Next lecture - online
HW extended till next Monday.

Data: \( D = \{ (X_i, Y_i) \}_{i=1}^N \)

Regression.

\[
P(Y | X, \Theta, \sigma^2) = \mathcal{N}(Y; \Theta^T X, \sigma^2)
\]

\[
x = (x_1, x_2, \ldots, x_d, 1) \in \mathbb{R}^d
\]

\[
y \in \text{scalar (R)}
\]

\[
f(x) = \Theta^T x + \Theta_0 + \ldots + \Theta_d x_d + \Theta_d
\]

If you pre-normalized \( y \Rightarrow y := y - \bar{y}_y \Rightarrow \sigma_d^2 = 0
\]

In order to assess the “importance” of different dimensions based on their weights, you may want to pre-normalize \( x_i \Rightarrow x_i := \frac{x_i - \bar{x}_i}{\sigma_i}
\]

ML formulation: \( \hat{\Theta}_{ML} = \arg \max_{\Theta} \ln(\Theta, \sigma^2; D)
\]

\[
\ln(\Theta, \sigma^2; D) = (X^T X)^{-1} (X^T Y)
\]

\[
X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} \in \mathbb{R}^{N \times d} 
\]

\[
Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^{N \times 1}
\]

\[
\frac{1}{N} X^T X = \frac{1}{N} \sum_{i=1}^N (x_i x_i^T) \in \mathbb{R}^{d \times d} 
\]

- Sample estimate of \( E(XX^T) \)
- Covariance sample of \( X \)

What is \( \frac{1}{N} X^T Y \)?

- Sample cross-covariance of \( X \) and \( Y \)
Also for $\sigma^2$

$$\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\Theta}_{ML}^T x_i)^2$$

This is in fact biased, just like our density estimation problem.

$$\hat{\sigma}_{\text{unbiased}}^2 = \frac{1}{N-d} \sum_{i=1}^{N} (y_i - \hat{\Theta}_{ML}^T x_i)^2$$

* Moving away from the "linear" assumption

$$f(x', x) = \alpha(x)^2 + b$$

$$= \Theta_1(x')^2 + \Theta_2(x')^2 + \Theta_3(x')^2 + \Theta_4(x') + \Theta_5(x') + \Theta_6$$

$$= [\Theta_1, \Theta_2, \ldots, \Theta_6] \begin{bmatrix} (x')^2 \\ x' \end{bmatrix} = \Theta^T \phi(x)$$

Linear regression on $\phi(x)$

Now extend to Bayesian formulation of linear regression

We'll stick with conjugate priors $\Rightarrow$ posterior will also be of the same form

$$P(\Theta | b^2) = \mathcal{N} (\Theta; \hat{\Theta}, \Sigma) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^\frac{1}{2}} \exp \left( -\frac{1}{2} (\Theta - \hat{\Theta})^T \Sigma^{-1} (\Theta - \hat{\Theta}) \right)$$

$$\Rightarrow \quad P(\Theta | b^2, \text{Data}) = \mathcal{N} (\Theta; (\Sigma^{-1} + X^T X)^{-1}(X^T Y + \Sigma^{-1} \hat{\Theta}), \sigma^2 (\Sigma^{-1} + X^T X)^{-1})$$

"Posterior"

$$\propto P(Y | X, \Theta, b^2) P(\Theta | b^2)$$

likelihood $\propto$ prior.
\( \hat{\Theta}_{\text{MAP}} = \arg \max \ P(\Theta \mid b^2, \text{data}) \)

\( \hat{\Theta}_{\text{Bayes}} = \mathbb{E}(\Theta \mid b^2, \text{data}) \)

In this case, they are the same:

\( \hat{\Theta}_{\text{MAP}} = \hat{\Theta}_{\text{Bayes}} = (\Sigma^{-1} + XX^T)(X^T \gamma + \Sigma^{-1} \vec{\Theta}) \)

Compare with

\( \hat{\Theta}_{M\ell} = (XX^T)^{-1}X^T \gamma \)

Most commonly use \( \Sigma \) and \( \vec{\Theta} \)

\( \Sigma = \frac{1}{\lambda} I \quad \vec{\Theta} = 0 \quad \text{"Ridge Regression"} \quad \lambda > 0 \)

\( \Rightarrow \hat{\Theta}_{\text{ridge}} = (\lambda I + XX^T)^{-1}X^T \gamma \)

To understand \( \lambda \), transform \( X \) based on its Singular Value Decomposition:

\[ X = UDV \]

\( U \) is an orthogonal matrix \( R^{n \times n} \)

\( D \) is a diagonal matrix \( R^{n \times n} \)

\( V \) is an orthogonal matrix \( R^{n \times d} \)

\( U^T U = I \)

- rotation & reflection

\( X^T X = (UDV)^T (UDV) \)

\( = V^T D^T U^T U V D V \)

\( = V^T D^T D V \)

\( = V^T \text{diag}(d_1^2) V \)

\( V X^T X V^T = \text{diag}(d_1^2) \)

\( (XX^T)^T V^T = \text{diag}(d_1^2) \)

\( V X^T = \left( \begin{array} {c} x_1^T \\ \vdots \\ x_n^T \end{array} \right) \)

\( V^T = (Vx_1 \ldots Vx_n)^T \)
\( \text{diag}(d_i^2) \) is the sample variance of \( X_1, \ldots, X_n \) after transforming the coordinate system by \( V \).  

- In the new coordinate system, all coordinates are UNCORRELATED with each other.

- If \( x' = x^2 \)  
  \[ d_i^2 = 0 \]  
  \[ \Rightarrow X^T X \text{ will be non-invertible} \]

Dimension with larger \( d_i^2 \) is more important.

Come back to \( \hat{\theta}_{\text{Ridge}} \)

\[
\hat{f}(x) = (X^T \hat{\theta}_{\text{Ridge}}) = X^T \hat{\theta}_{\text{Ridge}} = UDV \hat{\theta}_{\text{Ridge}}
\]

\[
= UDV \left( \lambda I + X^T X \right)^{-1} X^T Y
\]

\[
= UDV \left( \lambda I + V^T \text{diag}(d_i^2) V \right)^{-1} V^T U^T Y
\]

\[
= UDV \left[ \lambda I + \text{diag}(d_i^2) \right] D^T U^T Y
\]

- Column vectors

\[
U = (U_1, U_2, \ldots, U_N)
\]

\[
\lambda \text{ shrinks the coordinates}
\]

\[
\lambda \text{ shrinks small coordinates with small variance}.
\]

\[
X \hat{\theta}_{\text{ML}} = \sum_{i=1}^{d} U_i U_i^T Y
\]