Linear regression

COMS 4771
1. Old Faithful and prediction functions
Prediction problem: Old Faithful geyser (Yellowstone)

**Task:** Predict the time of next eruption.
Statistical model for time between eruptions

Historical records of eruptions:

\[ a_0 \quad b_0 \quad a_1 \quad b_1 \quad a_2 \quad b_2 \quad a_3 \quad b_3 \quad \ldots \]
Statistical model for time between eruptions

Historical records of eruptions:

\[
\begin{align*}
  a_0 & b_0 & a_1 & b_1 & a_2 & b_2 & a_3 & b_3 & \ldots \\
  \quad Y_1 \quad & \quad Y_2 \quad & \quad Y_3 \quad & \quad \Rightarrow
\end{align*}
\]

“Galton board” iid model: \(Y_1, \ldots, Y_n, Y \sim_{iid} N(\mu, \sigma^2)\).

- **Data:** \(Y_i := a_i - b_{i-1}, i = 1, \ldots, n\).

(Admittedly not a great model, since durations are non-negative. Better: take \(Y_i\) from exponential distribution.)
Historical records of eruptions:

\[ \ldots a_{n-1} \ b_{n-1} \ a_n \ b_n \ \ldots \ \underline{t} \]

\[ \ldots \underline{\text{data}} \ldots \underline{Y} \]

“Galton board” iid model: \( Y_1, \ldots, Y_n, Y \sim \text{iid } \mathcal{N}(\mu, \sigma^2) \).

- **Data:** \( Y_i := a_i - b_{i-1}, \ i = 1, \ldots, n \).

(Admittedly not a great model, since durations are non-negative. Better: take \( Y_i \) from exponential distribution.)

**Task:**
At later time \( t \) (when an eruption ends), predict time of next eruption \( t + Y \).
Statistical model for time between eruptions

Historical records of eruptions:

\[ \ldots a_{n-1} \ b_{n-1} \ a_n \ b_n \ldots \ t \]

\[ \ldots \begin{array}{l} \vdots \\ \text{data} \end{array} \]

“Galton board” iid model: \( Y_1, \ldots, Y_n, Y \sim_{\text{iid}} N(\mu, \sigma^2) \).

- **Data:** \( Y_i := a_i - b_{i-1}, \ i = 1, \ldots, n. \)

(Admittedly not a great model, since durations are non-negative.
Better: take \( Y_i \) from exponential distribution.)

**Task:**
At later time \( t \) (when an eruption ends), predict time of next eruption \( t + Y \).

**On “Old Faithful” data:**
Statistical model for time between eruptions

Historical records of eruptions:

\[ \ldots a_{n-1} \quad b_{n-1} \quad a_n \quad b_n \quad \ldots \quad t \]

\[ \begin{array}{c}
\text{data} \\
\hline
\end{array} \]

"Galton board" iid model: \( Y_1, \ldots, Y_n, Y \sim_{iid} N(\mu, \sigma^2) \).

\[ \text{Data: } Y_i := a_i - b_{i-1}, \quad i = 1, \ldots, n. \]

(Admittedly not a great model, since durations are non-negative. Better: take \( Y_i \) from exponential distribution.)

Task:
At later time \( t \) (when an eruption ends), predict time of next eruption \( t + Y \).

On "Old Faithful" data:

\[ \text{Using 136 past observations, we form estimate } \hat{\mu} = 70.7941. \]
Statistical model for time between eruptions

Historical records of eruptions:

\[ \ldots a_{n-1} \ b_{n-1} \ a_n \ b_n \ldots \ t \]

\[ \begin{array}{ccccccc}
\text{data} & - & - & - & - & - & - & Y \\
\end{array} \]

“Galton board” iid model: \( Y_1, \ldots, Y_n, Y \sim \text{iid } \mathcal{N}(\mu, \sigma^2) \).

- **Data:** \( Y_i := a_i - b_{i-1}, \ i = 1, \ldots, n. \)

(Admittedly not a great model, since durations are non-negative. Better: take \( Y_i \) from exponential distribution.)

