1 Ordinary Least Squares

Ordinary least squares (OLS) is one of the simplest regression problems, but it is well-understood and practically useful. In this setup, we assume that we have access to a dataset \( D = \{(x_i, y_i)\}_{i=1}^n \), where each \( x_i \in \mathbb{R}^d \) is a vector of numerical values, usually called features or predictors, and each \( y_i \in \mathbb{R} \) is a scalar that we want to predict using those features. Let us assume \( n \geq d \), which is typically the case in practice.

We want to fit a linear model, i.e., one such that \( y_i \approx x_i^\top w \) for some vector of weights \( w \in \mathbb{R}^d \).

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, \( X \in \mathbb{R}^{n \times d} \) has the datapoint \( x_i \) as its \( i \)th row. We typically assume that \( X \) is full rank, but later we will see options for when this is not the case.

There will in general be no exact solution to the equation \( y = Xw \) (consider how many equations and variables there are). But if we choose some metric for the quality of the model, we can attempt to find the “best-fit” model. In OLS, the objective is to minimize the sum (or equivalently, the mean) of the squared errors:

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

Now that we have formulated an optimization problem, we want to go about solving it. It turns out that the particular structure of OLS allows us to compute a closed-form solution for \( w_{\text{OLS}}^* \).

1.1 Approach 1: Vector calculus

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

\[
f(w) = \| Xw - y \|_2^2
\]

\[
= (Xw - y)^\top (Xw - y)
\]

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

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 \( f \) is easily seen to be
\[
\nabla f(w) = \nabla_w (w^\top X^\top Xw - 2w^\top X^\top y + y^\top y)
\]
\[
= \nabla_w (w^\top X^\top Xw) - 2\nabla_w (w^\top X^\top y) + \nabla_w (y^\top y)
\]
\[
= 2X^\top Xw - 2X^\top y
\]

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 and solving for \( w \), we have that the optimum \( w_{\text{OLS}}^* \) satisfies

\[
w_{\text{OLS}}^* = (X^\top X)^{-1}X^\top y
\]

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

\[
X^\top Xw = X^\top y
\]

In this derivation we have used the condition \( \nabla f(w^*) = 0 \), 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 \( f \) is convex, it suffices to compute the Hessian of \( f \), which in this case is

\[
\nabla^2 f(w) = X^\top X
\]

and show that this is positive semi-definite:

\[
\forall w \neq 0, \ w^\top X^\top Xw = (Xw)^\top Xw = \|Xw\|_2^2 \geq 0
\]

1.2 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 $S$ is the linear operator that maps $v$ to $v_S$ in the decomposition above. An important property of the orthogonal projection is that

$$\|v - P_Sv\| \leq \|v - s\|$$

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

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

**Proof.** By the Pythagorean theorem,

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

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

Here is a visual representation of the argument above:

![Diagram of orthogonal projection](image)

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

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

Let us write $P_{\text{range}(X)}$ as $P$ for brevity. Since we assume that $X$ is full rank with $n \geq d$, the columns of $X$ are linearly independent, so there is a one-to-one correspondence between $w$ and $Xw$. It follows that $Py = Xw^*$, where $w^*$ is the optimum for the right-hand side. To solve for this $w^*$, we need the following fact\(^1\):

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

\(^1\) This result is often stated as part of the Fundamental Theorem of Linear Algebra.
Since we are projecting onto \( \text{range}(X) \), the orthogonality condition for optimality is that \( y - Py \perp \text{range}(X) \), i.e. \( y - Xw^* \in \text{null}(X^\top) \). This leads to the equation

\[
X^\top(y - Xw^*) = 0
\]

which is equivalent to

\[
X^\top X w^*_{\text{OLS}} = X^\top y
\]
as before.

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 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, \( \text{rank}(X^\top X) = \text{rank}(X) = d \). 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 \( X \):

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

\[
= (V \begin{bmatrix} \Sigma_d & 0 \end{bmatrix} I \begin{bmatrix} \Sigma_d & 0 \end{bmatrix} V^\top)^{-1}
\]

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

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:

$$f(w) = \|Xw - y\|_2^2 + \lambda \|w\|_2^2 = w^\top X^\top Xw - 2w^\top X^\top y + y^\top y + \lambda w^\top w$$

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

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

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

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

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

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

$$\nabla^2 f(w) = X^\top X + \lambda I$$

and show that this is **positive definite (PD)**:

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

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**, (otherwise known as Tikhonov regularization in the statistics community). 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:

$$(X^\top X + \lambda I)^{-1} = \left( \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \right)^{-1} \left( \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \right)^{-1}$$
\[
\begin{align*}
&= \left( V \begin{bmatrix} \Sigma^2_r & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} V^T + \lambda I \right)^{-1} \\
&= \left( V \begin{bmatrix} \Sigma^2_r & 0 \\ 0 & 0 \end{bmatrix} V^T + V(\lambda I)V^T \right)^{-1} \\
&= \left( V \left( \begin{bmatrix} \Sigma^2_r & 0 \\ 0 & 0 \end{bmatrix} + \lambda I \right) V^T \right)^{-1} \\
&= \left( V \begin{bmatrix} \Sigma^2_r + \lambda I & 0 \\ 0 & \lambda I \end{bmatrix} V^T \right)^{-1} \\
&= (V^T)^{-1} \begin{bmatrix} \Sigma^2_r + \lambda I & 0 \\ 0 & \lambda I \end{bmatrix}^{-1} V^{-1} \\
&= V \begin{bmatrix} (\Sigma^2_r + \lambda I)^{-1} & 0 \\ 0 & \frac{1}{\lambda} I \end{bmatrix} V^T
\end{align*}
\]

Now with our slight tweak, the matrix $X^T X + \lambda I$ 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 penalized the weights corresponding to complex features that only serve to fine tune the model and fit noise in the data.