameters associated with every class, which amounts to a matrix \(W \in \mathbb{R}^{K \times (d + 1)}\). For each input \(x_i \in \mathbb{R}^{d + 1}\), each class \(k\) is given a “score”

\[ z_k = w_k^\top x_i \]

Where \(w_k \in \mathbb{R}^{d + 1}\) is the \(k\)’th row of the \(W\) matrix. In total there are \(K\) raw linear scores for an arbitrary input \(x\):

\[
\begin{bmatrix}
w_1^\top x & w_2^\top x & \ldots & w_K^\top x
\end{bmatrix}
\]

The higher the score for a class, the more likely logistic regression will pick that class. Now that we have a score system, we must transform all of these scores into a posterior probability distribution \(P(\hat{Y})\). For binary logistic regression, we used the logistic function, which takes the value \(w^\top x\) and squashes it to a value between 0 and 1. The generalization to the the logistic function for the multi-class case is the **softmax function**. The softmax function takes as input all \(K\) scores (formally known as **logits**) and an index \(j\), and outputs the probability that the corresponding softmax distribution takes value \(j\):

\[
\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}
\]

The logits induce a **softmax distribution**, which we can verify is indeed a probability distribution:

1. The entries are between 0 and 1.
2. The entries add up to 1.

Using the softmax function, we generate probabilities for each class:

\[
\hat{y} = \sigma(Wx) = [\sigma(Wx)_1 \ \sigma(Wx)_2 \ \ldots \ \sigma(Wx)_K]
\]

and our prediction is the class with the maximum probability:

\[
\hat{y} = \max_k \sigma(Wx)_k
\]

On inspection, this softmax distribution is reasonable, because the higher the score of a class, the higher its probability. In fact, we can verify that the logistic function used in the binary case is a special case of the softmax function used in the multiclass case. Assuming that the corresponding parameters for class 0 and 1 are \(w_0\) and \(w_1\), we have that:

\[
P(\hat{Y}_i = 1 \mid x_i, W) = \sigma(Wx_i)_1 = \frac{e^{w_1^\top x_i}}{e^{w_0^\top x_i} + e^{w_1^\top x_i}} = s((w_1 - w_0)^\top x_i)
\]

\[
P(\hat{Y}_i = 0 \mid x_i, W) = \sigma(Wx_i)_0 = \frac{e^{w_0^\top x_i}}{e^{w_0^\top x_i} + e^{w_1^\top x_i}} = 1 - s((w_1 - w_0)^\top x_i)
\]

In the 2-class case, because we are only interested in the difference between \(w_1\) and \(w_0\), we just use a change of variables \(w = w_1 - w_0\). We don’t need to know \(w_1\) and \(w_0\) individually, because once we know \(P(\hat{Y}_i = 1)\), we know by default that \(P(\hat{Y}_i = 0) = 1 - P(\hat{Y}_i = 1)\).
Loss Function

Let’s derive the loss function for multiclass logistic regression, first using the information-theoretic perspective. The “true” or more formally the target distribution in this case is \( P(Y_i = j) = \delta_{j,y_i} \). In other words, the entire distribution is concentrated on the label for the training example. The estimated distribution \( P(\hat{Y}_i) \) comes from multiclass logistic regression, and in this case is the softmax distribution:

\[
P(\hat{Y}_i = j) = \frac{e^{w_j^\top x_i}}{\sum_{k=1}^{K} e^{w_k^\top x_i}}
\]

Now let’s proceed to deriving the loss function. The objective, as always, is to minimize the sum of the KL divergences contributed by all of the training examples.

\[
\hat{W}_{mclr} = \arg \min_W \sum_{i=1}^{n} D_{KL}(P(Y_i) \parallel P(\hat{Y}_i))
\]

\[
= \arg \min_W \sum_{i=1}^{n} \sum_{j=1}^{K} P(Y_i = j) \ln \left( \frac{P(Y_i = j)}{P(\hat{Y}_i = j)} \right)
\]

\[
= \arg \min_W \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \ln \left( \frac{\delta_{j,y_i}}{\sigma(Wx_i)_j} \right)
\]

\[
= \arg \min_W \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \ln \delta_{j,y_i} - \delta_{j,y_i} \ln (\sigma(Wx_i)_j)
\]

\[
= \arg \min_W - \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \ln (\sigma(Wx_i)_j)
\]

\[
= \arg \min_W - \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \ln \left( \frac{e^{w_j^\top x_i}}{\sum_{k=1}^{K} e^{w_k^\top x_i}} \right)
\]

\[
= \arg \min_W \sum_{i=1}^{n} H(P(Y_i), P(\hat{Y}_i))
\]

Just like binary logistic regression, we can justify the loss function with MLE as well:

\[
\hat{W}_{mclr} = \arg \max_W P(\hat{Y}_1 = y_1, \ldots, \hat{Y}_n = y_n \mid x_1, \ldots, x_n, W)
\]

\[
= \arg \max_W \prod_{i=1}^{n} P(\hat{Y}_i = y_i \mid x_i, W)
\]

\[
= \arg \max_W \prod_{i=1}^{n} \prod_{j=1}^{K} P(\hat{Y}_i = j \mid x_i, W)^{\delta_{j,y_i}}
\]

\[
= \arg \max_W \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \ln P(\hat{Y}_i = j \mid x_i, W)
\]

\[
= \arg \max_W \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \ln \left( \frac{e^{w_j^\top x_i}}{\sum_{k=1}^{K} e^{w_k^\top x_i}} \right)
\]
5.3. LOGISTIC REGRESSION

\[ \hat{\mathbf{y}}_i = \underset{\mathbf{w}}{\arg\min} \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \ln \left( \frac{e^{\mathbf{w}^\top \mathbf{x}_i}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^\top \mathbf{x}_i}} \right) \]

We conclude that the loss function for multiclass logistic regression is

\[ L(\mathbf{W}) = -\sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \cdot \ln P(\hat{Y}_i = j \mid \mathbf{x}_i, \mathbf{W}) \]

Training

The logistic regression loss function has no known analytic closed-form solution. Therefore, in order to minimize it, we can use gradient descent, either in batch form or stochastic form. Let’s examine the case for batch gradient descent.

Binary Case

Recall the loss function

\[ L(\mathbf{w}) = -\sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln (1 - p_i) \]

where

\[ p_i = s(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}_i}} \]

\[ \nabla_{\mathbf{w}} L(\mathbf{w}) = \nabla_{\mathbf{w}} \left( -\sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln (1 - p_i) \right) \]

\[ = -\sum_{i=1}^{n} y_i \nabla_{\mathbf{w}} \ln p_i + (1 - y_i) \nabla_{\mathbf{w}} \ln (1 - p_i) \]

\[ = -\sum_{i=1}^{n} \frac{y_i}{p_i} \nabla_{\mathbf{w}} p_i - \frac{1 - y_i}{1 - p_i} \nabla_{\mathbf{w}} p_i \]

\[ = -\sum_{i=1}^{n} \left( \frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right) \nabla_{\mathbf{w}} p_i \]

Note that \( \nabla_z s(z) = s(z)(1 - s(z)) \), and from the chain rule we have that

\[ \nabla_{\mathbf{w}} p_i = \nabla_{\mathbf{w}} s(\mathbf{w}^\top \mathbf{x}_i) = s(\mathbf{w}^\top \mathbf{x}_i)(1 - s(\mathbf{w}^\top \mathbf{x}_i)) \mathbf{x}_i = p_i(1 - p_i) \mathbf{x}_i \]

Plugging in this gradient value, we have

\[ \nabla_{\mathbf{w}} L(\mathbf{w}) = -\sum_{i=1}^{n} \left( \frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right) \nabla_{\mathbf{w}} p_i \]
\[
\begin{align*}
&= -\sum_{i=1}^{n} \left( \frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right) p_i (1 - p_i) x_i \\
&= -\sum_{i=1}^{n} \left( y_i (1 - p_i) - (1 - y_i) (p_i) \right) x_i \\
&= -\sum_{i=1}^{n} (y_i - p_i) x_i
\end{align*}
\]

The gradient descent update is thus

\[
\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \epsilon \sum_{i=1}^{n} (y_i - p_i) x_i
\]

It does not matter what initial values we pick for \( \mathbf{w} \), because the loss function \( L(\mathbf{w}) \) is convex and does not have any local minima. Let’s prove this, by first finding the Hessian of the loss function. The \( k,l \)th entry of the Hessian is the partial derivative of the gradient with respect to \( w_k \) and \( w_\ell \):

\[
H_{k\ell} = \frac{\partial^2 L(\mathbf{w})}{\partial w_k \partial w_\ell} = \frac{\partial}{\partial w_k} - \sum_{i=1}^{n} (y_i - p_i) x_{i\ell}
\]

\[
= \sum_{i=1}^{n} \frac{\partial}{\partial w_k} p_i x_{i\ell}
\]

\[
= \sum_{i=1}^{n} p_i (1 - p_i) x_{i\ell} x_{i\ell}
\]

We conclude that

\[
H(\mathbf{w}) = \sum_{i=1}^{n} p_i (1 - p_i) \mathbf{x}_i \mathbf{x}_i^\top
\]

To prove that \( L(\mathbf{w}) \) is convex in \( \mathbf{w} \), we need to show that \( \forall \mathbf{v}, \mathbf{v}^\top H(\mathbf{w}) \mathbf{v} \geq 0 \):

\[
\mathbf{v}^\top H(\mathbf{w}) \mathbf{v} = \mathbf{v}^\top \sum_{i=1}^{n} p_i (1 - p_i) \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v} = \sum_{i=1}^{n} (\mathbf{v}^\top \mathbf{x}_i)^2 p_i (1 - p_i) \geq 0
\]

Multiclass Case

Instead of finding the gradient with respect to all of the parameters of the matrix \( \mathbf{W} \), let’s find them with respect to one row of \( \mathbf{W} \) at a time:

\[
\nabla_{\mathbf{w}_i} L(\mathbf{W}) = \nabla_{\mathbf{w}_i} - \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \cdot \ln \left( \frac{e^{\mathbf{w}_j^\top \mathbf{x}_i}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^\top \mathbf{x}_i}} \right)
\]

\[
= - \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \cdot \nabla_{\mathbf{w}_i} \left( \ln \frac{e^{\mathbf{w}_j^\top \mathbf{x}_i}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^\top \mathbf{x}_i}} \right)
\]
\[\begin{align*}
&= -\sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \cdot \left( \nabla_{w_j} w_j^\top x_i - \nabla_{w_i} \ln \sum_{k=1}^{K} e^{w_k^\top x_i} \right) \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{j,y_i} \cdot \left( \delta_{\ell,y_i} - \frac{e^{w_{\ell}^\top x_i}}{\sum_{k=1}^{K} e^{w_k^\top x_i}} \right) x_i \\
&= -\sum_{i=1}^{n} \left( \delta_{\ell,y_i} - P(\hat{Y}_i = \ell) \right) x_i
\end{align*}\]

The gradient descent update for \(w_\ell\) is thus

\[w_{\ell}^{(t+1)} = w_{\ell}^{(t+1)} + \epsilon \sum_{i=1}^{n} \left( \delta_{\ell,y_i} - P(\hat{Y}_i = \ell) \right) x_i\]

Just as with binary logistic regression, it does not matter what initial values we pick for \(W\), because the loss function \(L(W)\) is convex and does not have any local minima.

### 5.4 Gaussian Discriminant Analysis

Recall the idea of **generative models**: we classify an arbitrary datapoint \(x\) with the class label that maximizes the joint probability \(p(x,Y)\) over the label \(Y\):

\[\hat{y} = \arg \max_k p(x,Y = k)\]

Generative models typically form the joint distribution by explicitly forming the following:

- A prior probability distribution over all classes:
  \[P(k) = P(\text{class} = k)\]

- A conditional probability distribution for each class \(k \in \{1, 2, \ldots, K\}\):
  \[p_k(X) = p(X|\text{class} \ k)\]

In total there are \(K + 1\) probability distributions: 1 for the prior, and \(K\) for all of the individual classes. Note that the prior probability distribution is a categorical distribution over the \(K\) discrete classes, whereas each class conditional probability distribution is a continuous distribution over \(\mathbb{R}^d\) (often represented as a Gaussian). Using the prior and the conditional distributions in conjunction, we have (from Bayes’ rule) that maximizing the joint probability over the class labels is equivalent to maximizing the posterior probability of the class label:

\[\hat{y} = \arg \max_k p(x,Y = k) = \arg \max_k P(k) \ p_k(x) = \arg \max_k P(Y = k|x)\]

Consider the example of digit classification. Suppose we are given dataset of images of handwritten digits each with known values in the range \(\{0,1,2,\ldots,9\}\). The task is, given an image of a handwritten digit, to classify it to the correct digit. A generative classifier for this this task would
effectively form a the prior distribution and conditional probability distributions over the 10 possible digits and choose the digit that maximizes posterior probability:

$$\hat{y} = \arg \max_{k \in \{0,1,2,...,9\}} p(\text{digit} = k | \text{image}) = \arg \max_{k \in \{0,1,2,...,9\}} P(\text{digit} = k) p(\text{image} | \text{digit} = k)$$

Maximizing the posterior will induce regions in the feature space in which one class has the highest posterior probability, and decision boundaries in between classes where the posterior probability of two classes are equal.

**Gaussian Discriminant Analysis (GDA)** is a specific generative method in which the class conditional probability distributions are Gaussian: $(X | Y = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$. (Caution: the term “discriminant” in GDA is misleading; GDA is a generative method, it is not a discriminative method!)

Assume that we are given a training set $D = \{(x_i, y_i)\}_{i=1}^n$ of $n$ points. Estimating the prior distribution is the same for any other generative model. The probability of a class $k$ is

$$P(k) = \frac{n_k}{n}$$

where $n_k$ is the number of training points that belong to class $k$. We can estimate the parameters of the conditional distributions with MLE. Once we form the estimated prior conditional distributions, we use Bayes’ Rule to directly solve the optimization problem

$$\hat{y} = \arg \max_k p(k | x)$$

$$= \arg \max_k P(k) p_k(x)$$

$$= \arg \max_k \ln(P(k)) + \ln(\sqrt{2\pi})^d p_k(x)$$

$$= \arg \max_k \ln(P(k)) - \frac{1}{2}(x - \hat{\mu}_k)^\top \hat{\Sigma}_k^{-1}(x - \hat{\mu}_k) - \frac{1}{2} \ln(|\hat{\Sigma}_k|) = Q_k(x)$$

For future reference, let’s use $Q_k(x) = \ln(\sqrt{2\pi})^d P(k) p_k(x)$ to simplify our notation.

We classify an arbitrary test point

$$\hat{y} = \arg \max_{k \in \{1,2,...,K\}} Q_k(x)$$

GDA comes in two flavors: Quadratic Discriminant Analysis (QDA) in which the decision boundary is quadratic, and Linear Discriminant Analysis (LDA) in which the decision boundary is linear. We will now present both and compare them in detail.

**QDA Classification**

In **Quadratic Discriminant Analysis (QDA)**, the class conditional probability distributions are independent Gaussians — namely, the covariance $\Sigma_k$ of class $k$ has no dependence/relation to that of the other classes.

Due to this independence property, we can estimate the true mean and covariance $\mu_k, \Sigma_k$ for each class conditional probability distribution $p_k(X)$ independently, with the $n_k$ samples in our training
data that are classified as class $k$. The MLE estimate for the parameters of $p_k(X)$ is:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i : y_i = k} x_i$$

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i : y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top$$

**LDA Classification**

While QDA is a reasonable approach to classification, we might be interested in simplifying our model to reduce the number of parameters we have to learn. One way to do this is through **Linear Discriminant Analysis (LDA)** classification. Just as in QDA, LDA assumes that the class conditional probability distributions are normally distributed with different means $\mu_k$, but LDA is different from QDA in that it requires all of the distributions to share the same covariance matrix $\Sigma$. This is a simplification which, in the context of the Bias-Variance tradeoff, increases the bias of our method but may help decrease the variance.

The training and classification procedures for LDA are almost identical that of QDA. To compute the within-class means, we still want to take the empirical mean. However, the empirical covariance for all classes is now computed as

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^\top$$

One way to understand this formula is as a weighted average of the within-class covariances. Here, assume we have sorted our training data by class and we can index through the $x_i$’s by specifying a class $k$ and the index within that class $j$:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^\top$$

$$= \frac{1}{n} \sum_{k=1}^{K} \sum_{i : y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top$$

$$= \frac{1}{n} \sum_{k=1}^{K} n_k \Sigma_k$$

$$= \sum_{k=1}^{K} \frac{n_k}{n} \Sigma_k$$

**LDA and QDA Decision Boundary**

Let’s now derive the form of the decision boundary for QDA and LDA. As we will see, the term *quadratic* in QDA and *linear* in LDA actually signify the shape of the decision boundary. We will prove this claim using binary (2-class) examples for simplicity (class $A$ and class $B$). An arbitrary
point \( x \) is classified according to the following cases:

\[
\hat{y} = \begin{cases} 
A & Q_A(x) > Q_B(x) \\
B & Q_A(x) < Q_B(x) \\
\text{Either } A \text{ or } B & Q_A(x) = Q_B(x)
\end{cases}
\]

The decision boundary is the set of all points in \( x \)-space that are classified according to the third case.

**Identical Conditional Distributions with Identical Priors**

The simplest case is when the two classes are equally likely in prior, and their conditional probability distributions are isotropic with identical covariances. Recall that isotropic Gaussian distributions have covariances of the form of \( \Sigma = \sigma^2 I \), which means that their isocontours are circles. In this case, \( p_A(X) \) and \( p_B(X) \) have identical covariances of the form \( \Sigma_A = \Sigma_B = \sigma^2 I \).

![Contour plot of two isotropic, identically distributed Gaussians in \( \mathbb{R}^2 \). The circles are the level sets of the Gaussians.](image)

Geometrically, we can see that the task of classifying a 2-D point into one of the two classes amounts simply to figuring out which of the means it’s closer to. Using our notation of \( Q_k(x) \) from before, this can be expressed mathematically as:

\[
\begin{align*}
\ln(P(A)) - \frac{1}{2}(x - \hat{\mu}_A)^\top \Sigma_A^{-1}(x - \hat{\mu}_A) - \frac{1}{2} \ln(|\Sigma_A|) &= \ln(P(B)) - \frac{1}{2}(x - \hat{\mu}_B)^\top \Sigma_B^{-1}(x - \hat{\mu}_B) - \frac{1}{2} \ln(|\Sigma_B|) \\
\ln\left(\frac{1}{2}\right) - \frac{1}{2}(x - \hat{\mu}_A)^\top \sigma^{-2} I(x - \hat{\mu}_A) - \frac{1}{2} \ln(|\sigma^2 I|) &= \ln\left(\frac{1}{2}\right) - \frac{1}{2}(x - \hat{\mu}_B)^\top \sigma^{-2} I(x - \hat{\mu}_B) - \frac{1}{2} \ln(|\sigma^2 I|) \\
(x - \hat{\mu}_A)^\top(x - \hat{\mu}_A) &= (x - \hat{\mu}_B)^\top(x - \hat{\mu}_B)
\end{align*}
\]

The decision boundary is the set of points \( x \) for which \( ||x - \hat{\mu}_A||_2 = ||x - \hat{\mu}_B||_2 \), which is simply the set of points that are equidistant from \( \hat{\mu}_A \) and \( \hat{\mu}_B \). This decision boundary is linear because
the set of points that are equidistant from \( \hat{\mu}_A \) and \( \hat{\mu}_B \) are simply the perpendicular bisector of the segment connecting \( \hat{\mu}_A \) and \( \hat{\mu}_B \).

