The least squares problem:
\[
\begin{align*}
\min_{\beta} \quad L(\beta) &= ||y - X\beta||_2^2, \\
\beta &= \arg \min_{\beta} \quad L(\beta)
\end{align*}
\]
\(X \in \mathbb{R}^{n \times p}\) data matrix
\(y \in \mathbb{R}^{n \times 1}\) target values
\(n\): number of sample points
\(p\): dimension of feature vectors.

Solve by setting the gradient to zero:
\[
\nabla_{\beta} L(\beta) = -2X^Ty + 2X^T\beta = 0
\]
\[
\Rightarrow \quad X^T\beta = X^Ty, \quad \text{called the "normal equation."}
\]

If \(X^TX\) is invertible, \(\hat{\beta} = (X^TX)^{-1}X^Ty\) is the unique solution.

Q: Under what condition is \((X^TX)\) invertible, or equivalent, of full rank?

Note: The rank of a square matrix is the max # of linearly independent rows (or columns).

A: Two cases: \(\oplus n < p \quad \ominus n \geq p\).

\[
\begin{pmatrix}
p \\
\end{pmatrix}
\begin{pmatrix}
X^TX \\
\end{pmatrix} =
\begin{cases}
p \begin{pmatrix}
X^T \\
\end{pmatrix} \begin{pmatrix}
X \\
\end{pmatrix}, & \text{if } n < p \\
\begin{pmatrix}
X^T \\
\end{pmatrix} \begin{pmatrix}
X \\
\end{pmatrix}, & \text{if } n \geq p
\end{cases}
\]

\(\text{rank}(X^TX) \leq n\), because every column of \(X^TX\) is a linear combination of at most \(n\) \(p\)-dimensional vectors.
When $n < p$, rank$(x^T x) < p$, so $x^T x$ not invertible, and the least square problem has multiple solutions.

When $n \geq p$, and there are $p$ linearly independent feature vectors in the data, (which is usually the case when $n \geq p$), $x^T x$ is invertible and 

$$\hat{\beta} = (x^T x)^{-1} x^T y$$

is the unique solution.
Ridge Regression

\[
\min_{\beta} \text{L}(\beta) = \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \\
= y^T y - 2y^T X \beta + \beta^T (X^T X + \lambda I) \beta
\]

\(\lambda > 0\) is regularization parameter,
\(I\) is a \(p\)-by-\(p\) identity matrix

Solve \(\nabla_{\beta} \text{L}(\beta) = 0\)

\[
\iff -2X^T y + 2(X^T X + \lambda I) \beta = 0
\]

\[
\iff (X^T X + \lambda I) \beta = X^T y .
\]

Thm: \(X^T X + \lambda I\) is always invertible

pf: Prove the following lemma first:

**Lemma:** \(\forall \ a \in \mathbb{R}^p, \ a \neq 0\),
\[ a^T (X^T X + \lambda I) a > 0 . \]

**Pr:**
\[
\begin{align*}
a^T (X^T X + \lambda I) a &= a^T X^T X a + \lambda a^T a \\
&= \|Xa\|_2^2 + \lambda a^T a > 0 . \text{ Since } a \neq 0 \text{ and } \lambda > 0
\end{align*}
\]

Then prove by contradiction: If \(X^T X + \lambda I\) is not invertible, its columns are not linearly independent, so there exists \(a \in \mathbb{R}^p, \ a \neq 0\) such that
\[
(X^T X + \lambda I) a = 0,
\]
which implies \(a^T (X^T X + \lambda I) a = 0\), a contradiction to the lemma.
So
\[ \hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'y \]

is the unique solution to the Ridge Regression problem.

Why ridge regression?

1. When \( n < p \), helps to get a unique solution.

2. When \( n \geq p \), even though \( \hat{\beta} \) usually exists and is unique, it may overfit the data. In terms of Bias and Variance,

\[
\text{bias}(\hat{\beta}_{\text{ridge}}) \geq \text{bias}(\hat{\beta}) = 0 \text{ under the linear model,}
\]
\[
\text{Variance}(\hat{\beta}_{\text{ridge}}) < \text{Variance}(\hat{\beta})
\]

As \( \lambda \uparrow \), \( \text{bias}(\hat{\beta}_{\text{ridge}}) \uparrow \) and \( \text{Variance}(\hat{\beta}_{\text{ridge}}) \downarrow \)

Use cross validation to decide \( \lambda \).
Consider the following family of p.d.f.s over the interval \([a, b]\):

\[
f(x) = \sum_{j=1}^{K} \frac{1}{\Delta_j} \mathbb{1} \{ x \in B_{ij} \} \, p_j, \quad p_j > 0\]

is the density in the \(j\)th bin.