**Task:**
At later time \( t \) (when an eruption ends), predict time of next eruption \( t + Y \).

**On “Old Faithful” data:**

- Using 136 past observations, we form estimate \( \hat{\mu} = 70.7941 \).
- Mean squared loss of \( \hat{\mu} \) on next 136 observations is 187.1894.
Statistical model for time between eruptions

Historical records of eruptions:

\[ \ldots a_{n-1} \quad b_{n-1} \quad a_n \quad b_n \quad \ldots \quad t \]

\[ \text{data} \]

“Galton board” iid model: \( Y_1, \ldots, Y_n, Y \sim_{iid} \mathcal{N}(\mu, \sigma^2) \).

- **Data:** \( Y_i := a_i - b_{i-1}, \ i = 1, \ldots, n. \)

(Admittedly not a great model, since durations are non-negative. Better: take \( Y_i \) from exponential distribution.)

**Task:**
At later time \( t \) (when an eruption ends), predict time of next eruption \( t + Y \).

**On “Old Faithful” data:**

- Using 136 past observations, we form estimate \( \hat{\mu} = 70.7941 \).
- Mean squared loss of \( \hat{\mu} \) on next 136 observations is 187.1894.
  (Easier to interpret the square root, 13.6817, which has same units as \( Y \).)
Naturalist Harry Woodward observed that time until the next eruption seems to be related to duration of last eruption.
Naturalist Harry Woodward observed that the time until the next eruption seems to be related to the duration of the last eruption.
Using side-information

At prediction time $t$, duration of last eruption is available as *side-information*.

\[
\begin{array}{ccccccc}
\cdots & a_{n-1} & b_{n-1} & a_n & b_n & \cdots & t \\
\hline
& & & & & & \\
\text{data} & & & & & & \\
\end{array}
\]
Using side-information

At prediction time $t$, duration of last eruption is available as *side-information*.

$$\ldots a_{n-1} \ b_{n-1} \ a_n \ b_n \ \ldots \ \ t$$

$$\ldots \ X_n \ Y_n \ \ldots$$

**IID model for supervised learning:**

$(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are iid random *pairs* (i.e., *labeled examples*).

$X$ takes values in $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}$), $Y$ takes values in $\mathbb{R}$. 
Using side-information

At prediction time $t$, duration of last eruption is available as side-information.

\[
\cdots a_{n-1} \ b_{n-1} \ a_n \ b_n \ \cdots \ \ t
\]
\[
\cdots \ X_n \ Y_n \ \cdots \ \ X \ Y
\]

**IID model** for *supervised learning*: 
$(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are iid random *pairs* (i.e., *labeled examples*).

$X$ takes values in $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}$), $Y$ takes values in $\mathbb{R}$.

1. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the choose a *prediction function* (a.k.a. *predictor*)

\[
\hat{f} : \mathcal{X} \rightarrow \mathbb{R},
\]

This is called "learning" or "training".
Using side-information

At prediction time $t$, duration of last eruption is available as *side-information*.

$$\cdots \ a_{n-1} \ b_{n-1} \ a_n \ b_n \ \cdots \ t$$

$$\cdots \ X_n \leftarrow Y_n \rightarrow \cdots \ X \leftarrow Y \rightarrow$$

**IID model** for *supervised learning*:

$(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are iid random *pairs* (i.e., *labeled examples*).

$X$ takes values in $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}$), $Y$ takes values in $\mathbb{R}$.

1. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the choose a *prediction function* (a.k.a. *predictor*)

$$\hat{f}: \mathcal{X} \rightarrow \mathbb{R},$$

This is called "learning" or "training".

2. At prediction time, observe $X$, and form prediction $\hat{f}(X)$. 
Using side-information

At prediction time $t$, duration of last eruption is available as *side-information*.

$\ldots a_{n-1} \ b_{n-1} \ a_n \ b_n \ \ldots \ t$

$IID$ model for *supervised learning*: $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are iid random *pairs* (i.e., *labeled examples*). $X$ takes values in $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}$), $Y$ takes values in $\mathbb{R}$.

1. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the choose a *prediction function* (a.k.a. *predictor*)

   $\hat{f} : \mathcal{X} \rightarrow \mathbb{R}$

   This is called "learning" or "training".