The next case is when the two classes are equally likely in prior, and their conditional probability distributions are anisotropic with identical covariances.

![Figure 5.7: Two anisotropic, identically distributed Gaussians in \( \mathbb{R}^2 \). The ellipses are the level sets of the Gaussians.](image)

The anisotropic case can be reduced to the isotropic case simply by performing a linear change of coordinates that transforms the ellipses back into circles, which induces a linear decision boundary both in the transformed and original space. Therefore, the decision boundary is still the set of points that are equidistant from \( \hat{\mu}_A \) and \( \hat{\mu}_B \).

**Identical Conditional Distributions with Different Priors**

Now, let’s find the decision boundary when the two classes still have identical covariances but are not necessarily equally likely in prior:

\[
Q_A(x) = Q_B(x) \\
\ln(P(A)) - \frac{1}{2}(x - \hat{\mu}_A)^\top \hat{\Sigma}^{-1}_A(x - \hat{\mu}_A) - \frac{1}{2} \ln(|\hat{\Sigma}_A|) = \ln(P(B)) - \frac{1}{2}(x - \hat{\mu}_B)^\top \hat{\Sigma}^{-1}_B(x - \hat{\mu}_B) - \frac{1}{2} \ln(|\hat{\Sigma}_B|) \\
\ln(P(A)) - \frac{1}{2}(x - \hat{\mu}_A)^\top \hat{\Sigma}^{-1}_A(x - \hat{\mu}_A) = \ln(P(B)) - \frac{1}{2}(x - \hat{\mu}_B)^\top \hat{\Sigma}^{-1}_B(x - \hat{\mu}_B) \\
2 \ln(P(A)) - x^\top \hat{\Sigma}^{-1}_A x + 2 x^\top \hat{\Sigma}^{-1}_A \hat{\mu}_A - \hat{\mu}_A^\top \hat{\Sigma}^{-1}_A \hat{\mu}_A = 2 \ln(P(B)) - x^\top \hat{\Sigma}^{-1}_B x + 2 x^\top \hat{\Sigma}^{-1}_B \hat{\mu}_B - \hat{\mu}_B^\top \hat{\Sigma}^{-1}_B \hat{\mu}_B \\
2 \ln(P(A)) + 2 x^\top \hat{\Sigma}^{-1}_A \hat{\mu}_A - \hat{\mu}_A^\top \hat{\Sigma}^{-1}_A \hat{\mu}_A = 2 \ln(P(B)) + 2 x^\top \hat{\Sigma}^{-1}_B \hat{\mu}_B - \hat{\mu}_B^\top \hat{\Sigma}^{-1}_B \hat{\mu}_B
\]

Simplifying, we have that

\[
x^\top \hat{\Sigma}^{-1}_A (\hat{\mu}_A - \hat{\mu}_B) + \left( \ln \left( \frac{P(A)}{P(B)} \right) - \frac{\hat{\mu}_A^\top \hat{\Sigma}^{-1}_A \hat{\mu}_A - \hat{\mu}_B^\top \hat{\Sigma}^{-1}_B \hat{\mu}_B}{2} \right) = 0
\]
The decision boundary is the level set of a linear function $f(x) = w^\top x - b$. In fact, the decision boundary is the level set of a linear function (which itself is linear) as long as the two class conditional probability distributions share the same covariance matrices. This is the reason for why LDA has a linear decision boundary.

**Nonidentical Conditional Distributions with Different Priors**

This is the most general case. We have that:

$$Q_A(x) = Q_B(x)$$

$$\ln(p(A)) - \frac{1}{2}(x - \mu_A)^\top \Sigma_A^{-1}(x - \mu_A) - \frac{1}{2} \ln(|\Sigma_A|) = \ln(p(B)) - \frac{1}{2}(x - \mu_B)^\top \Sigma_B^{-1}(x - \mu_B) - \frac{1}{2} \ln(|\Sigma_B|)$$

Here, unlike in LDA when $\Sigma_A = \Sigma_B$, we cannot cancel out the quadratic terms in $x$ from both sides of the equation, and thus our decision boundary is now represented by the level set of an arbitrary quadratic function.

It should now make sense why QDA is short for *quadratic* discriminant analysis and LDA is short for *linear* discriminant analysis!

**LDA and Logistic Regression**

As it turns out, LDA and logistic regression share the same type of posterior distribution. We already showed that the posterior distribution in logistic regression is

$$P(Y = A|x) = \frac{1}{1 + e^{w^\top x - b}} = s(w^\top x - b)$$

for some appropriate vector $w$ and bias $b$. Now let’s derive the posterior distribution for LDA. From Bayes’ rule we have that

$$P(Y = A|x) = \frac{p(x|Y = A)P(Y = A)}{p(x|Y = B)P(Y = B) + p(x|Y = B)P(Y = B)}$$

$$= \frac{e^{Q_A(x)}}{e^{Q_A(x)} + e^{Q_B(x)}}$$

$$= \frac{1}{1 + e^{Q_A(x) - Q_B(x)}}$$

We already showed the the decision boundary in LDA is linear — it is the set of points $x$ such that

$$Q_A(x) - Q_B(x) = w^\top x - b = 0$$

for some appropriate vector $w$ and bias $b$. We therefore have that

$$P(Y = A|x) = \frac{1}{1 + e^{w^\top x - b}} = s(w^\top x - b)$$

As we can see, even though logistic regression is a discriminative method and LDA is a generative method, both methods complement each other, arriving at the same form for the posterior distribution.
5.5 SUPPORT VECTOR MACHINES

5.5.1 Generalizing to Multiple Classes

The analysis on the decision boundary in QDA and LDA can be extended to the general case when there are more than two classes. In the multiclass setting, the decision boundary is a collection of linear boundaries in LDA and quadratic boundaries in QDA. The following Voronoi diagrams illustrate the point:

![Figure 5.8: LDA (left) vs QDA (right): a collection of linear vs quadratic level set boundaries. Source: Professor Shewchuk’s notes](image)

5.5 Support Vector Machines

So far we’ve explored generative classifiers (LDA) and discriminative classifiers (logistic regression), but in both of these methods, we tasked ourselves with modeling some kind of probability distribution. One observation about classification is that in the end, if we only care about assigning each data point a class, all we really need to know do is find a “good” decision boundary, and we can skip thinking about the distributions. Support Vector Machines (SVMs) are an attempt to model decision boundaries directly in this spirit.

Here’s the setup for the problem. We are given a training dataset $D = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. Our goal is to find a $d - 1$ dimensional hyperplane decision boundary $H$ which separates the +1’s from the −1’s.

**Motivation for SVMs**

In order to motivate SVMs, we first have to understand the simpler perceptron algorithm and its shortcomings. Given that the training data is linearly separable, the perceptron algorithm finds a $d - 1$ dimensional hyperplane that perfectly separates the +1’s from the −1’s. Mathematically, the goal is to learn a set of parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, that satisfy the linear separability constraints:

$$\forall i, \begin{cases} w^T x_i - b \geq 0 \quad \text{if } y_i = 1 \\ w^T x_i - b \leq 0 \quad \text{if } y_i = -1 \end{cases}$$

Equivalently,

$$\forall i, \quad y_i (w^T x_i - b) \geq 0$$

The resulting decision boundary is a hyperplane $H = \{x : w^T x - b = 0\}$. All points on the positive side of the hyperplane are classified as +1, and all points on the negative side are classified as −1.
Perceptrons have two major shortcomings that as we shall see, SVMs can overcome. First of all, if the data is not linearly separable, the perceptron fails to find a stable solution. As we shall see, soft-margin SVMs fix this issue by allowing best-fit decision boundaries even when the data is not linearly separable. Second, if the data is linearly separable, the perceptron could find infinitely many hyperplanes that the perceptron could pick — if \((w, b)\) is a pair that separates the data points, then the perceptron could also end up choosing a slightly different \((w, b + \epsilon)\) pair that still separates the data points. Some hyperplanes are better than others, but the perceptron cannot distinguish between them. This leads to generalization issues.

![Several possible decision boundaries under the perceptron. The X's and C's represent the +1's and -1's respectively.](image)

In the figure above, we consider three potential linear separators that satisfy the constraints. To the eyes of the perceptron algorithm, all three are perfectly valid linear separators. Ideally, we should not treat all linear separators equally — some are better than others. One could imagine that if we observed new test points that are nearby the region of C’s (or X’s) in the training data, they should also be of class C (or X). The two separators close to the training points would incorrectly classify some of these new test points, while the third separator which maintains a large distance to the points would classify them correctly. The perceptron algorithm does not take this reasoning into account, and may find a classifier that does not generalize well to unseen data.

**Hard-Margin SVMs**

**Hard-Margin SVMs** address the generalization problem of perceptrons by maximizing the margin, formally defined as the minimum distance from the decision boundary to the training points.
Figure 5.10: The optimal decision boundary (as shown) maximizes the margin.

Intuitively, maximizing the margin allows us to generalize better to unseen data, because the decision boundary with the maximum margin is as far away from the training data as possible and the boundary cannot be violated unless the unseen data contains outliers.

Simply put, the goal of hard-margin SVMs is to find a hyperplane $H$ that maximizes the margin $m$. Let’s formalize an optimization problem for hard-margin SVMs. The variables we are trying to optimize over are the margin $m$ and the parameters of the hyperplane, $w$ and $b$. The objective is to maximize the margin $m$, subject to the following constraints:

- All points classified as $+1$ are to the positive side of the hyperplane and their distance to $H$ is greater than the margin.
- All points classified as $-1$ are to the negative side of the hyperplane and their distance to $H$ is greater than the margin.
- The margin is non-negative.

Let’s express the first two constraints mathematically. First, note that the vector $w$ is perpendicular to the hyperplane $H = \{x : w^\top x - b = 0\}$.

Proof: consider any two points on $H$, $x_0$ and $x_1$. We will show that $(x_1 - x_0) \perp w$. Note that:

\[
(x_1 - x_0)^\top(w) = (x_1 - x_0)^\top((x_1 + w) - x_1) = x_1^\top w - x_0^\top w = b - b = 0
\]
Since \( \mathbf{w} \) is perpendicular to \( H \), the (shortest) distance from any arbitrary point \( \mathbf{z} \) to the hyperplane \( H \) is determined by a scaled multiple of \( \mathbf{w} \). If we take any point on the hyperplane \( \mathbf{x}_0 \), the distance from \( \mathbf{z} \) to \( H \) is the length of the projection from \( \mathbf{z} - \mathbf{x}_0 \) to the vector \( \mathbf{w} \), which is

\[
D = \frac{|\mathbf{w}^\top (\mathbf{z} - \mathbf{x}_0)|}{\|\mathbf{w}\|_2} = \frac{|\mathbf{w}^\top \mathbf{z} - \mathbf{w}^\top \mathbf{x}_0|}{\|\mathbf{w}\|_2} = \frac{|\mathbf{w}^\top \mathbf{z} - b|}{\|\mathbf{w}\|_2}
\]

![Figure 5.12: Shortest distance from \( \mathbf{z} \) to \( H \) is determined by projection of \( \mathbf{z} - \mathbf{x}_0 \) onto \( \mathbf{w} \)](image)

Therefore, the distance from any of the training points \( \mathbf{x}_i \) to \( H \) is

\[
\frac{|\mathbf{w}^\top \mathbf{x}_i - b|}{\|\mathbf{w}\|_2}
\]

In order to ensure that positive points are on the positive side of the hyperplane outside a margin of size \( m \), and that negative points are on the negative side of the hyperplane outside a margin of size \( m \), we can express the constraint

\[
y_i \frac{(\mathbf{w}^\top \mathbf{x}_i - b)}{\|\mathbf{w}\|_2} \geq m
\]

Putting everything together, we have the following optimization problem:

\[
\max_{m, \mathbf{w}, b} m \\
\text{s.t. } y_i \frac{(\mathbf{w}^\top \mathbf{x}_i - b)}{\|\mathbf{w}\|_2} \geq m \quad \forall i \\
m \geq 0
\]

Maximizing the margin \( m \) implies that there exists at least one point on the positive side of the hyperplane and at least one point on the negative side whose distance to the hyperplane is exactly equal to \( m \). These points are the support vectors, hence the name “support vector machines.” They are called support vectors because they literally hold/support the margin planes in place.

Through a series of optimization steps, we can simplify the problem by removing the margin variable and just optimizing the parameters of the hyperplane. Note that the current optimization formulation does not induce a unique choice of \( \mathbf{w} \) and \( b \): if \((m^*, \mathbf{w}^*, b^*)\) is a solution, then \((m^*, \alpha \mathbf{w}^*, \alpha b^*)\) is also a solution, for any \( \alpha > 0 \). In order to ensure that \( \mathbf{w} \) and \( b \) are unique (without changing the nature of the optimization problem), we can add an additional constraint for the norm of \( \mathbf{w} \):
∥w∥_2 = α, for some α > 0. In particular, we can add the constraint ∥w∥_2 = \frac{1}{m} or equivalently, 
m = \frac{1}{∥w∥_2}:

\begin{align*}
\max_{m,w,b} \quad & m \\
\text{s.t.} \quad & y_i \left( \frac{w^\top x_i - b}{∥w∥_2} \right) \geq m \quad ∀i \\
\quad & m \geq 0 \\
\quad & m = \frac{1}{∥w∥_2}
\end{align*}  \tag{5.2}

Now, we can substitute \( m = \frac{1}{∥w∥_2} \) and eliminate \( m \) from the optimization:

\begin{align*}
\max_{w,b} \quad & \frac{1}{∥w∥_2} \\
\text{s.t.} \quad & y_i (w^\top x_i - b) \geq 1 \quad ∀i
\end{align*}  \tag{5.3}

At last, we have formulated the hard-margin SVM optimization problem! The standard formulation of hard-margin SVMs is

\begin{align*}
\min_{w,b} \quad & \frac{1}{2} ∥w∥^2 \\
\text{s.t.} \quad & y_i (w^\top x_i - b) \geq 1 \quad ∀i
\end{align*}  \tag{5.4}

**Soft-Margin SVMs**

The hard-margin SVM optimization problem has a unique solution only if the data are linearly separable, but it has no solution otherwise. This is because the constraints are impossible to satisfy if we can’t draw a hyperplane that separates the +1’s from the −1’s. In addition, hard-margin SVMs are very sensitive to outliers — for example, if our data is class-conditionally distributed Gaussian such that the two Gaussians are far apart, if we witness an outlier from class +1 that crosses into the typical region for class −1, then hard-margin SVM will be forced to compromise a more generalizable fit in order to accommodate for this point. Our next goal is to come up with a classifier that is not sensitive to outliers and can work even in the presence of data that is not linearly separable. To this end, we’ll talk about **Soft-Margin SVMs**.

A soft-margin SVM modifies the constraints from the hard-margin SVM by allowing some points to violate the margin. It introduces **slack variables** \( ξ_i \), one for each training point, into the constraints:

\[
y_i (w^\top x_i - b) \geq 1 - ξ_i \\
ξ_i \geq 0
\]

The constraints are now a less-strict, *softer* version of the hard-margin SVM constraints, because each point \( x_i \) need only be a “distance” of \( 1 - ξ_i \) of the separating hyperplane instead of a hard “distance” of 1.

(By the way, the Greek letter ξ is spelled “xi” and pronounced “zai.” \( ξ_i \) is pronounced “zai-eye.”)

These constraints would be fruitless if we didn’t bound the values of the \( ξ_i \)’s — by setting them to large values, we are saying that any point may violate the margin by an arbitrarily large distance, which makes our choice of \( w \) meaningless. Therefore we modify the objective function to penalize
the slacks:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

Where $C$ is a hyperparameter tuned through cross-validation. Putting the objective and constraints together, the soft-margin SVM optimization problem is

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(w^\top x_i - b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i \quad \text{(5.5)}$$

The table below compares the effects of having a large $C$ versus a small $C$. As $C$ goes to infinity, the penalty for having non-zero $\xi_i$ goes to infinity, and thus we force the $\xi_i$’s to be zero, which is exactly the setting of the hard-margin SVM.

<table>
<thead>
<tr>
<th></th>
<th>small $C$</th>
<th>large $C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desire</td>
<td>maximize margin</td>
<td>keep $\xi_i$’s small or zero</td>
</tr>
<tr>
<td>Danger</td>
<td>underfitting</td>
<td>overfitting</td>
</tr>
<tr>
<td>Outliers</td>
<td>less sensitive</td>
<td>more sensitive</td>
</tr>
</tbody>
</table>

**SVMs as Tikhonov Regularization Learning**

The constrained version of soft-margin SVM optimization problem

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(w^\top x_i - b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i \quad \text{(5.6)}$$

can equivalently be expressed in an unconstrained fashion:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \max(1 - y_i(w^\top x_i - b), 0) + \lambda \|w\|^2$$

Let’s see why. Manipulating the first constraint of constraints, we have that

$$\xi_i \geq 1 - y_i(w^\top x_i - b)$$

Combining with the constraint $\xi_i \geq 0$, we have that

$$\xi_i \geq \max(1 - y_i(w^\top x_i - b), 0)$$

At the optimal value of the optimization problem, these inequalities must be tight. Otherwise, we could lower each $\xi_i$ to equal $\max(1 - y_i(w^\top x_i - b), 0)$ and decrease the value of the objective function. Thus we can rewrite the soft-margin SVM optimization problem as

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad \xi_i = \max(1 - y_i(w^\top x_i - b), 0) \quad \forall i \quad \text{(5.7)}$$
5.6 Generalization and Stability

Simplifying further, we can remove the constraints:

$$\min_{w, b} C \sum_{i=1}^{n} \max(1 - y_i(w^\top x_i - b), 0) + \frac{1}{2} \|w\|^2$$  \hspace{1cm} (5.8)

If we divide by $Cn$ (which does not change the optimal solution of the optimization problem), we can see that this formulation is equivalent to the regularized regression problem, with $\lambda = \frac{1}{2Cn}$. Thus we have two interpretations of soft-margin SVM: either as finding a max-margin hyperplane that is allowed to make some mistakes via slack variables $\xi_i$, or as regularized empirical risk minimization.

5.6 Generalization and Stability

In general, most classification problems are an optimization over the objective

$$\min_{w, b} \frac{1}{n} \sum_{i=1}^{n} L(y_i, w^\top x_i - b) + \lambda \|w\|^2$$

over some specified loss function and a possible regularization term (sometimes we may set $\lambda = 0$). The most simple loss function that we can optimize is the 0-1 step loss:

$$L_{\text{STEP}}(y, w^\top x - b) = \begin{cases} 1 & y(w^\top x - b) < 0 \\ 0 & y(w^\top x - b) \geq 0 \end{cases}$$

The 0-1 loss is 0 if $x$ is correctly classified and 1 otherwise. Minimizing $\frac{1}{n} \sum_{i=1}^{n} L(y_i, w^\top x_i - b)$ directly minimizes classification error on the training set. However, the 0-1 loss is difficult to optimize: it is neither convex nor differentiable (see Figure 7.1). Furthermore, if we exclude the regularization term, we do not penalize the classifier for being close to the training points, which leads to generalization issues.