Let \(\Delta_1, \Delta_2, \ldots, \Delta_K\) be the sizes of the \(K\) bins, so \(\sum_{j=1}^{K} \Delta_j = b - a\) and

\[
\Pr(x \in B_{ij}) = \int_{a}^{b} \mathbb{1} \{ x \in B_{ij} \} \, f(x) \, dx = p_j \Delta_j.
\]

Since \(f(x)\) is a p.d.f., we have

\[
\int_{a}^{b} f(x) \, dx = \sum_{j=1}^{K} p_j \Delta_j = 1
\]

Given an \(i.i.d\) sample \(\{x_1, x_2, \ldots, x_n\}\) drawn from some \(f\) in this family, we want to estimate the densities \(p_1, p_2, \ldots, p_K\). We do MLE estimation.

**Likelihood:**

\[
\mathcal{L}(p_1, \ldots, p_K) = \prod_{i=1}^{n} \left( p_j \Delta_j \right)^{\mathbb{1} \{ x_i \in B_{ij} \}}
\]

**Log likelihood:**

\[
\log \mathcal{L}(p_1, \ldots, p_K) = \sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{1} \{ x_i \in B_{ij} \} \log (p_j \Delta_j)
\]

\[
= \sum_{j=1}^{K} \frac{1}{\Delta_j} \sum_{i=1}^{n} \mathbb{1} \{ x_i \in B_{ij} \} \log (p_j \Delta_j)
\]

Call \(\hat{n}_j\), \# of points in \(B_{ij}\).

Concave in \(p_1, \ldots, p_K\)

Solve \(\max \mathcal{L}(p_1, \ldots, p_K)\) s.t. \(\sum_{j=1}^{K} p_j \Delta_j = 1\) by setting the gradient of the Lagrangian function to zero:

\[
\nabla p_j \left[ \mathcal{L}(p_1, \ldots, p_K) - \lambda \left( \sum_{j=1}^{K} p_j \Delta_j - 1 \right) \right] = 0 \quad \Rightarrow \quad \frac{\Delta \hat{n}_j}{\hat{p}_j \Delta_j} - \lambda \Delta_j = 0
\]

\[
\hat{p}_j = \frac{\hat{n}_j}{\lambda \Delta_j}, \quad \text{since} \quad \sum_{j=1}^{K} \hat{p}_j \Delta_j = 1, \quad \lambda \text{ must be } \sum_{j=1}^{K} \hat{n}_j = n, \quad \text{and}
\]

\[
\hat{p}_j = \frac{\hat{n}_j}{n \Delta_j}, \quad \text{the histogram density estimate}.
\]
\textit{L_0 penalty is non-convex.}

For simplicity, consider one dimensional case.

\textbf{Def.} A function \( f \) is convex if
\[
\alpha f(x_1) + (1-\alpha) f(x_2) \geq f(\alpha x_1 + (1-\alpha) x_2) \quad \forall 0 \leq \alpha \leq 1 \quad \text{and} \quad \forall x_1, x_2 \text{ in domain of } f.
\]

\begin{itemize}
  \item \textbf{Ex.} convex
  \begin{itemize}
    \item \text{Graph 1:} \( f(x) \) is convex if the line connecting any two points on the graph lies above the graph.
    \item \text{Graph 2:} \( f(x) \) is non-convex if the line connecting any two points on the graph is below the graph.
  \end{itemize}
\end{itemize}

The \( L_0 \) penalty in 1-d:
\[
L_0(\beta) = \text{1 if } \beta \neq 0 \text{ if } \beta = 0
\]

\begin{itemize}
  \item \text{Graph 1:} \( f(\alpha x_1 + (1-\alpha) x_2) = 1 \)
  \item \text{Graph 2:} \( \alpha f(x_1) + (1-\alpha) f(x_2) = 1 - \alpha < 1 \quad \Rightarrow \text{non-convex.} \)
\end{itemize}