2. At prediction time, observe $X$, and form prediction $\hat{f}(X)$.

3. Outcome is $Y$, and *squared loss* is $(\hat{f}(X) - Y)^2$.
At prediction time $t$, duration of last eruption is available as side-information.

IID model for supervised learning:
$(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are iid random pairs (i.e., labeled examples).

$X$ takes values in $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}$), $Y$ takes values in $\mathbb{R}$.

1. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$, and choose a prediction function (a.k.a. predictor)
\[
\hat{f} : \mathcal{X} \rightarrow \mathbb{R},
\]
This is called “learning” or “training”.

2. At prediction time, observe $X$, and form prediction $\hat{f}(X)$.

3. Outcome is $Y$, and squared loss is $(\hat{f}(X) - Y)^2$.

How should we choose $\hat{f}$ based on data?
2. Optimal predictors and linear regression models
Distributions over labeled examples

\( \mathcal{X} \): Space of possible side-information (feature space).
\( \mathcal{Y} \): Space of possible outcomes (label space or output space).
Distributions over labeled examples

\( \mathcal{X} \): Space of possible side-information (feature space).
\( \mathcal{Y} \): Space of possible outcomes (label space or output space).

Distribution \( P \) of random pair \((X, Y)\) taking values in \( \mathcal{X} \times \mathcal{Y} \) can be thought of in two parts:
$\mathcal{X}$: Space of possible side-information (feature space).

$\mathcal{Y}$: Space of possible outcomes (label space or output space).

Distribution $P$ of random pair $(X, Y)$ taking values in $\mathcal{X} \times \mathcal{Y}$ can be thought of in two parts:

1. **Marginal distribution** $P_X$ of $X$:
   
   $P_X$ is a probability distribution on $\mathcal{X}$. 

2. **Conditional distribution** $P_{Y|X=x}$ of $Y$ given $X=x$ for each $x \in \mathcal{X}$:
   
   $P_{Y|X=x}$ is a probability distribution on $\mathcal{Y}$. 

This lecture: $Y=R$ (regression problems).
Distributions over labeled examples

\( \mathcal{X} \): Space of possible side-information (feature space).
\( \mathcal{Y} \): Space of possible outcomes (label space or output space).

Distribution \( P \) of random pair \((X, Y)\) taking values in \( \mathcal{X} \times \mathcal{Y} \) can be thought of in two parts:

1. **Marginal distribution** \( P_X \) of \( X \):
   \( P_X \) is a probability distribution on \( \mathcal{X} \).

2. **Conditional distribution** \( P_{Y|X=x} \) of \( Y \) given \( X = x \) for each \( x \in \mathcal{X} \):
   \( P_{Y|X=x} \) is a probability distribution on \( \mathcal{Y} \).
$\mathcal{X}$: Space of possible side-information (feature space).

$\mathcal{Y}$: Space of possible outcomes (label space or output space).

Distribution $P$ of random pair $(X, Y)$ taking values in $\mathcal{X} \times \mathcal{Y}$ can be thought of in two parts:

1. **Marginal distribution** $P_X$ of $X$:
   
   $P_X$ is a probability distribution on $\mathcal{X}$.

2. **Conditional distribution** $P_{Y|X=x}$ of $Y$ given $X = x$ for each $x \in \mathcal{X}$:
   
   $P_{Y|X=x}$ is a probability distribution on $\mathcal{Y}$.

**This lecture**: $\mathcal{Y} = \mathbb{R}$ (regression problems).
Optimal predictor

What function $f : \mathcal{X} \to \mathbb{R}$ has smallest (squared loss) risk

$$
\mathcal{R}(f) := \mathbb{E}[(f(X) - Y)^2]
$$

Conditional on $X = x$, the minimizer of conditional risk $\hat{y} \mapsto \mathbb{E}[(\hat{y} - Y)^2 | X = x]$ is the conditional mean $\mathbb{E}[Y | X = x]$.

Therefore, the function $f^\star : \mathbb{R} \to \mathbb{R}$ where $f^\star(x) = \mathbb{E}[Y | X = x]$, $x \in \mathbb{R}$ has the smallest risk.