Another loss function that we have seen is the logistic loss, which is used in logistic regression:

$$L_{LR}(y, w^\top x - b) = y \ln\left(\frac{1}{s(w^\top x - b)}\right) + (1 - y) \ln\left(\frac{1}{1 - s(w^\top x - b)}\right)$$

Figure 5.13: Step (0-1) loss, hinge loss, and logistic loss. Logistic loss is convex and differentiable, hinge loss is only convex, and step loss is neither.
The logistic loss is convex and differentiable, and is optimized using gradient descent methods. The logistic loss is the basis for logistic regression, and it works well without regularization:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} L_{LR}(y_i, w^\top x_i - b)$$

The hinge loss modifies the 0-1 loss to be convex. The points with $y(w^\top x - b) \geq 0$ should remain at 0 loss, but we may consider allowing a linear penalty “ramp” for misclassified points. This leads us to the hinge loss, as illustrated in Figure 7.1:

$$L_{Hinge}(y, w^\top x - b) = \max(1 - y(w^\top x - b), 0)$$

The ramp ensures that misclassified points that are close to the boundary are penalized less than misclassified points that are far from the boundary. The perceptron algorithm optimizes over the sum of hinge losses contributed from all of the training points:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} L_{Hinge}(y_i, w^\top x_i - b)$$

The SVM formulation is an optimization over the same problem, with the addition of a regularization term:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} L_{Hinge}(y_i, w^\top x_i - b) + \lambda \|w\|^2$$

The regularization term allows for better generalization, in this case by penalizing choices of $w$ for which the margin is small.

### 5.7 Duality

As we have seen in our discussion of kernels, ridge regression can be viewed in two ways: (1) an optimization problem over the weights $w \in \mathbb{R}^d$ which scales according to the dimensionality of the augmented feature space, and (2) an optimization problem over the weights $\alpha \in \mathbb{R}^n$ which scales according to the number of training points. These two viewpoints give rise to two equivalent solutions:

$$w^* = (X^\top X + \lambda I)^{-1} X^\top y \quad \text{and} \quad w^* = X^\top (XX^\top + \lambda I)^{-1} y$$

The second (kernelized) expression is much more efficient to calculate when the number of training points $n$ is significantly smaller than the number of augmented features $d$. Recall that the derivation for the kernelized expression relied on invoking the fundamental theorem of linear algebra and solving for a set of dual variables. While this approach is certainly valid, it may not be applicable for kernelizing all problems. Rather, a more principled approach is to apply Lagrangian duality and solve the dual problem. In this section we will introduce duality for arbitrary optimization problems, and then use duality to derive the kernelized versions for ridge regression and SVMs.

### Primal and Dual Problem

All optimization problems can be expressed in the standard form

$$\begin{align*}
\min_{x} & \quad f_0(x) \\
\text{s.t.} & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad g_j(x) = 0, \quad j = 1, \ldots, n
\end{align*}$$

(5.9)
For the purposes of our discussion, assume that $x \in \mathbb{R}^d$. The components of an optimization problem are:

- The **objective function** $f_0(x)$
- The **inequality constraints**: expressions involving $f_i(x)$
- The **equality constraints**: expressions involving $h_j(x)$

Working with the constraints can be cumbersome and challenging to manipulate, and it would be ideal if we could somehow turn this constrained optimization problem into an unconstrained one. One idea is to re-express the optimization problem into

$$\min_x L(x)$$

where

$$L(x) = \begin{cases} f_0(x) & \text{if } f_i(x) \leq 0, \forall i \in \{1 \ldots m\} \text{ and } h_j(x) = 0, \forall j \in \{1 \ldots n\} \\ \infty & \text{otherwise} \end{cases} \quad (5.10)$$

Note that the unconstrained optimization problem above is equivalent to the original constrained problem. Even though the unconstrained problem considers values that violate the constraints (and therefore are not in the feasible set for the constrained optimization problem), it will effectively ignore them because they are treated as $\infty$ in a minimization problem.

Even though we are now dealing with an unconstrained problem, it still is difficult to solve the optimization problem, because we still have to deal with all of the casework in the objective function $L(x)$. In order to solve this issue, we have to introduce dual variables, specifically one set of dual variables for the equality constraints, and one set for the inequality constraints. If we only take into account the dual variables for the equality constraints, the optimization problem now becomes

$$\min_{x} \max_{\nu} L(x, \nu)$$

where

$$L(x, \nu) = \begin{cases} f_0(x) + \sum_{j=1}^{n} \nu_j h_j(x) & \text{if } f_i(x) \leq 0, \forall i \in \{1 \ldots m\} \\ \infty & \text{otherwise} \end{cases} \quad (5.11)$$

We are still working with an unconstrained optimization problem, except that now, we are optimizing over two sets of variables: the **primal variables** $x \in \mathbb{R}^d$ and the **dual variables** $\nu \in \mathbb{R}^n$.

Also note that the optimization problem has now become a nested one, with an inner optimization problem the maximizes over the dual variables, and an outer optimization problem that minimizes over the primal variables. Let’s examine why this optimization problem is equivalent to the original constrained optimization problem:

- Any $x$ that violates the inequality constraints is still treated as $\infty$ by the outer minimization problem over $x$ and therefore ignored
- For any $x$ that violates the equality constraints (meaning that $\exists j$ s.t. $h_j(x) \neq 0$), the inner maximization problem over $\nu$ can choose $\nu_j$ as $\infty$ if $h_j(x) > 0$ (or $\nu_j$ as $-\infty$ if $h_j(x) < 0$) to cause the inner maximization to go to $\infty$, therefore being ignored by the outer minimization over $x$
• For any \( x \) that does not violate any of the equality or inequality constraints, the inner maximization problem over \( \nu \) is simply equal to \( f_0(x) \)

This solution comes at a cost — in an effort to remove the equality constraints, we had to add in dual variables, one for each equality constraint. With this in mind, let’s try to do the same for the inequality constraints. Adding in dual variable \( \lambda_i \) to represent each inequality constraint, we now have

\[
\begin{align*}
\min_{x} \max_{\lambda, \nu} \mathcal{L}(x, \lambda, \nu) &= f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \nu_j h_j(x) \\
\text{s.t.} \quad &\lambda_i \geq 0 \quad i = 1, \ldots, m
\end{align*}
\]

(5.12)

For convenience, we can place the constraints involving \( \lambda \) into the optimization variable.

\[
\begin{align*}
\min_{x} \max_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu) &= f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \nu_j h_j(x) \\
\end{align*}
\]

(5.13)

This optimization problem above is otherwise known as the **primal** (not to be confused with the **primal variables**), and its optimal value is indeed equivalent to that of the original constrained optimization problem.

\[
p^* = \min_{x} \max_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu)
\]

(5.14)

We can verify that this is indeed the case:

• For any \( x \) that violates the inequality constraints (meaning that \( \exists i \in \{1 .. m\} \) s.t. \( f_i(x) > 0 \)), the inner maximization problem over \( \lambda \) can choose \( \lambda_i \) as \( \infty \) to cause the inner maximization go to \( \infty \), therefore being ignored by the outer minimization over \( x \)

• For any \( x \) that violates the equality constraints (meaning that \( \exists j \) s.t. \( h_j(x) \neq 0 \)), the inner maximization problem over \( \nu \) can choose \( \nu_j \) as \( \infty \) if \( h_j(x) > 0 \) (or \( \nu_j \) as \( -\infty \) if \( h_j(x) < 0 \)) to cause the inner maximization go to \( \infty \), therefore being ignored by the outer minimization over \( x \)

• For any \( x \) that does not violate any of the equality or inequality constraints, in the inner maximization problem over \( \nu \), the expression \( \sum_{j=1}^{n} \nu_j h_j(x) \) evaluates to 0 no matter what the value of \( \nu \) is, and in the inner maximization problem over \( \lambda \), the expression \( \sum_{i=1}^{m} \lambda_i f_i(x) \) can at maximum be 0, because \( \lambda_i \) is constrained to be non-negative, and \( f_i(x) \) is non-positive. Therefore, at best, the maximization problem sets \( \lambda_i f_i(x) = 0 \), and

\[
\max_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu) = f_0(x)
\]

In its full form, the objective \( \mathcal{L}(x, \lambda, \nu) \) is called the **Lagrangian**, and it takes into account the unconstrained set of primal variables \( x \in \mathbb{R}^d \), the constrained set of dual variables \( \lambda \in \mathbb{R}^n \) corresponding to the inequality constraints, and the unconstrained set of dual variables \( \nu \in \mathbb{R}^m \) corresponding to the equality constraints. Note that our dual variables \( \lambda_i \) are in fact constrained, so ultimately we were not able to turn the original optimization problem into an unconstrained one, but our constraints are much simpler than before.

The **dual** of this optimization problem is still over the same optimization objective, except that now we swap the order of the maximization of the dual variables and the minimization of the primal variables.

\[
d^* = \max_{\lambda \geq 0, \nu} \min_{x} \mathcal{L}(x, \lambda, \nu) = \max_{\lambda \geq 0, \nu} g(\lambda, \nu)
\]

(5.15)
The dual is effectively a maximization problem (over the dual variables):

\[ d^* = \max_{\lambda \geq 0, \nu} g(\lambda, \nu) \]  

(5.16)

where

\[ g(\lambda, \nu) = \min_x \mathcal{L}(x, \lambda, \nu) \]  

(5.17)

The dual is very useful to work with, because now the inner optimization problem over \( x \) is an unconstrained problem! Furthermore, the dual \( g(\lambda, \nu) \) is always a concave function, regardless of the primal objective function or its constraints. This is because the dual is a pointwise minimum of concave functions, which itself is a concave function. Specifically \( g(\lambda, \nu) = \min_x \mathcal{L}(x, \lambda, \nu) \) is a pointwise minimum of functions \( \mathcal{L}(x, \lambda, \nu) \) that are affine in the dual variables (which are both concave and convex at the same time).

**Strong Duality and KKT Conditions**

Let’s examine the relationship between the primal and dual problem. It is always true that the solution to the primal problem is at least as large as the solution to the dual problem:

\[ p^* \geq d^* \]  

(5.18)

This condition is know as **weak duality**.

**Proof.** We know that

\[ \forall x, \lambda \geq 0, \nu \quad \max_{\tilde{\lambda} \geq 0, \tilde{\nu}} \mathcal{L}(x, \tilde{\lambda}, \tilde{\nu}) \geq \mathcal{L}(x, \lambda, \nu) \geq \min_x \mathcal{L}(\tilde{x}, \lambda, \nu) \]

More compactly,

\[ \forall x, \lambda \geq 0, \nu \quad \max_{\tilde{\lambda} \geq 0, \tilde{\nu}} \mathcal{L}(x, \tilde{\lambda}, \tilde{\nu}) \geq \min_x \mathcal{L}(\tilde{x}, \lambda, \nu) \]

Since this is true for all \( x, \lambda \geq 0, \nu \) this is true in particular when we set

\[ x = \arg \min_x \max_{\tilde{\lambda} \geq 0, \tilde{\nu}} \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \]

and

\[ \lambda, \nu = \arg \max_{\tilde{\lambda} \geq 0, \tilde{\nu}} \min_x \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \]

We therefore know that

\[ p^* = \min_x \max_{\tilde{\lambda} \geq 0, \tilde{\nu}} \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \geq \max_{\tilde{\lambda} \geq 0, \tilde{\nu}} \min_x \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) = d^* \]

\[ \square \]

The difference \( p^* - d^* \) is known as the **duality gap**. In the case of **strong duality**, the duality gap is 0. That is, we can swap the order of the minimization and maximization and up with the same optimal value:

\[ p^* = d^* \]  

(5.19)

There are several useful theorems detailing the existence of strong duality, such as **Slater’s theorem**, which states that if the primal problem is convex, and there exists an \( x \) that can strictly meet the inequality constraints and meet the equality constraints, then strong duality holds. Given that strong duality holds, the **Karush-Kuhn-Tucker (KKT) conditions** can help us find the solution to the dual variables of the optimization problem. The KKT conditions are composed of:
1. Primal feasibility (inequalities)
   \[ f_i(x) \leq 0, \forall i \in \{1 \ldots m\} \]

2. Primal feasibility (equalities)
   \[ h_j(x) = 0, \forall j \in \{1 \ldots n\} \]

3. Dual feasibility
   \[ \lambda_i \geq 0, \forall i \in \{1 \ldots m\} \]

4. Complementary Slackness
   \[ \lambda_i f_i(x) = 0, \forall i \in \{1 \ldots m\} \]

5. Stationarity
   \[ \nabla_x f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla_x f_i(x) + \sum_{j=1}^{n} \nu_j \nabla_x h_j(x) = 0 \]

Let’s see how the KKT conditions relate to strong duality.

**Theorem 1.** If \( x^* \) and \( \lambda^*, \nu^* \) are the primal and dual solutions respectively, with zero duality gap (i.e. strong duality holds), then \( x^*, \lambda^*, \nu^* \) also satisfy the KKT conditions.

**Proof.** KKT conditions 1, 2, 3 are trivially true, because the primal solution \( x^* \) must satisfy the primal constraints, and the dual solution \( \lambda^*, \nu^* \) must satisfy the dual constraints. Now, let’s prove conditions 4 and 5. We know that since strong duality holds, we can say that

\[
p^* = f_0(x^*) = g(\lambda^*, \nu^*) = d^*
\]

\[
= \min_x \mathcal{L}(x, \lambda^*, \nu^*)
\]

\[
\leq \mathcal{L}(x^*, \lambda^*, \nu^*)
\]

\[
= f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{j=1}^{n} \nu_j^* h_j(x^*)
\]

\[
= f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*)
\]

\[
\leq f_0(x^*)
\]

We cancel the terms involving \( h_j(x^*) \) because we know that the primal solution must satisfy \( h_j(x^*) = 0 \). Furthermore, we know that \( \lambda_i^* f_i(x^*) \leq 0 \), because \( \lambda_i^* \geq 0 \) in order to satisfy the dual constraints, and \( f_i(x^*) \leq 0 \) in order to satisfy the primal constraints. Since we established that \( f_0(x^*) = \min_x \mathcal{L}(x, \lambda^*, \nu^*) \leq \mathcal{L}(x^*, \lambda^*, \nu^*) \leq f_0(x^*) \), we know that all of the inequalities hold with equality and therefore \( \mathcal{L}(x^*, \lambda^*, \nu^*) = \min_x \mathcal{L}(x, \lambda^*, \nu^*) \). This implies KKT condition 5 (stationarity), that

\[
\nabla_x f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_x f_i(x^*) + \sum_{j=1}^{n} \nu_j^* \nabla_x h_j(x^*) = 0
\]

Finally, note that due to the equality \( f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) = f_0(x^*) \), we know that \( \sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0 \). This combined with the fact that \( \forall i \lambda_i^* f_i(x^*) \leq 0 \), establishes KKT condition 4 (complementary slackness):

\[ \lambda_i^* f_i(x^*) = 0, \forall i \in \{1 \ldots m\} \]
The theorem above establishes that in the presence of strong duality, if the solutions are optimal, then they satisfy the KKT conditions. Let's prove a statement that is almost (but not quite) the converse, which will be much more helpful for solving optimization problems.

**Theorem 2.** If \( \bar{x} \) and \( \bar{\lambda}, \bar{\nu} \) satisfy the KKT conditions, and the primal problem is convex, then they are the optimal solutions to the primal and dual problems with zero duality gap.

**Proof.** If \( \bar{x} \) and \( \bar{\lambda}, \bar{\nu} \) satisfy KKT conditions 1, 2, 3 we know that they are at least feasible for the primal and dual problem. From the KKT stationarity condition we know that

\[
\nabla_x f_0(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla_x f_i(\bar{x}) + \sum_{j=1}^{n} \bar{\nu}_j \nabla_x h_j(\bar{x}) = 0
\]

Since the primal problem is convex, we know that \( \mathcal{L}(x, \lambda, \nu) \) is convex in \( x \), and if the gradient of \( \mathcal{L}(x, \lambda, \nu) \) at \( \bar{x} \) is 0, we know that

\[
\bar{x} = \arg \min_x \mathcal{L}(x, \bar{\lambda}, \bar{\nu})
\]

Therefore, we know that the optimal primal values for the primal problem optimize the inner optimization problem of the dual problem, and

\[
g(\bar{\lambda}, \bar{\nu}) = f_0(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i f_i(\bar{x}) + \sum_{j=1}^{n} \bar{\nu}_j h_j(\bar{x})
\]

By the primal feasibility conditions for \( h_j(x) \) and the complementary slackness condition, we know that

\[
g(\bar{\lambda}, \bar{\nu}) = f_0(\bar{x})
\]

Now, all we have to do is to prove that \( \bar{x} \) and \( \bar{\lambda}, \bar{\nu} \) are primal and dual optimal, respectively. Note that since weak duality always holds, we know that

\[
p^* \geq d^* = \max_{\lambda \geq 0, \nu} g(\lambda, \nu) \geq g(\bar{\lambda}, \bar{\nu}), \quad \forall \bar{\lambda} \geq 0, \bar{\nu}
\]

Since we know that \( p^* \geq g(\lambda, \nu) \), we can also say that

\[
f_0(x) - p^* \leq f_0(x) - g(\lambda, \nu)
\]

And if we have that \( f_0(\bar{x}) = g(\bar{\lambda}, \bar{\nu}) \) as we deduced earlier, then

\[
f_0(\bar{x}) - p^* \leq f_0(\bar{x}) - g(\bar{\lambda}, \bar{\nu}) = 0 \implies p^* \geq f_0(\bar{x})
\]

Since \( p^* \) is the minimum value for the primal problem, we can go further by saying that \( p^* \geq f_0(\bar{x}) \) holds with equality and

\[
p^* = f_0(\bar{x}) = g(\bar{\lambda}, \bar{\nu}) \leq d^*
\]

since it always holds that \( p^* \geq d^* \) we conclude that

\[
p^* = f_0(\bar{x}) = g(\bar{\lambda}, \bar{\nu}) = d^*
\]

Therefore, we have proven that \( \bar{x} \) and \( \bar{\lambda}, \bar{\nu} \) are primal and dual optimal respectively, with zero duality gap. We eventually arrived at the conclusion that strong duality does indeed hold. \( \square \)
Let’s pause for a second to understand what we’ve found so far. Given an optimization problem, its primal problem is an optimization problem over the primal variables, and its dual problem is an optimization problem over the dual variables. If strong duality holds, then we can solve the dual problem and arrive at the same optimal value. In order to solve the dual, we have to first solve the unconstrained inner optimization problem over the primal variables and then solve the constrained outer optimization problem over the dual variables. But how do we even know in the first place that strong duality holds? This is where KKT comes into play. If the primal problem is convex and the KKT conditions hold, we can solve for the dual variables easily and also verify strong duality does indeed hold. We shall do just that, in our discussion of dual ridge regression and dual SVMs.

**Dual Ridge Regression**

Let’s derive kernel ridge regression again, using duality this time. Recall the unconstrained ridge regression formulation:

$$\min_w \|Xw - y\|^2 + \lambda \|w\|^2$$

This formulation is not conducive to dualization, because it lacks constraints. We will add constraints by introducing a dummy variable $z = Xw - y$ that corresponds to equality constraints:

$$\min_{w,z} \|z\|^2 + \lambda \|w\|^2$$

$$\text{s.t. } z = Xw - y \quad (5.26)$$

Now we proceed to forming the dual problem. For the purposes of notation, note that we are using $\alpha$ in place of $\nu$, and there are no dual variables corresponding to $\lambda$ because there are no inequality constraints. The Lagrangian is

$$\mathcal{L}(w, z, \alpha) = \|z\|^2 + \lambda \|w\|^2 + \alpha^\top (Xw - y - z)$$

The dual problem is

$$\max_\alpha g(\alpha) \quad (5.27)$$

where

$$g(\alpha) = \min_{w,z} \|z\|^2 + \lambda \|w\|^2 + \alpha^\top (Xw - y - z) \quad (5.28)$$

Since the $g(\alpha)$ is a convex minimization problem over the variables $w$ and $z$, we can simply set the derivative to $0$ w.r.t. $w$ and $z$:

- $\nabla_w \mathcal{L} = 2\lambda w + X^\top \alpha = 0 \implies w^*(\alpha) = -\frac{1}{2\lambda} X^\top \alpha$. This tells us that $w^*$ is going to be a linear combination of the $x_i$’s.

- $\nabla_z \mathcal{L} = 2z - \alpha = 0 \implies z^*(\alpha) = \frac{1}{2} \alpha$.

Plugging these optimal values back into the optimization problem, we have that

$$g(\alpha) = \min_{w,z} \mathcal{L}(w, z, \alpha)$$

$$= \mathcal{L}(w^*(\alpha), z^*(\alpha), \alpha)$$

$$= \left\| \frac{1}{2} \alpha \right\|^2 + \lambda \left\| \frac{1}{2\lambda} X^\top \alpha \right\|^2 + \alpha^\top \left( X \left( -\frac{1}{2\lambda} X^\top \alpha \right) - y - \frac{1}{2} \alpha \right) \quad (5.31)$$
5.7. DUALITY

\[ -\frac{1}{4} \alpha^\top \alpha - \frac{1}{4\lambda} \alpha^\top XX^\top \alpha - \alpha^\top y \]  \tag{5.32}

Now, the dual problem is

\[ \max_{\alpha} g(\alpha) = \max_{\alpha} -\frac{1}{2} \alpha^\top \alpha - \frac{1}{2\lambda} \alpha^\top XX^\top \alpha - \alpha^\top y \]

Note that this problem is a maximization over a concave problem (similar to a minimization over a convex problem) and we can take the derivative w.r.t \( \alpha \) and set it to 0:

\[ \nabla_\alpha g(\alpha) = -\frac{1}{2} \alpha - \frac{1}{2\lambda} XX^\top \alpha - y = 0 \implies \alpha^* = -\frac{2}{\lambda}(XX^\top + \lambda I)^{-1} y \]

The optimal \( w^* \) is therefore given by

\[ w^* = \frac{1}{2\lambda} X^\top \alpha^* = X^\top (XX^\top + \lambda I)^{-1} y \]

Which exactly matches the expression we previously derived for kernel ridge regression! Note that while this solution is dual optimal, it may not be optimal for the primal problem. In order to ensure that it is primal optimal, we need to establish that strong duality holds. In this case the primal problem is convex, so we simply need to ensure that the KKT conditions hold. Since we are not dealing with any inequality conditions here, the only applicable conditions are primal feasibility for the equalities and stationarity. Indeed the primal equality constraints are met, since

\[ Xw^* - y - z^* = -\frac{1}{2\lambda} XX^\top \alpha^* - y - \frac{1}{2} \alpha^* \\
= -\frac{1}{2\lambda} (XX^\top + \lambda I)\alpha^* - y \\
= (XX^\top + \lambda I)(XX^\top + \lambda I)^{-1} y - y \\
= 0 \]

We already showed the stationarity conditions are met, when we were solving \( g(\alpha) = \min_{w,z} L(w, z, \alpha) \).

We conclude that \( w^* \) is indeed the optimal solution to the primal problem.

Dual SVMs

Previously in our investigation of SVMs, we formulated a constrained optimization problem that we can solve to find the optimal parameters for our hyperplane decision boundary. Recall the setup of soft-margin SVMs:

- \( y_i \)'s: \( \pm 1 \), representing positive or negative class
- \( x_i \)'s: feature vectors in \( \mathbb{R}^d \)
- \( \xi_i \)'s: slack variables representing how much an \( x_i \) is allowed to violate the margin
- \( C \): a hyperparameter describing how severely we penalize slack
- The optimization problem for \( w \in \mathbb{R}^d \) and \( b \in \mathbb{R} \), the parameters of the SVM:

\[ \min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \]

s.t. \[ y_i (w^\top x_i - b) \geq 1 - \xi_i \quad \forall i \]

\[ \xi_i \geq 0 \quad \forall i \]  \tag{5.33}
Now, let’s investigate the dual of this problem. The primal problem in standard form is

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

s.t. 

$$(1 - \xi_i) - y_i(w^T x_i - b) \leq 0 \quad \forall i$$

$$-\xi_i \leq 0 \quad \forall i$$

(5.34)

Let’s identify the primal and dual variables for the SVM problem. We will have

- **Primal variables** $w$, $b$, and $\xi_i$
- **Dual variables** $\alpha_i$ corresponding to each constraint of the form $y_i(w^T x_i - b) \geq 1 - \xi_i$
- **Dual variables** $\beta_i$ corresponding to each constraint of the form $\xi_i \geq 0$

For the purposes of notation, note that we are using $\alpha$ and $\beta$ in place of $\lambda$, and there are no dual variables corresponding to $\nu$ because there are no equality constraints. The Lagrangian for the SVM problem is

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i((1 - \xi_i) - y_i(w^T x_i - b)) + \sum_{i=1}^{n} \beta_i(-\xi_i)$$

(5.35)

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i y_i(w^T x_i - b) + \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} (C - \alpha_i - \beta_i) \xi_i$$

(5.36)

Thus, the dual is

$$\max_{\alpha \geq 0, \beta \geq 0} g(\alpha, \beta)$$

(5.37)

where

$$g(\alpha, \beta) = \min_{w,b,\xi} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i y_i(w^T x_i - b) + \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} (C - \alpha_i - \beta_i) \xi_i$$

(5.38)

Let’s use the KKT conditions to find the optimal dual variables. Verify that the primal problem is convex in the primal variables. We know that from the stationarity conditions, evaluated at the optimal dual variables, we can

- $\nabla_w L = w^* - \sum_{i=1}^{n} \alpha_i^* y_i x_i = 0 \implies w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$. This tells us that $w^*$ is going to be a weighted combination of the positive-class $x_i$’s and negative-class $x_i$’s.
- $\frac{\partial L}{\partial b} = 0$. This tells us that the weights $\alpha_i^*$ will be equally distributed among positive- and negative-class training points.
- $\frac{\partial L}{\partial \xi_i} = C - \alpha_i^* - \beta_i^* = 0 \implies 0 \leq \alpha_i^* \leq C$. This tells us that the weights $\alpha_i^*$ are restricted to being less than the hyperparameter $C$.

Verify that the other KKT also hold, establishing strong duality. Using these observations, we can eliminate some terms of the dual problem.

$$L(w, b, \xi, \alpha^*, \beta^*) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i^* y_i(w^T x_i - b) + \sum_{i=1}^{n} \alpha_i^* + \sum_{i=1}^{n} (C - \alpha_i^* - \beta_i^*) \xi_i$$

(5.39)
\[
\frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i^* y_i (w^\top x_i) + b \sum_{i=1}^{n} \alpha_i^* y_i + \sum_{i=1}^{n} \alpha_i^* + \sum_{i=1}^{n} (C - \alpha_i^* - \beta_i^*) \xi_i \tag{5.40}
\]

\[
= \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i^* y_i (w^\top x_i) + \sum_{i=1}^{n} \alpha_i^*
\tag{5.41}
\]

Since the primal problem is convex, from the KKT conditions we have that the optimal primal variables \(w^*, b^*, \xi^*\) minimize \(L(w, b, \xi, \alpha^*, \beta^*)\):

\[
g(\alpha^*, \beta^*) = \min_{w, b, \xi} L(w, b, \xi, \alpha^*, \beta^*) \tag{5.42}
\]

\[
= \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i^* y_i x_i \right\|^2 - \sum_{i=1}^{n} \alpha_i^* y_i \left( \left( \sum_{j=1}^{n} \alpha_j^* y_j x_j \right)^\top x_i \right) + \sum_{i=1}^{n} \alpha_i^* \tag{5.43}
\]

\[
= \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i^* y_i x_i \right\|^2 - \sum_{i=1}^{n} \left( \sum_{j=1}^{n} \alpha_j^* y_j x_j \right)^\top \left( \sum_{j=1}^{n} \alpha_j^* y_j x_j \right) + \sum_{i=1}^{n} \alpha_i^* \tag{5.44}
\]

\[
= \alpha^\top 1 - \frac{1}{2} \alpha^\top Q \alpha^* \tag{5.45}
\]

where \(Q_{ij} = y_i (x_i^\top x_j) y_j\) (and \(Q = (\text{diag } y)XX^\top (\text{diag } y)\)).

Now, we can write the final form of the dual, which is only in terms of \(\alpha\) and \(X\) and \(y\) (Note that we have eliminated all references to \(\beta\)):

\[
\max_{\alpha} \alpha^\top 1 - \frac{1}{2} \alpha^\top Q \alpha
\]

s.t.

\[
\sum_{i=1}^{n} \alpha_i y_i = 0
\]

\[
0 \leq \alpha_i \leq C \quad i = 1, \ldots, n
\tag{5.47}
\]

Remember to account for the constraints \(\sum_{i=1}^{n} \alpha_i y_i = 0\) and \(0 \leq \alpha_i \leq C\) that arise from the stationarity conditions. After all of this effort, we have managed to turn a minimization problem over the primal variables into a maximization problem over the dual variables. One might ask, why go through the effort to formulate and solve the dual problem instead? For one, the dual is an optimization problem over the number of training points \(n\) rather than the number of augmented features \(d\), making it particularity attractive when \(n \ll d\). Second, it incorporates the term \(XX^\top\) which is simply the Gram matrix \(K\) of kernel evaluations among all pairs of training points. We can apply the kernel trick to form this Gram matrix, effectively relying on the the dimensionality of the raw feature space rather than the augmented feature space. These are more or less the exact same justifications for kernel ridge regression.

**Geometric intuition**

We’ve formulated the dual SVM problem and used the KKT conditions to formulate an equivalent optimization problem, but what do these dual values \(\alpha_i\) even mean? That’s a good question!

We know that given optimal primal and dual values, the following KKT conditions are enforced:

- Stationarity

\[
C - \alpha_i^* - \beta_i^* = 0
\]
• Complementary slackness

\[ \alpha_i^* \cdot ((1 - \xi_i^*) - y_i(w^*^T x_i - b^*)) = 0 \]

\[ \beta_i^* \cdot \xi_i^* = 0 \]

Here are some noteworthy relationships between \( \alpha_i \) and the properties of the SVMs:

• **Case 1:** \( \alpha_i^* = 0 \). In this case, we know \( \beta_i^* = C \), which is nonzero, and therefore \( \xi_i^* = 0 \). That is, if for point \( i \) we have that \( \alpha_i^* = 0 \) by the dual problem, then we know that there is no slack given to this point. Looking at the other complementary slackness condition, this makes sense because if \( \alpha_i^* = 0 \), then \( y_i(w^*^T x_i - b^*) - (1 - \xi_i^*) \) may be any value, and if we’re minimizing the sum of our \( \xi_i \)'s, we should have \( \xi_i^* = 0 \). So, point \( i \) lies on or outside the margin.

• **Case 2:** \( \alpha_i^* \) is nonzero. If this is the case, then we know \( \beta_i^* = C - \alpha_i^* \geq 0 \)

  - **Case 2.1:** \( \alpha_i^* = C \). If this is the case, then we know \( \beta_i^* = 0 \), and therefore \( \xi_i^* \) may be exactly 0 or nonzero. So, point \( i \) lies on or violates the margin.

  - **Case 2.2:** \( 0 < \alpha_i^* < C \). In this case, then \( \beta_i^* \) is nonzero and \( \xi_i^* = 0 \). But this is different from Case 1 because with \( \alpha_i^* \) nonzero, we can divide by \( \alpha_i^* \) in the complementary slackness condition and arrive at the fact that \( 1 - y_i(w^*^T x_i - b^*) = 0 \) \( \Rightarrow \) \( y_i(w^*^T x_i - b^*) = 1 \), which means \( x_i \) lies exactly on the margin determined by \( w^* \) and \( b^* \). So, point \( i \) lies on the margin.

Using this information, let’s reconstruct the optimal primal values \( w^*, b^*, \xi_i^* \) from the optimal dual
values $\alpha^*$:

$$
\begin{align*}
\mathbf{w}^* &= \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i \\
b^* &= \mathbf{w}^{*\top} \mathbf{x}_i - y_i \quad \text{if } 0 < \alpha_i^* < C \\
\xi_i^* &= \begin{cases} 
1 - y_i (\mathbf{w}^{*\top} \mathbf{x}_i - b^*) & \text{if } \alpha_i^* = C, \\
0 & \text{otherwise}
\end{cases}
\end{align*}
$$

(5.48)

The principal takeaway is that the optimal $\mathbf{w}^*$ is a linear combination of the training points for which the corresponding dual weight $\alpha_i$ is non-zero. Such points are called support vectors, because they determine the optimal $\mathbf{w}^*$. There is a special relationship between the values of $\alpha_i$ and the position of $\mathbf{x}_i$ relative to the margin. All training points that violate the decision boundary have $\alpha_i > 0$ and are thus support vectors, while all training points that strictly do not violate the decision boundary (meaning that they do not lie on the boundary) have $\alpha_i = 0$ and are not support vectors. For training points which lie exactly on the boundary, some may have $\alpha_i > 0$ and some may have $\alpha_i = 0$; only the points that are critical to determining the decision boundary have $\alpha_i > 0$ and are thus support vectors. Intuitively, there are very few support vectors compared to the total number of training points, meaning that the dual vector $\alpha^*$ is sparse. This is advantageous when predicting class for a test point:

$$
\mathbf{w}^{*\top} \phi(\mathbf{x}) + b^* = \sum_{i=1}^{n} \alpha_i^* y_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}) + b^* = \sum_{i=1}^{n} \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}) + b^*
$$

We only have to make $m \ll n$ kernel evaluations to predict a test point, where $m$ is the number of support vectors. It should now be clear why the dual SVM problem is so useful: it allows us to use the kernel trick to eliminate dependence on the dimensionality of the argument feature space, while also allowing us to discard most training points because they have dual weight 0.

## 5.8 Nearest Neighbor Classification

In classification, it is reasonable to conjecture that data points that are sufficiently close to one another should be of the same class. For example, in fruit classification, perturbing a few pixels in an image of a banana should still result in something that looks like a banana. The k-nearest-neighbors (k-NN) classifier is based on this observation. Assuming that there is no preprocessing of the training data, the training time for k-NN is effectively $O(1)$. To train this classifier, we simply store our training data for future reference.\(^1\) For this reason, k-NN is sometimes referred to as “lazy learning.” The major work of k-NNs in done at testing time: to predict on a test data point $\mathbf{z}$, we compute the $k$ closest training data points to $\mathbf{z}$, where “closeness” can be quantified in some distance function such as Euclidean distance — these are the $k$ nearest neighbors to $\mathbf{z}$. We then find the most common class $y$ among these $k$ neighbors and classify $\mathbf{z}$ as $y$ (that is, we perform a majority vote). For binary classification, $k$ is usually chosen to be odd so we can break ties cleanly. Note that k-NN can also be applied to regression tasks — in that case k-NN would return the average label of the $k$ nearest points.

\(^1\)Sometimes we store the data in a specialized structure called a k-d tree. This data structure is out of scope for this course, but it usually allows for faster (average-case $O(\log n)$) nearest neighbors queries.
Choosing \( k \)

Nearest neighbors can produce very complex decision functions, and its behavior is highly dependent on the choice of \( k \).

Choosing \( k = 1 \), we achieve an optimal training error of 0 because each training point will classify as itself, thus achieving 100% accuracy on itself. However, \( k = 1 \) overfits to the training data, and is a terrible choice in the context of the bias-variance tradeoff. Increasing \( k \) leads to an increase in training error, but a decrease in testing error and achieves better generalization. At one point, if \( k \) becomes too large, the algorithm will underfit the training data, and suffer from huge bias. In general, in order to select \( k \) we use cross-validation.
**Bias-Variance Analysis**

Let’s justify this reasoning formally for k-NN applied to regression tasks. Suppose we are given a training dataset $D = \{(x_i, y_i)\}_{i=1}^n$, where the labels $y_i$ are real valued scalars. We model our hypothesis $h(z)$ as

$$h(z) = \frac{1}{k} \sum_{i=1}^n N(x_i, z, k)$$

where the function $N$ is defined as

$$N(x_i, z, k) = \begin{cases} y_i & \text{if } x_i \text{ is one of the } k \text{ closest points to } z \\ 0 & \text{o.w.} \end{cases}$$

Suppose also we assume our labels $y_i = f(x_i) + \epsilon$, where $\epsilon$ is the noise that comes from $\mathcal{N}(0, \sigma^2)$ and $f$ is the true function. Without loss of generality, let $x_1 \ldots x_k$ be the $k$ closest points. Let’s first derive the bias$^2$ of our model for the given dataset $D$.

$$\left( \mathbb{E}[h(z)] - f(z) \right)^2 = \left( \mathbb{E} \left[ \frac{1}{k} \sum_{i=1}^n N(x_i, z, k) \right] - f(z) \right)^2 = \left( \mathbb{E} \left[ \frac{1}{k} \sum_{i=1}^k y_i \right] - f(z) \right)^2$$

$$= \left( \frac{1}{k} \sum_{i=1}^k \mathbb{E}[y_i] - f(z) \right)^2 = \left( \frac{1}{k} \sum_{i=1}^k \mathbb{E}[f(x_i) + \epsilon] - f(z) \right)^2$$

$$= \left( \frac{1}{k} \sum_{i=1}^k f(x_i) - f(z) \right)^2$$

When $k \to \infty$, then $\frac{1}{k} \sum_{i=1}^k f(x_i)$ goes to the average label for $x$. When $k = 1$, then the bias is simply $f(x_1) - f(z)$. Assuming $x_1$ is close enough to $f(z)$, the bias would likely be small when
$k = 1$ since it’s likely to share a similar label. Meanwhile, when $k \to \infty$, the bias doesn’t depend on the training points at all which like will restrict it to be higher.