$f^\star$ is called the regression function or conditional mean function.
What function $f : \mathcal{X} \to \mathbb{R}$ has smallest (squared loss) risk

$$\mathcal{R}(f) := \mathbb{E}[(f(X) - Y)^2]?$$

- Conditional on $X = x$, the minimizer of conditional risk

$$\hat{y} \mapsto \mathbb{E}[(\hat{y} - Y)^2 | X = x]$$

is the conditional mean

$$\mathbb{E}[Y | X = x].$$

(Recall Galton board!)
What function \( f : \mathcal{X} \to \mathbb{R} \) has smallest (squared loss) \textit{risk}

\[
\mathcal{R}(f) := \mathbb{E}[(f(X) - Y)^2]
\]

- Conditional on \( X = x \), the minimizer of \textit{conditional risk}

\[
\hat{y} \mapsto \mathbb{E}[(\hat{y} - Y)^2 \mid X = x]
\]

is the conditional mean

\[
\mathbb{E}[Y \mid X = x].
\]

(Recall Galton board!)

- Therefore, the function \( f^* : \mathbb{R} \to \mathbb{R} \) where

\[
f^*(x) = \mathbb{E}[Y \mid X = x], \quad x \in \mathbb{R}
\]

has the smallest risk.
What function \( f : \mathcal{X} \rightarrow \mathbb{R} \) has smallest (squared loss) \( \text{risk} \)

\[
\mathcal{R}(f) := \mathbb{E}[(f(X) - Y)^2]?
\]

- Conditional on \( X = x \), the minimizer of \textit{conditional risk}

\[
\hat{y} \mapsto \mathbb{E}[(\hat{y} - Y)^2 | X = x]
\]

is the conditional mean

\[
\mathbb{E}[Y | X = x].
\]

(Recall Galton board!)

- Therefore, the function \( f^* : \mathbb{R} \rightarrow \mathbb{R} \) where

\[
f^*(x) = \mathbb{E}[Y | X = x], \quad x \in \mathbb{R}
\]

has the smallest risk.

- \( f^* \) is called the \textit{regression function} or \textit{conditional mean function}. 

- Optimal predictor
Linear regression models

When side-information is encoded as vectors of real numbers \( \mathbf{x} = (x_1, \ldots, x_d) \) (called \textit{features} or \textit{variables}), it is common to use a \textit{linear regression model}, such as the following:

\[
Y \mid \mathbf{X} = \mathbf{x} \sim N(\mathbf{x}^T \beta, \sigma^2), \quad \mathbf{x} \in \mathbb{R}^d.
\]
Linear regression models

When side-information is encoded as vectors of real numbers $\mathbf{x} = (x_1, \ldots, x_d)$ (called \textit{features} or \textit{variables}), it is common to use a \textit{linear regression model}, such as the following:

$$Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\mathbf{x}^\top \mathbf{\beta}, \sigma^2), \quad \mathbf{x} \in \mathbb{R}^d.$$ 

- Parameters: $\mathbf{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$, $\sigma^2 > 0$. 
Linear regression models

When side-information is encoded as vectors of real numbers \( \mathbf{x} = (x_1, \ldots, x_d) \) (called *features* or *variables*), it is common to use a *linear regression model*, such as the following:

\[
Y \mid \mathbf{X} = \mathbf{x} \sim N(\mathbf{x}^\top \beta, \sigma^2), \quad \mathbf{x} \in \mathbb{R}^d.
\]

▶ Parameters: \( \beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d, \sigma^2 > 0. \)

▶ \( \mathbf{X} = (X_1, \ldots, X_d) \), a *random vector* (i.e., a vector of random variables).
Linear regression models

When side-information is encoded as vectors of real numbers \( x = (x_1, \ldots, x_d) \) (called \textit{features} or \textit{variables}), it is common to use a \textit{linear regression model}, such as the following:

\[
Y \mid X = x \sim N(x^\top \beta, \sigma^2), \quad x \in \mathbb{R}^d.
\]