Now, let’s derive the variance of our model.

$$\text{Var}[h(z)] = \text{Var} \left[ \frac{1}{k} \sum_{i=1}^{k} y_i \right] = \frac{1}{k^2} \sum_{i=1}^{k} \text{Var}[f(x_i) + \epsilon]$$

$$= \frac{1}{k^2} \sum_{i=1}^{k} \text{Var}[\epsilon]$$

$$= \frac{1}{k^2} \sum_{i=1}^{k} \sigma^2 = \frac{1}{k^2} k \sigma^2 = \frac{\sigma^2}{k}$$

The variance goes to 0 when $k \to \infty$, and is maximized at $k = 1$.

Properties

Computational complexity: We require $O(n)$ space to store a training set of size $n$. There is no runtime cost during training if we do not use specialized data structures to store the data. However, predictions take $O(n)$ time, which is costly. There has been research into approximate nearest neighbors (ANN) procedures that quickly find an approximation for the nearest neighbor - some common ANN methods are Locality-Sensitive Hashing and algorithms that perform dimensionality reduction via randomized (Johnson-Lindenstrauss) distance-preserving projections.

Flexibility: When $k > 1$, k-NN can be modified to output predicted probabilities $P(Y|X)$ by defining $P(Y|X)$ as the proportion of nearest neighbors to $X$ in the training set that have class $Y$. k-NN can also be adapted for regression — instead of taking the majority vote, take the average of the $y$ values for the nearest neighbors. k-NN can learn very complicated, non-linear decision boundaries.

Non-parametric: k-NN is a non-parametric method, which means that the number of parameters in the model grows with $n$, the number of training points. This is as opposed to parametric methods, for which the number of parameters is independent of $n$. Some examples of parametric models include linear regression, LDA, and neural networks.

Behavior in high dimensions: k-NN does not behave well in high dimensions. As the dimension increases, data points drift farther apart, so even the nearest neighbor to a point will tend to be very far away.

Theoretical properties: 1-NN has impressive theoretical guarantees for such a simple method. Cover and Hart, 1967 prove that as the number of training samples $n$ approaches infinity, the expected prediction error for 1-NN is upper bounded by $2\epsilon^*$, where $\epsilon^*$ is the Bayes (optimal) error. Fix and Hodges, 1951 prove that as $n$ and $k$ approach infinity and if $\frac{k}{n} \to 0$, then the $k$ nearest neighbor error approaches the Bayes error.

Curse of Dimensionality

To understand why k-NN does not perform well in high-dimensional space, we first need to understand the properties of metric spaces. In high-dimensional spaces, much of our low-dimensional

---

2ANN methods are beyond the scope of this course, but are useful in real applications.
intuition breaks down. Here is one classical example. Consider a ball in $\mathbb{R}^d$ centered at the origin with radius $r$, and suppose we have another ball of radius $r - \epsilon$ centered at the origin. In low dimensions, we can visually see that much of the volume of the outer ball is also in the inner ball. In general, the volume of the outer ball is proportional to $r^d$, while the volume of the inner ball is proportional to $(r - \epsilon)^d$. Thus the ratio of the volume of the inner ball to that of the outer ball is

$$\frac{(r - \epsilon)^d}{r^d} = \left(1 - \frac{\epsilon}{r}\right)^d \approx e^{-\epsilon d/r} \quad d \to \infty$$

Hence as $d$ gets large, most of the volume of the outer ball is concentrated in the annular region $\{x : r - \epsilon < x < r\}$ instead of the inner ball.

High dimensions also make Gaussian distributions behave counter-intuitively. Suppose $X \sim \mathcal{N}(0, \sigma^2 I)$. If $X_i$ are the components of $X$ and $R$ is the distance from $X$ to the origin, then $R^2 = \sum_{i=1}^d X_i^2$. We have $E(R^2) = d\sigma^2$, so in expectation a random Gaussian will actually be reasonably far from the origin. If $\sigma = 1$, then $R^2$ is distributed chi-squared with $d$ degrees of freedom. One can show that in high dimensions, with high probability $1 - O(e^{-d})$, this multivariate Gaussian will lie within the annular region $\{X : |R^2 - E(R^2)| \leq d^{1/2+\epsilon}\}$ where $E(R^2) = d\sigma^2$ (one possible approach is to note that as $d \to \infty$, the chi-squared approaches a Gaussian by the CLT, and use a Chernoff bound to show exponential decay). This phenomenon is known as concentration of measure. Without resorting to more complicated inequalities, we can show a simple, weaker result:

**Theorem:** If $X_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, ..., d$ are independent and $R^2 = \sum_{i=1}^d X_i^2$, then for every $\epsilon > 0$, the following holds:

$$\lim_{d \to \infty} P(|R^2 - E(R^2)| \geq d^{\frac{1}{2} + \epsilon}) = 0$$

Thus in the limit, the squared radius is concentrated about its mean.

**Proof.** From the formula for the variance of a chi-squared distribution, we see that $\text{Var}(R^2) = 2d\sigma^4$. Applying a Chebyshev bound yields

$$P(|R^2 - E(R^2)| \geq d^{\frac{1}{2} + \epsilon}) \leq \frac{2d\sigma^4}{d^{1+2\epsilon}} \to 0 \quad d \to \infty$$

Thus a random Gaussian will lie within a thin annular region away from the origin in high dimensions with high probability, even though the mode of the Gaussian bell curve is at the origin. This
illustrates the phenomenon in high dimensions where random data is spread very far apart. The k-NN classifier was conceived on the principle that nearby points should be of the same class - however, in high dimensions, even the nearest neighbors that we have to a random test point will tend to be far away, so this principle is no longer useful.

**Improving k-NN**

There are two main ways to improve k-NN and overcome the shortcomings we have discussed.

1. Obtain more training data.

2. Reduce the dimensionality of the features and/or pick better features. Consider other choices of distance function.

One example of reducing the dimensionality in image space is to lower the resolution of the image — while this is throwing some of the original pixel features away, we may still be able to get the same or better performance with a nearest neighbors method.

We can also modify the distance function. For example, we have a whole family of Minkowski distances that are induced by the $L^p$ norms:

$$D_p(x, z) = \left( \sum_{i=1}^{d} |x_i - z_i|^p \right)^{\frac{1}{p}}$$

Without preprocessing the data, 1-NN with the $L^3$ distance outperforms 1-NN with $L^2$ on MNIST.

We can also use kernels to compute distances in a different feature space. For example, if $k$ is a kernel with associated feature map $\Phi$ and we want to compute the Euclidean distance from $\Phi(x)$ to $\Phi(z)$, then we have

$$\|\Phi(x) - \Phi(z)\|_2^2 = \Phi(x)^\top\Phi(x) - 2\Phi(x)^\top\Phi(z) + \Phi(z)^\top\Phi(z)$$

$$= k(x, x) - 2k(x, z) + k(z, z)$$

Thus if we define $D(x, z) = \sqrt{k(x, x) - 2k(x, z) + k(z, z)}$, then we can perform Euclidean nearest neighbors in $\Phi$-space without explicitly representing $\Phi$ by using the kernelized distance function $D$. 
Chapter 6

Clustering

In the problem of clustering, we are given a dataset comprised only of input features without labels. We wish to assign to each data point a discrete label indicating which “cluster” it belongs to, in such a way that the resulting cluster assignment “fits” the data. We are given flexibility to choose our notion of goodness of fit for cluster assignments.

![Figure 6.1: Left: unclustered raw data; Right: clustered data](image)

Figure 6.1: Left: unclustered raw data; Right: clustered data

![Figure 6.2: A nonspherical clustering assignment. Possible outliers are shown in black.](image)

Figure 6.2: A nonspherical clustering assignment. Possible outliers are shown in black.¹

In our discussion of LDA and QDA, we assumed that we had data which was conditionally Gaussian

given a discrete class label. When we observed a data point, we observed both its input features and its class label. These are supervised learning methods, which deal with prediction of observed outputs from observed inputs. Clustering is an example of unsupervised learning, where we are not given labels and desire to infer something about the underlying structure of the data. Another example of unsupervised learning is dimensionality reduction, where we desire to learn important features from the data.

Clustering is most often used in exploratory data visualization, as it allows us to see the different groups of similar data points within the data. Combined with domain knowledge, these clusters can have a physical interpretation - for example, different clusters can represent different species of plant in the biological setting, or types of consumers in a business setting. If desired, these clusters can be used as pre-processing to make the data more compact. Clustering is also used for outlier detection, as in Figure 6.2: data points that do not seem to belong in their assigned cluster may be flagged as outliers.

In order to create an algorithm for clustering, we first must determine what makes a good clustering assignment. Here are some possible desired properties:

1. High intra-cluster similarity - points within a given cluster are very similar.
2. Low inter-cluster similarity - points in different clusters are not very similar.

Of course, this depends on our notion of similarity. For now, we will say that points in \( \mathbb{R}^d \) are similar if their \( L^2 \) distance is small, and dissimilar otherwise. A generalization of this notion is provided in the appendix.

## 6.1 K-means Clustering

Let \( X \) denote the set of \( N \) data points \( x_i \in \mathbb{R}^d \). A cluster assignment is a partition \( C_1, \ldots, C_K \subseteq X \) such that the sets \( C_k \) are disjoint and \( X = C_1 \cup \cdots \cup C_K \). A data point \( x \in X \) is said to belong to cluster \( k \) if it is in \( C_k \).

One approach to the clustering problem is to represent each cluster \( C_k \) by a single point \( c_k \in \mathbb{R}^d \) in the input space - this is called the centroid approach. K-means is an example of centroid-based clustering where we choose centroids and a cluster assignment such that the total distance of each point to its assigned centroid is minimized. In this regard, K-means optimizes for high intra-cluster similarity, but the clusters do not necessarily need to be far apart, so we may also have high inter-cluster similarity.

Formally, K-means solves the following problem:

\[
\arg \min_{\{C_k\}_{k=1}^{K}, \{c_k\}_{k=1}^{K}: X = C_1 \cup \cdots \cup C_K} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - c_k\|^2
\]

It has been shown that this problem is NP hard, so solving it exactly is intractable. However, we can come up with a simple algorithm to compute a candidate solution. If we knew the cluster assignment \( C_1, \ldots, C_K \), then we would only need to determine the centroid locations. Since the choice of centroid location \( c_i \) does not affect the distances of points in \( C_j \) to \( c_j \) for \( i \neq j \), we can consider each cluster separately and choose the centroid that minimizes the sum of squared distances to points in that cluster. The centroid we compute, \( \hat{c}_k \), is

\[
\hat{c}_k = \arg \min_{c_k} \sum_{x \in C_k} \|x - c_k\|^2
\]
But this is simply the mean of the data in $C_k$, that is,

$$\hat{c}_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$$

Similarly, if we knew the centroids $c_k$, in order to choose the cluster assignment $C_1, ..., C_K$ that minimizes the sum of squared distances to the centroids, we simply assign each data point $x$ to the cluster represented by its closest centroid, that is, we assign $x$ to

$$\arg \min_k \|x - c_k\|^2$$

Now we can perform alternating minimization - on each iteration of our algorithm, we update the clusters using the current centroids, and then update the centroids using the new clusters. This algorithm is sometimes called Lloyd’s Algorithm.

**Algorithm 8: K-means Algorithm**

Initialize $c_k, k = 1, ..., K$

while K-means objective has not converged do

- Update partition $C_1 \cup \cdots \cup C_K$ given the $c_k$ by assigning each $x \in X$ to the cluster represented by its nearest centroid
- Update centroids $c_k$ given $C_1 \cup \cdots \cup C_K$ as $c_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$

This algorithm will always converge to some value. To show this, note the following facts:

1. There are only finitely many (say, $M$) possible partition/centroid pairs that can be produced by the algorithm. This is true since each partition chosen at some iteration in the algorithm completely determines the subsequent centroid assignment in that iteration.

2. Each update of the cluster assignment and centroids does not increase the value of the objective. This is true since each of these updates is a minimization of the objective which we solve exactly.

If the value of the objective has not converged after $M$ iterations, then we have cycled through all the possible partition/centroid pairs attainable by the algorithm. On the next iteration, we would obtain a partition and centroid assignment that we have already seen, say on iteration $t \leq M$. But this means that the value of the objective at time $M + 1$ is the same as at time $t$, and because the value of the objective function never increases throughout the algorithm, the value is the same as at time $M$, so we have converged.

In practice, it is common to run the K-means algorithm multiple times with different initialization points, and the cluster corresponding to the minimum objective value is chosen. There are also ways to choose a smarter initialization than a random seed, which can improve the quality of the local optimum found by the algorithm.\(^2\) It should be emphasized that no efficient algorithm for solving the K-means optimization is guaranteed to give a good cluster assignment, as the problem is NP hard and there are local optima.

Choosing the number of clusters $k$ is similar to choosing the number of principal components for PCA - we can compute the value of the objective for multiple values of $k$ and find the “elbow” in the curve.

\(^2\)For example, K-means++.
We have noted that the main algorithm for solving K-means does not have to produce a good solution. Let us step back and consider some shortcomings of the K-means objective function itself:

1. There is no likelihood attached to K-means, which makes it harder to understand what assumptions we are making on the data.
2. Each feature is treated equally, so the clusters produced by K-means will look spherical. We can also infer this by looking at the sum of squares in the objective function, which we have seen to be related to spherical Gaussians.
3. Each cluster assignment in the optimization is a hard assignment - each point belongs in exactly one cluster. A soft assignment would assign each point to a distribution over the clusters, which can encode not only which cluster a point belongs to, but also how far it was from the other clusters.

**Soft K-means**

We can introduce soft assignments to our algorithm easily using the familiar softmax function. Recall that if \( z \in \mathbb{R}^d \), then the softmax function \( \sigma \) is defined as:

\[
\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{d} e^{z_k}}
\]

To compute the cluster assignment of a data point \( x_i \) in K-means, we computed

\[
\arg \min_k \|x_i - c_k\|^2 = \arg \max_k -\|x_i - c_k\|^2
\]

In soft K-means, we instead compute a soft assignment \( r_i(k), k = 1, ..., K \) where \( \sum_k r_i(k) = 1 \) as the softmax of the vector of \( z := -\beta \|x_i - c_k\|^2, k = 1, ..., K \):

\[
r_i(k) = \sigma(z)_k
\]

Here, \( \beta \) is a tunable parameter indicating the level of “softness” desired.

Once we have computed the soft assignments, we may use them to determine the centroids by using a weighted average. Before, we defined the new centroids as

\[
\hat{c}_k = \arg \min_{c_k} \sum_{x \in C_k} \|x - c_k\|^2
\]

Now, we apply the \( r_i(k) \) as weights for the minimization:

\[
\hat{c}_k = \arg \min_{c_k} \sum_{i=1}^{N} r_i(k)\|x_i - c_k\|^2 = \frac{\sum_{i=1}^{N} r_i(k)x_i}{\sum_{i=1}^{N} r_i(k)} \tag{6.1}
\]

This is now a weighted average of the \( x_i \) - the weights reflect how much we believe each data point belongs to a particular cluster, and because we are using this information, our algorithm should not jump around between clusters, resulting in better convergence speed.

There are still a few issues with soft K-means. One is the choice of \( \beta \) - it is not so clear how to set this hyperparameter. Another issue is that our clusters are still spherical, since we are still weighting all features the same (note that we have weighted each data point differently with soft K-means). To solve these issues, we will use a fully probabilistic model.
6.2 Mixture of Gaussians

Suppose $\mu_k \in \mathbb{R}^d, \Sigma_k \in \mathbb{R}^{d \times d}$ are fixed parameters for $k = 1, \ldots, K$. Consider the following experiment: draw a value $z$ from some distribution on the set of indices $\{1, \ldots, K\}$, and then draw $x \in \mathbb{R}^d$ from the Gaussian distribution $\mathcal{N}(\mu_z, \Sigma_z)$. We can interpret $x$ as belonging to cluster $z$. This model is called Mixture of Gaussians (MoG), also known as a Gaussian mixture model.

If we have fit a MoG model to data (i.e. we have determined values for $\mu_k, \Sigma_k$, and the prior on $z$), then to perform clustering, we can use Bayes’ rule to determine the posterior $P(z = k | x)$ and assign $x$ to the cluster $k$ that maximizes this quantity. In fact, this is exactly our decision rule with QDA using a prior - the difference is that QDA, a supervised method, is given labels to fit the mixture model, while in the unsupervised clustering setting we must fit the mixture model without the aid of labels. When $\Sigma_k$ are not multiples of the identity, we can obtain non-spherical clusters, which was not possible with K-means.

MoG is an example of a latent variable model. A latent variable model is a probabilistic model in which some variables can be directly observed or measured, while other latent (hidden) variables cannot be observed directly; rather, we observe them indirectly through their influence on the observed variables. When we try to fit a MoG model to data, we only observe the data $x_i$, which we presume to have been generated based on the latent variable $z_i$, the cluster assignment. Latent variable models are modular and can be used to model complex dependencies in a probabilistic model - however, the added flexibility can lead to difficulty in learning its parameters.

To illustrate this, we will examine the likelihood function for MoG. Suppose $x_i$ has distribution $p(x_i; \theta)$, where $\theta$ is a set of all $\mu_k, \Sigma_k, p(z_i = k)$. The likelihood for the single data point $x_i$ is

$$L_i(\theta; x_i) = p(x_i; \theta) = \sum_{k=1}^{K} p(x_i, z_i = k; \theta) = \sum_{k=1}^{K} p(x_i | z_i = k; \theta) p(z_i = k; \theta)$$

Over all data points, the likelihood is

$$L(\theta; x) = \prod_{i=1}^{N} L_i(\theta; x_i) = \prod_{i=1}^{N} \sum_{k=1}^{K} p(x_i | z_i = k; \theta) p(z_i = k; \theta)$$

Hence the log likelihood $\ell(\theta; x)$ is given by

$$\ell(\theta; x) = \sum_{i=1}^{N} \log \left( \sum_{k=1}^{K} p(x_i | z_i = k; \theta) p(z_i = k; \theta) \right)$$ (6.2)

When we perform QDA, we know the $z_i$ are known, deterministic quantities and thus the likelihood (6.2) reduces to

$$\ell(\theta; X) = \sum_{i=1}^{N} \log p(x_i | z_i; \theta)$$
Maximizing this is equivalent to fitting the individual class-conditional Gaussians via maximum likelihood, which is consistent with how we have described QDA in the past. When we fit the MoG without knowledge of the latent variables, the parameters $\theta$ in (6.2) are now coupled together inside the log, which complicates the likelihood. While it is still possible to find the MLE by working out the gradient and using our descent methods, there is an alternative approach called Expectation-Maximization (EM), which takes advantage of the latent variable structure.

### 6.3 Expectation Maximization (EM) Algorithm

The **Expectation-Maximization (EM) Algorithm** is used to compute the MLE for latent variable models, such as MoG. First recall that soft K-means consisted of the following two alternating steps:

1. For each data point, compute a soft assignment $r_i(k)$ to the clusters - that is, a probability distribution over clusters. The soft assignment is obtained by using a softmax.
2. Update the centroids in an optimal way given the soft assignments computed in the first step. The resulting updates are a weighted average of the data points.

One could rephrase these updates as:

1. Soft imputation of the data - fill in the missing data (“impute”) with a probability distribution over all its possible values (“soft”).
2. Parameter updates given the imputed data.

The EM algorithm alternates between these two steps in the same way as soft K-means, but the updates for each step are performed in a principled way to maximize certain “components” of the log likelihood (we will make this precise later). We will derive the following updates for EM when computing the MLE for MoG:

1. Using the current parameter estimates, estimate $p(z_i|x_i; \theta)$. That is, perform soft imputation of the latent cluster variable.
2. Estimate the parameters via MLE, using the estimates of $p(z_i|x_i; \theta)$ to make the computation tractable.

Both soft K-means and EM estimate $p(z_i|x_i; \theta)$, though EM will do so in a more principled way. In the second alternating step, we will see that the EM update of the mean is exactly the same as the soft K-means centroid update. However, EM will also update the covariance estimates, which captures ellipsoidal structure in the data.

To derive the EM algorithm, it will be helpful to introduce the notion of the **complete log likelihood**, which we define as

$$L_c(x_i, z_i; \theta) := \log p(x_i, z_i; \theta)$$

If we assumed $z_i$ to be known, this would be the log likelihood of the data. In practice, we do not know the values of $z_i$, but if we are given a distribution $q(z_i|x_i)$ over the latents $z_i$, we can marginalize over the possible values of $z_i$ by taking the expectation

$$E_q(L_c(x_i, z_i; \theta)) = \sum_{k=1}^{K} L_c(x_i, z_i = k; \theta)q(z_i = k|x_i)$$
6.3. EXPECTATION MAXIMIZATION (EM) ALGORITHM

We call this the **expected complete log likelihood**. The distribution \( q \) is an estimate of the true conditional distribution \( p(z_i|x_i; \theta) \); in the EM algorithm, we will alternate between updating our \( q \) distribution to better estimate \( p(z_i|x_i; \theta) \) and maximizing the expected complete log likelihood in place of the true likelihood function.

We are now in a position to derive the algorithm. We will need a well-known result called **Jensen’s Inequality**: 

**Theorem 3.** If \( X \) is a random variable and \( f \) is convex, then \( f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)) \).

If \( f \) is concave, then using the fact \(-f\) is convex immediately yields the conclusion \( \mathbb{E}(f(X)) \leq f(\mathbb{E}(X)) \). In particular, since \( \log \) is concave, we have \( \mathbb{E}(\log(X)) \leq \log(\mathbb{E}(X)) \).

Now we can derive the EM algorithm. Suppose \( x_i \) are random variables depending on \( z_i \), and \( \theta \) are the parameters of interest. Given any conditional distribution over the latents \( q(z_i = k|x_i) \), the log likelihood for the \( i \)-th data point is

\[
\ell_i(\theta; x_i) = \log p(x_i; \theta) \\
= \log \sum_{k=1}^{K} p(x_i, z_i = k; \theta) \\
= \log \sum_{k=1}^{K} q(z_i = k|x_i) p(x_i, z_i = k; \theta) / q(z_i = k|x_i) \\
= \log \mathbb{E}_q \left[ \frac{p(x_i, z_i; \theta)}{q(z_i|x_i)} \right] \quad \text{(Jensen)} \\
\geq \mathbb{E}_q \left[ \log \frac{p(x_i, z_i; \theta)}{q(z_i|x_i)} \right] \\
= - \sum_{k=1}^{K} q(z_i = k|x_i) \log[q(z_i = k|x_i)] + \sum_{k=1}^{K} q(z_i = k|x_i) \log p(x_i, z_i = k; \theta) \\
=: F_i(q, \theta)
\]

We will define

\[
H(q(z_i|x_i)) := - \sum_{k=1}^{K} q(z_i = k|x_i) \log[q(z_i = k|x_i)]
\]

and

\[
L_c(x_i, z_i; \theta) := \log p(x_i, z_i; \theta)
\]

so that the above lower bound can be written as

\[
F_i(q, \theta) = H(q(z_i|x_i)) + \mathbb{E}_q(L_c(x_i, z_i; \theta))
\]

The first term \( H(q(z_i|x_i)) \) has an information-theoretic interpretation - it is the **entropy** of the distribution \( q(z_i|x_i) \), a non-negative quantity that measures the amount of disorder encoded in the distribution. As mentioned earlier, the term \( L_c(x_i, z_i; \theta) \) is called the **complete log likelihood** of \( x_i \) - it will be easier to optimize \( \mathbb{E}_q(L_c(x_i, z_i; \theta)) \) than the original log likelihood, since we will not need to deal with the “marginalization problem” in the original log likelihood from Equation 6.2.
Since the log likelihood of the full data is the sum of the individual log likelihoods, we can take
sums and find
\[
\ell(\theta; X) \geq \sum_{i=1}^{N} F_i(q, \theta) = H(q(z|X)) + \mathbb{E}_q(L_c(X, z; \theta)) =: F(q, \theta)
\]  
(6.3)

Here, \(X\) denotes the full dataset and \(z\) denotes the length \(N\) vector of latent variables. It is easy to
check that if \(q(z_i|x_i) = p(z_i|x_i; \theta)\) for all \(i\), then the inequality (6.3) is tight (set \(q(z_i|x_i) = p(x_i, z_i; \theta)\)
in the application of Jensen’s inequality and observe both sides of the inequality are equal). Thus it
makes sense to perform an alternating maximization scheme, where we iteratively update \(q(z_i|x_i)\)
to \(p(z_i|x_i; \theta)\) to make the inequality tight and then maximize over \(\theta\). Formally, the algorithm is as follows:

1. Initialize \(\theta^0\)
2. Expectation (E) step (soft imputation): set \(q^{t+1} = \arg\max_q F(q, \theta^t)\), that is,
   \[
   q^{t+1}(z_i = k|x_i) := p(z_i = k|x_i; \theta^t)
   \]
   This value of \(q\) is used to compute \(\mathbb{E}_q(L_c(X, z; \theta^t))\).
3. Maximization (M) step (parameter estimation): set
   \[
   \theta^{t+1} = \arg\max_{\theta} F(q^{t+1}, \theta) = \arg\max_{\theta} \mathbb{E}_{q^{t+1}}(L_c(X, z; \theta))
   \]
4. Repeat steps 2, 3 until convergence

A few remarks: when we maximize over \(q\) in the E step, there are \(nK\) values to be updated - one
value of \(q(z_i = k|x_i)\) for every data index \(i\) and latent index \(k\). After the E step, \(q^{t+1}\) is fixed and
does not depend on \(\theta\), so the entropy term does not depend on \(\theta\) and maximizing \(F(q^{t+1}, \theta)\) in
the subsequent M step amounts to maximizing the expected complete log likelihood. The E step is
what we have previously described as soft imputation of the latents: we fill in values for the hidden
variables \(z_i\) by determining a conditional distribution \(q(z_i|x_i)\). The M step assumes the E step
has done a reasonable job at imputing the data and uses this additional information to maximize
the likelihood. Observe the connections between K-means, soft K-means, and EM - all perform
alternating steps of data imputation and subsequent parameter optimization given the imputed
data. In the data imputation step for K-means, each data point is given a hard assignment to a
latent variable value, while in soft K-means and EM, each data point gets assigned a distribution
over the latent variables.

From our derivation of EM, we can see that the value of the likelihood never decreases during
the execution of the algorithm. It turns out that EM will converge to a parameter estimate with
zero gradient, but will not necessarily find the global optimum. When the clusters are sufficiently
separated, EM can exhibit Newtonian (second-order) convergence speed - however, if the clusters
are close together, the posteriors will be very flat and EM can take longer than gradient descent
methods to converge.

**EM for MoG**

As a concrete example, we derive the EM updates for fitting a mixture of Gaussians. Recall the
MoG model
\[
x|z \sim \mathcal{N}(\mu_z, \Sigma_z), \quad p(z = k) =: \alpha_k
\]
6.3. EXPECTATION MAXIMIZATION (EM) ALGORITHM

We define the parameter set $\theta$ as the set of all $\mu_k, \Sigma_k, \alpha_k$. Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be our observed data. Define $q^t_{ki} := q^t(z_i = k|x_i)$. The EM updates, derived below, are

**E step:**

$$q^{t+1}_k(z_i = k|x_i) = \frac{\alpha_k^t p(x_i|z_i = k; \theta^t)}{\sum_{j=1}^K \alpha_j^t p(x_i|z_i = j; \theta^t)}$$

**M step:**

$$\mu_k^{t+1} = \frac{\sum_{i=1}^N q^{t+1}_{ki} x_i}{\sum_{i=1}^N q^{t+1}_{ki}}$$

$$\Sigma_k^{t+1} = \frac{\sum_{i=1}^N q^{t+1}_{ki} (x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T}{\sum_{i=1}^N q^{t+1}_{ki}}$$

$$\alpha_k^{t+1} = \frac{1}{N} \sum_{i=1}^N q^{t+1}_{ki}$$

In the E step, we assign to each $x_i$ a probability distribution over latents (that is, a soft assignment). This assignment is $p(x_i|z_i = k; \theta^t)$, the Gaussian likelihood of the data, but reweighted by the prior and normalized. In the M step, we are essentially computing the usual maximum likelihood estimates of the parameters, but weighted by the posterior on $z$; indeed, if we set $q^{t+1}_{ki} = \frac{1}{N}$, then we are using the usual MLE. The update for $\mu_k$ is entirely analogous to the update to the centroids for soft k-means (6.1). The main difference is that now we are also updating estimates of covariances and the prior and synthesizing this information in our posterior estimates in the E step, which will in turn influence the $\mu_k$ assignments in the next E step.

We now derive these updates. Recall the log likelihood is

$$\ell(\theta; x) = \sum_{i=1}^N \log \left( \sum_{k=1}^K \pi_k p(x_i|z_i = k; \theta) p(z_i = k; \theta) \right)$$

For the E step, we set

$$q^{t+1}_k(z_i = k|x_i) = p(z_i = k|x_i; \theta^t)$$

$$= \frac{p(z_i = k; x_i; \theta^t)}{p(x_i; \theta^t)}$$

$$= \frac{p(x_i|z_i = k; \theta^t) p(z_i = k; \theta^t)}{\sum_{j=1}^K p(x_i, z_i = j; \theta^t)}$$

$$= \frac{p(x_i|z_i = k; \theta^t) p(z_i = k; \theta^t)}{\sum_{j=1}^K p(x_i|z_i = j; \theta^t) p(z_i = j; \theta^t)}$$

For MoG, $p(x|z = j; \theta^t)$ is the pdf of $N(\mu_j, \Sigma_j)$ evaluated at $x$.

For the M step, we need to maximize the expected complete log-likelihood $\ell_q = \mathbb{E}_{q^{t+1}}(L_c(X, z; \theta))$. The parameters to estimate are $\mu_k$, $\Sigma_k$, and $\alpha_k$, the prior. The expected complete log likelihood is

$$\mathbb{E}_{q^{t+1}}(L_c(x, z; \theta)) = \sum_{i=1}^N \sum_{k=1}^K q^{t+1}_{ki} \left[ \log \alpha_k - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) - \frac{1}{2} \log \left( (2\pi)^d |\Sigma_k| \right) \right]$$

$$= \sum_{i=1}^N \sum_{k=1}^K q^{t+1}_{ki} \left[ \log \alpha_k - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \frac{1}{2} \log \left( (2\pi)^{-d} |\Sigma_k^{-1}| \right) \right]$$
We can take gradients with respect to the parameters:

\[
\frac{\partial \ell_q}{\partial \mu_k} = \Sigma_k^{-1} \sum_{i=1}^N q_{ki}^{t+1} (x_i - \mu_k)
\]

\[
\frac{\partial \ell_q}{\partial \Sigma_k^{-1}} = \frac{1}{2} \sum_{i=1}^N q_{ki}^{t+1} [\Sigma_k - (x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T]
\]

Setting the gradients to zero and solving these equations gives us the updates for \( \mu_k, \Sigma_k \) shown above. To obtain the update for \( \alpha_k \), we need to introduce the constraint \( \sum_{k=1}^K \alpha_k = 1 \) via Lagrange multipliers. We thus maximize the Lagrangian \( \ell'_q = E_{q_{t+1}}[L_c(x, z; \theta)] - \lambda (\sum_{k=1}^K \alpha_k - 1) \). Taking gradients, we get

\[
\frac{\partial \ell'_q}{\partial \alpha_k} = -\lambda + \frac{1}{\alpha_k} \sum_{i=1}^N q_{ki}^{t+1} = 0
\]

We can rearrange to write

\[\alpha_k \lambda = \sum_{i=1}^N q_{ki}^{t+1}\]

Summing over all \( k \), we obtain

\[\lambda \sum_{k=1}^K \alpha_k = \sum_{k=1}^K \sum_{i=1}^N q_{ki}^{t+1} = \sum_{i=1}^N \sum_{k=1}^K q^{t+1} (z_i = k|x_i) = \sum_{i=1}^N 1 = N\]

using the fact that \( q^{t+1} \) is a distribution over \( z_i \). Since \( \sum_{k=1}^K \alpha_k = 1 \), the left hand side reduces to \( \lambda \), so we conclude \( \lambda = N \). Substituting into the original gradient, we have

\[
\frac{\partial \ell'_q}{\partial \alpha_k} = -N + \frac{1}{\alpha_k} \sum_{i=1}^N q_{ki}^{t+1}
\]

Setting this to zero gives us the desired updates for \( \alpha_k \).

**Appendix: Jensen’s Inequality Proof**

Let \( f \) be convex, and \( X \) be a random variable. Construct the tangent line \( L(X) = aX + b \) to \( f \) at \( E(X) \) for some \( a, b \) - this means \( L(E(X)) = f(E(X)) \). By convexity, \( L(X) \leq f(X) \) for all \( X \).\(^3\) Then by monotonicity of expectation, we have

\[f(E(X)) = L(E(X)) = aE(X) + b = E(aX + b) = E(L(X)) \leq E(f(X))\]

\(^3\)Actually, it takes some extra work to prove this intuitive fact. We will take it for granted here.
Appendix: Distance and Similarity

In our discussion of K-means, we restricted ourselves to using $L^2$ distance as a distance function. Formally, a distance function $d(x, y)$ is defined as a non-negative function that satisfies the following properties:

1. $d(x, y) = 0$ iff $x = y$
2. $d(x, y) = d(y, x)$ for all $x, y$
3. $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z$ (triangle inequality)

A dissimilarity measure $d(x, y)$ is a function satisfying the above properties except possibly the triangle inequality. One possible similarity measure $s(x, y)$ can be defined as $-d(x, y)$. In clustering, we are free to choose our notion of similarity. Different algorithms may or may not work differently depending on which similarity measure we choose.

For example, it does not make sense to use the K-means algorithm if we care about $L^1$ distances instead of $L^2$. However, we can apply the same principles used to derive the K-means algorithm: if we replace the $L^2$ norm in the objective by $L^1$ and do a similar alternating minimization, then on the centroid assignment step, we will set each centroid to the median of the data instead of the mean. On the cluster assignment step, we will assign each point to the closest centroid in $L^1$ distance. This variation is called K-medians.

Deciding which similarity measure to use is a modeling choice that typically depends on the data and desired clustering properties. For example, K-medians may be of use if the data has outliers and we desire a more robust estimator of the clusters. In certain domains, such as computer vision, the $L^p$ distances are not appropriate measures of dissimilarity, so other measures may be used.
Chapter 7

Decision Tree Learning

7.1 Decision Trees

A decision tree is a model that makes predictions by posing a series of simple tests on the given point. This process can be represented as a tree, with the non-leaf nodes of the tree representing these tests. At the leaves of this tree are the values to be predicted after descending the tree in the manner prescribed by each node’s test. Decision trees can be used for both classification and regression, but we will focus exclusively on classification.

In the simple case which we consider here, the tests are of the form “Is feature $j$ of this point less than the value $v$?”\footnote{One could use more complicated decision procedures. For example, in the case of a categorical variable, there could be a separate branch in the tree for each value that the variable could take on. Or multiple features could be used, e.g. “Is $x_1 > x_2$?” However, using more complicated decision procedures complicates the learning process. For simplicity we consider the single-feature binary case here.} These tests carve up the feature space in a nested rectangular fashion:
Given sufficiently many splits, a decision tree can represent an arbitrarily complex classifier and perfectly classify any training set.\(^2\)

**Training**

Decision trees are trained in a greedy, recursive fashion, downward from the root. After creating a node with some associated split, the children of that node are constructed by the same tree-growing procedure. However, the data used to train left subtree are only those points satisfying \(x_j < v\), and similarly the right subtree is grown with only those points with \(x_j \geq v\). Eventually we must each a base case where no further splits are to be made, and a prediction (rather than a split) is associated with the node.

We can see that the process of building the tree raises at least the following questions:

- How do we pick the split-feature, split-value pairs?
- How do we know when to stop growing the tree?

Typically decision trees will pick a split by considering all possible splits and choosing the one that is the best according to some criterion. We will discuss possible criteria later, but first it is worth asking what we mean by “all possible splits”. It is clear that we should look at all features, but what about the possible values? Observe that in the case where tests are of the form \(x_j < v\), there are infinitely many values of \(v\) we could choose, but only finitely many different resulting splits (since there are finitely many training points). Therefore it suffices to perform a one-dimensional sweep: sort the datapoints by their \(x_j\) values and only consider these values as split candidates.

Now let us revisit the criterion. Intuitively, we want to choose the split which most reduces our classifier’s uncertainty about which class points belongs to. In the ideal case, the hyperplane \(x_j = v\)

\(^2\) Unless two training points of different classes coincide.
perfectly splits the given data points such that all the instances of the positive class lie on one side and all the instances of the negative class lie on the other.

**Entropy and information**

One way to quantify the aforementioned “uncertainty”, we’ll use the ideas of surprise and entropy. The surprise of observing that a discrete random variable $Y$ takes on value $k$ is:

$$\log \frac{1}{P(Y = k)} = - \log P(Y = k)$$

As $P(Y = k) \to 0$, the surprise of observing that value approaches $\infty$, and conversely as $P(Y = k) \to 1$, the surprise of observing that value approaches 0.

The entropy of $Y$, denoted $H(Y)$, is the expected surprise:

$$H(Y) = \mathbb{E}[- \log P(Y)]$$

$$= - \sum_k P(Y = k) \log P(Y = k)$$

Here’s a graph of entropy vs. $p$ for a Bernoulli($p$) random variable:

![Graph of entropy vs. $p$ for a Bernoulli($p$) random variable](image)

Observe that as a function of $p$ (the probability of the variable being 1), the entropy is strictly concave. Moreover, it is maximized at $p = \frac{1}{2}$, when the probability distribution is uniform with respect to the outcomes. That is to say, a coin that is completely fair ($P(Y = 0) = P(Y = 1) = \frac{1}{2}$) has more entropy than a coin that is biased. This is because we are less sure of the outcome of the fair coin than the biased coin overall. Even though we are more surprised when the biased coin comes up as its more unlikely outcome, the way that entropy is defined gives a higher uncertainty score to the fair coin. Generally speaking, a random variable has more entropy when the distribution of its outcomes is closer to uniform and less entropy when the distribution is highly skewed to one outcome.

This definition is for random variables, but in practice we work with data. The distribution is empirically defined by our training points $\{(x_i, y_i)\}_{i=1}^n$. Concretely, the probability of class $k$ is the proportion of datapoints having class $k$:

$$P(Y = k) = \frac{\left| \{i \mid y_i = k\} \right|}{n}$$

We know that when we choose a split-feature, split-value pair, we want to reduce entropy in some way. Let $X_{j,v}$ be an indicator variable which is 1 when $x_j < v$, and 0 otherwise. There are a few entropies to consider:
• $H(Y)$
• $H(Y|X_{j,v} = 1)$, the entropy of the distribution of points such that $x_j < v$.
• $H(Y|X_{j,v} = 0)$, the entropy of the distribution of points such that $x_j \geq v$.

$H(Y)$ is not really under our control: we start with the set of points with labels represented by $Y$, this distribution has some entropy, and now we wish to carve up those points in a way to minimize the entropy remaining. Thus, the quantity we want to minimize is a weighted average of the two sides of the split, where the weights are (proportional to) the sizes of two sides:

$$\text{minimize } H(Y|X_{j,v}) := P(X_{j,v} = 1)H(Y|X_{j,v} = 1) + P(X_{j,v} = 0)H(Y|X_{j,v} = 0)$$

This quantity $H(Y|X_{j,v})$ is known as the conditional entropy of $Y$ given $X_{j,v}$. An equivalent way of seeing this is that we want to maximize the information we’ve learned, which is represented by how much entropy is reduced after learning whether or not $x_j < v$:

$$\text{maximize } I(X_{j,v};Y) := H(Y) - H(Y|X_{j,v})$$

This quantity $I(X_{j,v};Y)$ is known as the mutual information between $X_{j,v}$ and $Y$. It is always nonnegative, and it’s zero iff the resulting sides of the split have the same distribution of classes as the original set of points. Let’s say you were using a decision tree to classify emails as spam and ham. For example, you gain no information if you take a set of (20 ham, 10 spam) and split it on some feature to give you sets of (12 ham, 6 spam); (8 ham, 4 spam) because the empirical distribution of those two resulting sets is equal to the original one.

Gini impurity

Another way to assess the quality of a split is Gini impurity, which measures how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. It as defined as

$$G(Y) = \sum_k P(Y = k)\sum_{j \neq k} P(Y = j) = \sum_k P(Y = k)(1 - P(Y = k)) = 1 - \sum_k P(Y = k)^2$$

Exactly as with entropy, we can define a version of this quantity which is dependent on the split. For example, $G(Y|X_{j,v} = 1)$ would be the Gini impurity computed only on those points satisfying $x_j < v$. And we can define an analogous quantity

$$G(Y|X_{j,v}) := P(X_{j,v} = 1)G(Y|X_{j,v} = 1) + P(X_{j,v} = 0)G(Y|X_{j,v} = 0)$$

which is to be minimized.

Empirically, the Gini impurity is found to produce very similar results to entropy, and it is slightly faster to compute because we don’t need to take logs.

Why not the misclassification rate?

Since we ultimately care about classification accuracy, it is natural to wonder why we don’t directly use the misclassification rate by plurality vote as the measure of impurity:

$$M(Y) = 1 - \max_k P(Y = k)$$
It turns out that this quantity is insensitive in the sense that the quantity it induces for evaluating
splits \( M(Y|X_{j,v}) \) may assign the same value to a variety of splits which are not, in fact, equally
good. Suppose the current node has 40 training points of class 1 and 40 of class 2. Here \( M(Y) = 1 - \frac{1}{2} = \frac{1}{2} \). Now consider two possible splits:

1. Separate into a region \( x_j < v \) with 30 points of class 1 and 10 of class 2 \( \frac{10}{40} = \frac{1}{4} \) misclassified, and a region \( x_j \geq v \) with 10 points of class 1 and 30 of class 2 \( \frac{10}{40} = \frac{1}{4} \) misclassified. Since \( \frac{40}{80} = \frac{1}{2} \) of the points lie satisfy \( x_j < v \) and \( \frac{40}{80} = \frac{1}{2} \) of the points lie satisfy \( x_j \geq v \),

   \[
   M(Y|X_{j,v}) = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{4}
   \]

2. Separate into a region \( x_j < v \) with 20 points of class 1 and 40 of class 2 \( \frac{20}{60} = \frac{1}{3} \) misclassified, and a region \( x_j \geq v \) with 20 points of class 1 and 0 of class 2 \( \frac{0}{20} = 0 \) misclassified. Since \( \frac{60}{80} = \frac{3}{4} \) of the points lie satisfy \( x_j < v \) and \( \frac{20}{80} = \frac{1}{4} \) of the points lie satisfy \( x_j \geq v \),

   \[
   M(Y|X_{j,v}) = \frac{3}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot 0 = \frac{1}{4}
   \]

We see that the criterion value is the same for both splits, even though the second one appears
to do a better job reducing our uncertainty for many of the points. And indeed, the conditional
entropy and Gini impurity scores are lower (so the information gain is higher) for the second split.\(^4\)

The limitation of this criterion can be understood mathematically in terms of strict concavity.
This property means that the graph of the function lies strictly below the tangent line\(^5\) at every
point (except the point of tangency, of course). Consider the plots of entropy and misclassification
rate for a binary classification problem:

---

\(^3\) This example is from https://sebastianraschka.com/faq/docs/decision-tree-binary.html

\(^4\) Left as exercise (or see the source URL).

\(^5\) Or, in higher dimensions, hyperplane.
Both curves are concave, but the one for misclassification rate is not strictly so; it has only two
unique tangent lines. Because the conditional quantity is a convex combination (since it is weighted
by probabilities/proportions) of the children’s values, the strictly convex functions always yield
positive information gain as long as the children’s distributions are not identical to the parent’s.
We have no such guarantee in either linear region of the misclassification rate curve; any convex
combination of points on a line also lies on the line, yielding zero information gain.

Stopping criteria

We mentioned earlier that sufficiently deep decision trees can represent arbitrarily complex decision
boundaries, but of course this will lead to overfitting if we are not careful. There are a number of
heuristics we may consider to decide when to stop splitting:

- Limited depth: don’t split if the node is beyond some fixed depth depth in the tree
- Node purity: don’t split if the proportion of training points in some class is sufficiently high
- Information gain criteria: don’t split if the gained information/purity is sufficiently close to
  zero

Note that these are not mutually exclusive, and the thresholds can be tuned with validation. As
an alternative (or addition) to stopping early, you can **prune** a fully-grown tree by re-combining
splits if doing so reduces validation error.

7.2 Random Forests

Another way to combat overfitting is to combine the predictions of many varied models into a
single prediction, typically by plurality vote in the case of classification and averaging in the case
of regression. This is a general technique known as **ensemble learning**. To understand the
motivation for averaging, consider a set of uncorrelated random variables \( \{Y_i\}_{i=1}^n \) with common
mean \( E[Y_i] = \mu \) and variance \( \text{Var}(Y_i) = \sigma^2 \). The average of these has the same expectation

\[
E \left[ \frac{1}{n} \sum_{i=1}^{n} Y_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[Y_i] = \frac{1}{n} \cdot n \mu = \mu
\]

but reduced variance compared to each of the individual \( Y_i \)'s:

\[
\text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \left( \frac{1}{n} \right)^2 \sum_{i=1}^{n} \text{Var}(Y_i) = \frac{1}{n^2} \cdot n \sigma^2 = \frac{\sigma^2}{n}
\]

In the context of ensemble methods, these \( Y_i \) are analogous to the prediction made by classifier \( i \).
The combined prediction has the same expected value as any individual prediction but lower vari-
ance. Real-world predictions will of course not be completely uncorrelated, but reducing correlation
will generally reduce the final variance, so this is a goal to aim for.

**Random forests** are a specific ensemble method where the individual models are decision trees
trained in a randomized way so as to reduce correlation among them. Because the basic decision
tree building algorithm is deterministic, it will produce the same tree every time if we give it the
same dataset and use the same algorithm hyperparameters (stopping conditions, etc.).

Random forests are typically randomized in the following ways:
7.3. BOOSTING

- Per-classifier **bagging** (short for **bootstrap aggregating**): sample some number \( m < n \) of datapoints uniformly with replacement, and use these as the training set.

- Per-split **feature randomization**: sample some number number \( k < d \) of features as candidates to be considered for this split

Both the size of the random subsample of training points and the number of features at each split are hyperparameters which should be tuned through cross-validation.

### 7.3 Boosting

We have seen that in the case of random forests, combining many imperfect models can produce a single model that works very well. This is the idea of **ensemble methods**. However, random forests treat each member of the forest equally, taking a plurality vote or an average over their outputs. The idea of **boosting** is to combine the models (typically called **weak learners** in this context) in a more principled manner. The key idea is as follows: to improve our combined model, we should focus on finding learners that correctly predict the points which the overall boosted model is currently predicting inaccurately. Boosting algorithms implement this idea by associating a weight with each training point and iteratively reweighting so that mispredicted points have relatively high weights. Intuitively, some points are “harder” to predict than others, so the algorithm should focus its efforts on those.

These ideas also connect to matching pursuit. In both cases, our overall predictor is an additive combination of pieces which are selected one-by-one in a greedy fashion. The algorithm keeps track of residual prediction errors, chooses the “direction” to move based on these, and then performs a sort of line search to determine how far along that direction to move.

#### AdaBoost

There are many flavors of boosting. We will discuss one of the most popular versions, known as **AdaBoost** (short for **adaptive boosting**), which is a method for binary classification. Its developers won the prestigious Gödel Prize for this work.

### Algorithm

We present the algorithm first, then derive it later. Assume access to a dataset \( \{(x_i, y_i)\}_{i=1}^n \), where \( x_i \in \mathbb{R}^d \) and \( y_i \in \{-1, 1\} \).

1. Initialize the weights \( w_i = \frac{1}{n} \) for all \( i = 1, \ldots, n \) training points.

2. Repeat for \( m = 1, \ldots, M \):
   
   - (a) Build a classifier \( G_m : \mathbb{R}^d \rightarrow \{-1, 1\} \), where in the training process the data are weighted according to \( w_i \).
   
   - (b) Compute the weighted error \( e_m = \sum_{i \text{ misclassified}} w_i \sum_i w_i \).
   
   - (c) Re-weight the training points as

   \[
   w_i \leftarrow w_i \cdot \begin{cases} \sqrt{\frac{1-e_m}{e_m}} & \text{if misclassified by } G_m \\ \sqrt{\frac{e_m}{1-e_m}} & \text{otherwise} \end{cases}
   \]
(d) Optional: normalize the weights \( w_i \) to sum to 1.

We first address the issue of step (a): how do we train a classifier if we want to weight different samples differently? One common way to do this is to resample from the original training set every iteration to create a new training set that is fed to the next classifier. Specifically, we create a training set of size \( n \) by sampling \( n \) values from the original training data with replacement, according to the distribution \( w_i \). (This is why we might renormalize the weights in step (d).) This way, data points with large values of \( w_i \) are more likely to be included in this training set, and the next classifier will place higher priority on such data points.

Suppose\(^6\) that our weak learners always produce an error \( e_m < \frac{1}{2} \). To make sense of the formulas we see in the algorithm, note that for step (c), if the \( i \)-th data point is misclassified, then the weight \( w_i \) gets increased by a factor of \( \sqrt{\frac{1-e_m}{e_m}} \) (more priority placed on sample \( i \)), while if it is classified correctly, the priority gets decreased. AdaBoost does have a practical weakness in that this aggressive reweighting can cause the classifier to focus too much on certain training examples – if the data contains outliers or a lot of noise, the boosting algorithm’s generalization performance may suffer as it overfits to a few challenging examples.

We have not yet discussed how to make a prediction on test points given our classifiers \( G_1, \ldots, G_M \). One conceivable method is to use logistic regression with \( G_m(x) \) as features. However, a smarter choice that is based on the AdaBoost algorithm is to set

\[
\alpha_m = \frac{1}{2} \ln \left( \frac{1 - e_m}{e_m} \right)
\]

and classify \( x \) by

\[
h(x) = \text{sgn} \left( \sum_{m=1}^{M} \alpha_m G_m(x) \right)
\]

Note that this choice of \( \alpha_m \) (derived later) gives high weight to classifiers that have low error:

- As \( e_m \to 0 \), \( \frac{1-e_m}{e_m} \to \infty \), so \( \alpha_m \to \infty \).
- As \( e_m \to 1 \), \( \frac{1-e_m}{e_m} \to 0 \), so \( \alpha_m \to -\infty \).

We now proceed to demystify the formulas in the algorithm above by presenting a matching pursuit interpretation of AdaBoost. This interpretation is also useful because it generalizes to a powerful technique called Gradient Boosting, of which AdaBoost is just one instance.

**Derivation of AdaBoost**

Suppose we have computed classifiers \( G_1, \ldots, G_{m-1} \) along with their corresponding weights \( \alpha_k \) and we want to compute the next classifier \( G_m \) along with its weight \( \alpha_m \). The output of our model so far is \( F_{m-1}(x) := \sum_{i=1}^{m-1} \alpha_i G_i(x) \), and we want to minimize the risk:

\[
\alpha_m, G_m = \arg \min_{\alpha, G} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \alpha G(x_i))
\]

\(^6\) This is a reasonable thing to ask. A classifier with error \( e_m \geq \frac{1}{2} \) is even worse than the trivial classifier which predicts the class with the most total weight without regard for the input \( x_i \).
for some suitable loss function \( L(y, \hat{y}) \). Loss functions we have previously used include mean squared error for linear regression, cross-entropy loss for logistic regression, and hinge loss for SVM. For AdaBoost, we use the exponential loss:

\[
L(y, \hat{y}) = e^{-y\hat{y}}
\]

This loss function is illustrated in Figure 7.1. Observe that if \( y\hat{y} > 0 \) (i.e. \( \hat{y} \) has the correct sign), the loss decreases exponentially in \( |\hat{y}| \), which should be interpreted as the confidence of the prediction. Conversely, if \( y\hat{y} < 0 \), our loss is increasing exponentially in the confidence of the prediction.

Figure 7.1: The exponential loss provides exponentially increasing penalty for confident incorrect predictions. This figure is from Cornell CS4780 notes.

Plugging the exponential loss into the general optimization problem above yields

\[
\alpha_m, G_m = \arg \min_{\alpha, G} \sum_{i=1}^{n} e^{-y_i(F_m - 1(x_i)) + \alpha G(x_i)}
\]

\[
= \arg \min_{\alpha, G} \sum_{i=1}^{n} e^{-y_i F_m - 1(x_i)} e^{-y_i \alpha G(x_i)}
\]

The term \( w_i^{(m)} := e^{-y_i F_{m-1}(x_i)} \) is a constant with respect to our optimization variables. We can split out this sum into the components with correctly classified points and incorrectly classified points:

\[
\alpha_m, G_m = \arg \min_{\alpha, G} \sum_{i=1}^{n} w_i^{(m)} e^{-y_i \alpha G(x_i)}
\]

\[
= \arg \min_{\alpha, G} \sum_{y_i = G(x_i)} w_i^{(m)} e^{-\alpha} + \sum_{y_i \neq G(x_i)} w_i^{(m)} e^{\alpha}
\]

\[
= \arg \min_{\alpha, G} e^{-\alpha} \left( \sum_{i=1}^{n} w_i^{(m)} - \sum_{y_i \neq G(x_i)} w_i^{(m)} \right) + e^{\alpha} \sum_{y_i \neq G(x_i)} w_i^{(m)}
\]

\[
= \arg \min_{\alpha, G} \left( e^{\alpha} - e^{-\alpha} \right) \sum_{y_i \neq G(x_i)} w_i^{(m)} + e^{-\alpha} \sum_{i=1}^{n} w_i^{(m)}
\]
To arrive at (*) we have used the fact that \( y_iG_m(x_i) \) equals 1 if the prediction is correct, and \(-1\) otherwise. For a fixed value of \( \alpha \), the second term in this last expression does not depend on \( G \). Thus we can see that the best choice of \( G_m(x) \) is the classifier that minimizes the total weight of the misclassified points. Let

\[
e_m = \frac{\sum_{y_i \neq G_m(x_i)} w_i^{(m)}}{\sum_i w_i^{(m)}}
\]

Once we have obtained \( G_m \), we can solve for \( \alpha_m \). Dividing (*) by the constant \( \sum_{i=1}^n w_i^{(m)} \), we obtain

\[
\alpha_m = \arg \min_\alpha (1 - e_m) e^{-\alpha} + e_m e^\alpha
\]

We can solve for the minimizer analytically using calculus. Setting the derivative of the objective function to zero gives

\[
0 = -(1 - e_m)e^{-\alpha} + e_m e^\alpha = -e^{-\alpha} + e_m(e^{-\alpha} + e^\alpha)
\]

Multiplying through by \( e^\alpha \) yields

\[
0 = -1 + e_m(1 + e^{2\alpha})
\]

Adding one to both sides and dividing by \( e_m \), we have

\[
\frac{1}{e_m} = 1 + e^{2\alpha}
\]

i.e.

\[
e^{2\alpha} = \frac{1}{e_m} - 1 = \frac{1 - e_m}{e_m}
\]

Taking natural log on both sides and halving, we arrive at

\[
\alpha_m = \frac{1}{2} \ln \left( \frac{1 - e_m}{e_m} \right)
\]

as claimed earlier. From the optimal \( \alpha_m \), we can derive the weights:

\[
w_i^{(m+1)} = \exp \left( -y_i F_m(x_i) \right)
\]

\[
= \exp \left( -y_i [F_{m-1}(x_i) + \alpha_m G_m(x_i)] \right)
\]

\[
= w_i^{(m)} \exp \left( -y_i G_m(x_i) \alpha_m \right)
\]

\[
= w_i^{(m)} \exp \left( -y_i G_m(x_i) \frac{1}{2} \ln \left( \frac{1 - e_m}{e_m} \right) \right)
\]

\[
= w_i^{(m)} \exp \left( \ln \left( \frac{1 - e_m}{e_m} \right) - \frac{1}{2} y_i G_m(x_i) \right)
\]

\[
= w_i^{(m)} \left( \frac{1 - e_m}{e_m} \right)^{-\frac{1}{2}} y_i G_m(x_i)
\]

\[
= w_i^{(m)} \sqrt{\frac{e_m}{1 - e_m}} y_i G_m(x_i)
\]

Here we see that the multiplicative factor is \( \sqrt{\frac{e_m}{1 - e_m}} \) when \( y_i = G_m(x_i) \) and \( \sqrt{\frac{1 - e_m}{e_m}} \) otherwise. This completes the derivation of the algorithm.
As a final note about the intuition, we can view these α updates as pushing towards a solution in some direction until we can no longer improve our performance. More precisely, whenever we compute α_m (and thus w^{(m+1)}), for the incorrectly classified entries, we have

\[ \sum_{y_i \neq G_m(x_i)} w_i^{(m+1)} = \sum_{y_i \neq G_m(x_i)} w_i^{(m)} \sqrt{\frac{1 - e_m}{e_m}} \]

Dividing the right-hand side by \( \sum_{i=1}^{n} w_i^{(m)} \), we obtain \( e_m \sqrt{\frac{1 - e_m}{e_m}} = \sqrt{e_m(1 - e_m)} \). Similarly, for the correctly classified entries, we have

\[ \frac{\sum_{y_i = G_m(x_i)} w_i^{(m+1)}}{\sum_{i=1}^{n} w_i^{(m)}} = (1 - e_m) \sqrt{\frac{e_m}{1 - e_m}} = \sqrt{e_m(1 - e_m)} \]

Thus these two quantities are the same once we have adjusted our α, so the misclassified and correctly classified sets both get equal total weight.

This observation has an interesting practical implication. Even after the training error goes to zero, the test error may continue to decrease. This may be counter-intuitive, as one would expect the classifier to be overfitting to the training data at this point. One interpretation for this phenomenon is that even though the boosted classifier has achieved perfect training error, it is still refining its fit in a max-margin fashion, which increases its generalization capabilities.

**Gradient Boosting**

AdaBoost assumes a particular loss function, the exponential loss function. **Gradient boosting** is a more general technique that allows an arbitrary differentiable loss function \( L(y, \hat{y}) \). Recall the general optimization problem we must solve when choosing the next model:

\[
\min_{\alpha, G} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \alpha G(x_i))
\]

Here \( G \) should no longer be assumed to be a classifier; it may be real-valued if we are solving a regression problem. By a Taylor expansion in the second argument,

\[
L(y_i, F_{m-1}(x_i) + \alpha G(x_i)) \approx L(y_i, F_{m-1}(x_i)) + \frac{\partial L}{\partial \hat{y}}(y_i, F_{m-1}(x_i)) \cdot \alpha G(x_i)
\]

We can view the collection of predictions that a model \( G \) produces for the training set as a single vector \( g \in \mathbb{R}^{n} \) with components \( g_i = G(x_i) \). Then the overall cost function is approximated to first order by

\[
\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i)) + \alpha \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{y}}(y_i, F_{m-1}(x_i)) \cdot g_i \cdot \langle \nabla_{\hat{y}}L(y, F_{m-1}(x)), g \rangle
\]

where, in abuse of notation, \( F_{m-1}(X) \) is a vector with \( F_{m-1}(x_i) \) as its ith element, and \( \nabla_{\hat{y}}L(y, \hat{y}) \) is a vector with \( \frac{\partial L}{\partial \hat{y}}(y_i, \hat{y}_i) \) as its ith element. To decrease the cost in a steepest descent fashion, we seek the direction \( g \) which maximizes

\[
\left| \langle -\nabla_{\hat{y}}L(y, F_{m-1}(X)), g \rangle \right|
\]
subject to $g$ being the output of some model $G$ in the model class we are considering.\footnote{The absolute value may seem odd, but consider that after choosing the direction $g_m$, we perform a line search to select $\alpha_m$. This search may choose $\alpha_m < 0$, effectively flipping the direction. The key is to maximize the magnitude of the inner product.}

Some comments are in order. First, observe that the loss need only be differentiable with respect to its inputs, not necessarily with respect to model parameters, so we can use non-differentiable models such as decision trees. Additionally, in the case of squared loss $L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ we have

$$\frac{\partial L}{\partial \hat{y}}(y_i, F_{m-1}(x_i)) = -(y_i - F_{m-1}(x_i))$$

so

$$-\nabla_{\hat{y}} L(y, F_{m-1}(X)) = y - F_{m-1}(X)$$

This means the algorithm will follow the residual, as in matching pursuit.
Chapter 8

Deep Learning

8.1 Convolutional Neural Networks

Neural networks have been successfully applied to many real-world problems, such as predicting stock prices and learning robot dynamics in autonomous systems. The most general type of neural network is multilayer perceptrons (MLPs), which we have already studied in detail and employed to classify $28 \times 28$ pixel images as digits (MNIST). While MLPs can be used to effectively classify small images (such as those in MNIST), they are impractical for large images. Let’s see why. Given a $W \times H \times 3$ image (over 3 channels — red, green, blue), an MLP would take the flattened image as input, pass it through several fully connected (FC) layers and non-linearities, and finally output a vector of probabilities for each of the classes.

Associated with each FC layer is a $n_i \times n_o$ weight matrix that “connects” each of the $n_i$ input neurons to each of the $n_o$ output neurons, hence the term “fully connected layer”. The first FC layer takes an image as input, with $n_i = W \times H \times 3$ input neurons. Assuming that there are $n_o \approx n_i$ output neurons, then there are $n_i \times n_o \approx W^2 \times H^2 \times 3^2$ weights — a prohibitively large number of weights (in the millions)! This analysis extends to all FC layers that have large inputs, not just the first layer. In the framework of image classification, MLPs are generally ineffective — not only are they computationally expensive to train (both in terms of time and memory usage), but they
also have high variance due to the large number of weights.

**Convolutional neural networks (CNNs, or ConvNets)** are a different neural network architecture that significantly reduces the number of weights, and in turn reduces variance. Like MLPs, CNNs use FC layers and non-linearities, but they introduce two new types of layers — convolutional and pooling layers. Let’s look at these two layers in detail.

### Convolutional Layers

A **convolutional layer** takes a $W \times H \times D$ dimensional input $I$ and convolves it with a $w \times h \times D$ dimensional filter (or kernel) $G$. The weights of the filter can be hand designed, but in the context of machine learning we tune them automatically, just like we tune the weights of an FC layer. Mathematically, the convolution operator is defined as

$$(I * G)[x, y] = \sum_{a=0}^{w-1} \sum_{b=0}^{h-1} \sum_{c \in \{1, \ldots, D\}} I_c[x+a, y+b] \cdot G_c[a, b]$$

The subscript in $I_c$ indexes into the depth of the image, in this case for depth $c$. We can view convolution as either:

1. a 2-D operator over the width/height of the image, “broadcast” over the depth
2. a 3-D operator over the weight/height/depth of the image, with the convolution over the depth spanning the whole image with no room to move.

The output $L = I * G$ is an array of $(W - w + 1) \times (H - h + 1) \times 1$ values.

![Convolution Example](image)

Figure 8.2: Convolving a filter with an image. In this example, we have $W = H = 7, w = h = 3, D = 1$. We extract $n_o = (W - w + 1) \times (H - h + 1) \times 1 = 25$ output values from $n_i = W \times H \times D = 49$ input values, via a filter with $3 \times 3 = 9$ weights.

What exactly is convolution useful for, and why do we use it in the context of image classification? In simple terms, convolutions help us **extract features**. On a low level, filters can be used to detect all kinds of edges in an image, and at a high level they can detect more complex shapes and objects that are critical to classifying an image. Consider a simple horizontal edge detector filter $[1 \ -1]$. This filter will produce large negative values for inputs in which the left pixel is bright and the right pixel is dark; conversely, it will produce large positive values for inputs in which the left pixel is dark and the right pixel is bright.
8.1. CONVOLUTIONAL NEURAL NETWORKS

In general, a filter will produce large positive values in the areas of the image which appear most similar to it. As another example, here is a filter detects edges at a positive 45-degree angle:

\[
\begin{bmatrix}
0 & 0.6 & 0 \\
0.2 & 0 & 0.2 \\
0 & 0.2 & 0.6 \\
\end{bmatrix}
\]

How do conv layers compare to FC layers? Let’s revisit the example from figure 8.2, the input is \(n_i = 49\) units, while the output is \(n_o = 25\) units. In an FC layer, we would have used \(n_i \times n_o = 1225\) weights, but in our conv layer the filter only has 9 weights! conv layers use a significantly smaller number of weights, which reduces the variance of the model significantly while retaining the expressiveness. This is because we make use of **weight sharing**:

1. the same weights are shared among all the pixels of the input
2. the individual units of the output layer were all determined by the same weights

Compare this to the fully-connected architecture where for each output unit, there is a separate weight to learn for each input unit. We can illustrate the point for a simple 1-D input-output example.

![FC vs. conv layer](image)

Figure 8.4: FC vs. conv layer. Conv layers are equivalent to FC layers, except that (1) all weights outside the receptive field are 0, and (2) the weights are shared.

This kind of strategy decreases the complexity of our model (there are fewer weights), but it makes sense for image processing because there are repeated patterns in images, and if we have one filter
which can detect some kind of phenomenon, then it would make sense to use it elsewhere in the image.

A few more subtleties about convolutional layers:

1. We can stack the outputs of multiple filters together to form a convolutional layer.

![Figure 8.5: Here, we slid 6 independent 5 × 5 × 3 filters across the original image to produce 6 activation maps in the next convolutional layer. Each activation map can also be referred to as a channel.](image)

2. To save memory, we can have our filter stride across the image by multiple pixels.

3. Zero-padding is sometimes used to control the exact dimensions of the convolutional layer.

4. As we add more convolutional layers, the effective **receptive field** of each successive layer increases. That is to say, as we go downstream (of the layers), the value of any single unit is informed by an increasingly large patch of the original image. For example. If we use two successive layers of 3 × 3 filters, any one unit in the first convolutional layer is informed by 9 separate image pixels. Any one unit in the second convolutional layer is informed by 9 separate units of the first convolutional layer, which could informed by up to 9 × 9 = 81 original pixels.
5. We can think of convolutional layers as a complexity-regularized version of fully-connected layers.

Pooling Layers

A pooling layer does not involve more weights. Instead, it is a layer whose purpose is solely to downsample (also known as pool, gather, consolidate) the previous layer. It does this by sliding a window of some size across the units of a layer of the CNN and choosing (somehow) one value to effectively “represent” all the units captured by the window. We can tweak the nature of a pooling layer in a few orthogonal ways.

1. How to pool? In max-pooling, the representative value just becomes the largest of all the units in the window. In average-pooling, the representative value is the average of all the units in the window.

2. In which direction to pool?

(a) Spatial pooling pools values within the same channel. This has the capability of introducing translational invariance to our model.

(b) Cross-channel pooling pools values across different channels. This has the capability of introducing transformational invariance to our model.
CHAPTER 8. DEEP LEARNING

Figure 8.8: Here, we have an example where our convolutional layer is represented by 3 filters. Suppose they were each good for detecting the number 5 in some degree of rotation. If we pooled across the three channels determined by these filters, then no matter what orientation of the number “5” we got as input to our CNN, the pooling layer would have a large response!

3. “Lossiness” of pooling. This is determined by the stride of the pooling window. If we stride by a large amount, we potentially lose more information, but we conserve memory.

Backpropagation for CNNs

Just like MLPs, we can use the Backpropagation algorithm to train CNNs as well. We simply have to compute partial derivatives for conv and pool layers, just as we did for FC layers and non-linearities.

Derivatives for conv layers

Let’s denote the error function as $f$. In classification tasks, this is typically cross entropy loss. The forward pass will give us the input $I$ to the CNN. The backward pass will compute the partial derivatives of the error $f$ with respect to the output of the layer $L$, $\frac{\partial f}{\partial L}$. Without additional knowledge about the error function after $L$, we can compute the derivatives with respect to elements in the input $I$ and filter $G$ using the chain rule. Specifically, we have

$$\frac{\partial f}{\partial G_{c}[x, y]} = \frac{\partial f}{\partial L} \frac{\partial L}{\partial G_{c}[x, y]}$$

and

$$\frac{\partial f}{\partial I_{c}[x, y]} = \frac{\partial f}{\partial L} \frac{\partial L}{\partial I_{c}[x, y]}$$

where $G_{c}[x, y]$ denotes the entry in the filter for color $c$ at position $(x, y)$ and similarly for $I_{c}[x, y]$. From the equation for discrete convolution, we can compute the derivatives for each entry $(i, j)$ in $L$ as

$$\frac{\partial L[i, j]}{\partial G_{c}[x, y]} = \frac{\partial}{\partial G_{c}[x, y]} \sum_{a=1}^{w} \sum_{b=1}^{h} \sum_{c \in \{r, g, b\}} I_{c}[i + a, j + b] \cdot G_{c}[a, b]$$

$$= I_{c}[i + x, j + y]$$

For the input image, we similarly compute the derivative as

$$\frac{\partial L[i, j]}{\partial I_{c}[x, y]} = \frac{\partial}{\partial I_{c}[x, y]} \sum_{a=1}^{w} \sum_{b=1}^{h} \sum_{c \in \{r, g, b\}} I_{c}[i + a, j + b] \cdot G_{c}[a, b]$$
8.2. CNN ARCHITECTURES

\[ = G_c[x - i, y - j] \]

where we have \( i + a = x \) and \( j + b = y \). When \( x - i \) or \( y - j \) go outside the boundary of the filter, we can treat the derivative as zero.

We can collect the derivatives of the filter parameter for all \( L[i, j] \) into a vector and multiply it by the derivatives we computed to get

\[
\frac{\partial f}{\partial L} \frac{\partial L}{\partial G_c[x, y]} = \sum_{i,j} \frac{\partial f}{\partial L[i,j]} I_c[i + x, j + y] \tag{8.1}
\]

\[
= \frac{\partial f}{\partial G_c[x, y]} \tag{8.2}
\]

and for the image

\[
\frac{\partial f}{\partial L} \frac{\partial L}{\partial I_c[x, y]} = \sum_{i,j} \frac{\partial f}{\partial L[i,j]} G_c[x - i, y - j]
\]

\[
= \frac{\partial f}{\partial I_c[x, y]}
\]

Derivatives for pool layers

8.2 CNN Architectures

Convolutional Neural Networks were first applied successfully to the ImageNet challenge in 2012 and continue to outperform computer vision techniques that do not use neural networks. Here are a few of the architectures that have been developed over the years.

LeNet (LeCun et al, 1998)

Key characteristics:

- Used to classify handwritten alphanumeric characters

AlexNet (Krizhevsky et al, 2012)

Key characteristics:

- Conv filters of varying sizes - for example, the first layer has \( 11 \times 11 \) conv filters
• First use of ReLU, which fixed the problem of saturating gradients in the predominant tanh activation.

• Several layers of convolution, max pooling, some normalization. Three fully connected layers at the end of the network (these comprise the majority of the weights in the network).

• Around 60 million weights, over half of which are in the first fully connected layer following the last convolution.

• Trained over two GPU’s — the top and bottom divisions in Figure 8.9 were due to the need to separate training onto two GPU’s. There was limited communication between the GPU’s, as illustrated by the arrows that go between the top and bottom.

• Dropout in first two FC layers — prevents overfitting

• Heavy data augmentation. One form is image translation and reflection: for example, an elephant facing the left is the same class as an elephant facing the right. The second form is altering the intensity of RGB color channels: different cameras can have different lighting on the same objects, so it is necessary to account for this.

VGGNet (Simonyan and Zisserman, 2014)

Reference paper: “Very Deep Convolutional Networks for Large-Scale Image Recognition,” ICLR 2015.\(^1\) Key characteristics:

• Only uses \(3 \times 3\) convolutional filters. Blocks of conv-conv-conv-pool layers are stacked together, followed by fully connected layers at the end (the number of convolutional layers between pooling layers can vary). Note that a stack of \(3 \times 3\) conv filters has the same effective receptive field as one \(7 \times 7\) conv filter. To see this, imagine sliding a \(3 \times 3\) filter over a \(7 \times 7\) image - the result is a \(5 \times 5\) image. Do this twice more and the result is a \(1 \times 1\) cell - sliding one \(7 \times 7\) filter over the original image would also result in a \(1 \times 1\) cell. The computational cost of the \(3 \times 3\) filters is lower - a stack of \(3\) such filters over \(C\) channels requires \(3 \times (3^2C)\) weights (not including bias weights), while one \(7 \times 7\) filter would incur a higher cost of \(7^2C\) learned weights. Deeper, more narrow networks can introduce more non-linearities than shallower, wider networks due to the repeated composition of activation functions.

\(^1\)VGG stands for the “Visual Geometry Group” at Oxford where this was developed.
8.2. CNN ARCHITECTURES

GoogLeNet (Szegedy et al, 2014)

Also codenamed as “Inception.”\(^2\) Published in CVPR 2015 as “Going Deeper with Convolutions.”

Key characteristics:

- Deeper than previous networks (22 layers), but more computationally efficient (5 million parameters - no fully connected layers).

- Network is composed of stacked sub-networks called “Inception modules.” The naive Inception module (a) runs convolutional layers in parallel and concatenates the filters together. However, this can be computationally inefficient. The dimensionality reduction Inception module (b) performs 1 × 1 convolutions that act as dimensionality reduction. This lowers the computational cost and makes it tractable to stack many Inception modules together.

ResNet (He et al, 2015)

\[ \mathcal{F}(x) \]

```
x
weight layer

\[ \text{relu} \]

weight layer

\[ \mathcal{F}(x) + x \]

\[ \text{relu} \]

identity
```

Figure 8.11: Building block for the ResNet from “Deep Residual Learning for Image Recognition,” CVPR 2016. If the desired function to be learned is \( \mathcal{H}(x) \), we instead learn the residual \( \mathcal{F}(x) := \mathcal{H}(x) - x \), so the output of the network is \( \mathcal{F}(x) + x = \mathcal{H}(x) \).

Key characteristics:

\(^2\)“In this paper, we will focus on an efficient deep neural network architecture for computer vision, codenamed Inception, which derives its name from the Network in network paper by Lin et al [12] in conjunction with the famous ‘we need to go deeper internet meme [1].’” The authors seem to be meme-friendly.
• Very deep (152 layers). Residual blocks (Figure 8.11) are stacked together - each individual weight layer in the residual block is implemented as a $3 \times 3$ convolution. There are no FC layers until the final layer.

• Residual blocks solve the “vanishing gradient” problem: the gradient signal diminishes in layers that are farther away from the end of the network. Let $L$ be the loss, $Y$ be the output at a layer, $x$ be the input. Regular neural networks have gradients that look like

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial Y} \frac{\partial Y}{\partial x}$$

but the derivative of $Y$ with respect to $x$ can be small. If we use a residual block where $Y = F(x) + x$, we have

$$\frac{\partial Y}{\partial x} = \frac{\partial F(x)}{\partial x} + 1$$

The $+x$ term in the residual block always provides some default gradient signal so the signal is still backpropagated to the front of the network. This allows the network to be very deep.

To conclude this section, we note that the winning ImageNet architectures have all increased in depth over the years. While both shallow and deep neural networks are known to be universal function approximators, there is growing empirical and theoretical evidence that deep neural networks can require fewer (even exponentially fewer) parameters than shallow nets to achieve the same approximation performance. There is also evidence that deep neural networks possess better generalization capabilities than their shallow counterparts. The performance, generalization, and optimization benefits of adding more layers is an ongoing component of theoretical research.

### 8.3 Visualizing and Understanding CNNs

We know that a convolutional net learns features, but these may not be directly useful to visualize. There are several methods available that enable us to better understand what convolutional nets actually learn. These include:

• Visualizing filters - can give an idea of what types of features the network learns, such as edge detectors. This only works in the first layer. Visualizing activations - can see sparsity in the responses as the depth increases. One can also visualize the feature map before a fully connected layer by conducting a nearest neighbor search in feature space. This helps to determine if the features learned by the CNN are useful - for example, in pixel space, an elephant on the left side of the image would not be a neighbor of an elephant on the right side of the image, but in a translation-invariant feature space these pictures might be neighbors.

• Reconstruction by deconvolution - isolate an activation and reconstruct the original image based on that activation alone to determine its effect.

• Activation maximization - Hubel and Wiesel’s experiment, but computationally

• Saliency maps - find what locations in the image make a neuron fire

• Code inversion - given a feature representation, determine the original image

• Semantic interpretation - interpret the activations semantically (for example, is the CNN determining whether or not an object is shiny when it is trying to classify?)