- Parameters: \( \beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d, \sigma^2 > 0. \)
- \( X = (X_1, \ldots, X_d) \), a \textit{random vector} (i.e., a vector of random variables).
- Conditional distribution of \( Y \) given \( X \) is normal.
Linear regression models

When side-information is encoded as vectors of real numbers \( x = (x_1, \ldots, x_d) \) (called features or variables), it is common to use a linear regression model, such as the following:

\[
Y \mid X = x \sim N(x^T \beta, \sigma^2), \quad x \in \mathbb{R}^d.
\]

- Parameters: \( \beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d, \sigma^2 > 0. \)
- \( X = (X_1, \ldots, X_d) \), a random vector (i.e., a vector of random variables).
- Conditional distribution of \( Y \) given \( X \) is normal.
- Marginal distribution of \( X \) not specified.
Linear regression models

When side-information is encoded as vectors of real numbers \( \mathbf{x} = (x_1, \ldots, x_d) \) (called features or variables), it is common to use a linear regression model, such as the following:

\[
Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\mathbf{x}^T \beta, \sigma^2), \quad \mathbf{x} \in \mathbb{R}^d.
\]

- Parameters: \( \beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d, \sigma^2 > 0. \)
- \( \mathbf{X} = (X_1, \ldots, X_d) \), a random vector (i.e., a vector of random variables).
- Conditional distribution of \( Y \) given \( \mathbf{X} \) is normal.
- Marginal distribution of \( \mathbf{X} \) not specified.

In this model, the regression function \( f^* \) is a linear function \( f_\beta : \mathbb{R}^d \to \mathbb{R}, \)

\[
f_\beta(\mathbf{x}) = \mathbf{x}^T \beta = \sum_{i=1}^{d} x_i \beta_i, \quad \mathbf{x} \in \mathbb{R}^d.
\]

(We’ll often refer to \( f_\beta \) just by \( \beta \).)
Linear functions might sound rather restricted, but actually they can be quite powerful if you are creative about side-information.

Examples:
1. Non-linear transformations of existing variables: for $x \in \mathbb{R}$, $\phi(x) = \ln(1 + x)$.
2. Logical formula of binary variables: for $x = (x_1, \ldots, x_d) \in \{0, 1\}^d$, $\phi(x) = (x_1 \land x_5 \land \neg x_{10}) \lor (\neg x_2 \land x_7)$.
3. Trigonometric expansion: for $x \in \mathbb{R}$, $\phi(x) = (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots)$.
4. Polynomial expansion: for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, $\phi(x) = (1, x_1, \ldots, x_d, x_1^2, \ldots, x_d^{d-1}, x_1 x_2, \ldots, x_1 x_d)$.
Enhancing linear regression models

Linear functions might sound rather restricted, but actually they can be quite powerful if you are creative about side-information.

Examples:

1. Non-linear transformations of existing variables: for $x \in \mathbb{R}$,
   \[ \phi(x) = \ln(1 + x). \]
Enhancing linear regression models

Linear functions might sound rather restricted, but actually they can be quite powerful if you are creative about side-information.

Examples:

1. Non-linear transformations of existing variables: for $x \in \mathbb{R}$,

   $$\phi(x) = \ln(1 + x).$$

2. Logical formula of binary variables: for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \{0, 1\}^d$,

   $$\phi(\boldsymbol{x}) = (x_1 \wedge x_5 \wedge \neg x_{10}) \lor (\neg x_2 \wedge x_7).$$
Enhancing linear regression models

Linear functions might sound rather restricted, but actually they can be quite powerful if you are creative about side-information.

Examples:

1. Non-linear transformations of existing variables: for \( x \in \mathbb{R} \),
   \[
   \phi(x) = \ln(1 + x).
   \]

2. Logical formula of binary variables: for \( \mathbf{x} = (x_1, \ldots, x_d) \in \{0, 1\}^d \),
   \[
   \phi(\mathbf{x}) = (x_1 \land x_5 \land \neg x_{10}) \lor (\neg x_2 \land x_7).
   \]

3. Trigonometric expansion: for \( x \in \mathbb{R} \),
   \[
   \phi(x) = (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots).
   \]
Enhancing linear regression models

Linear functions might sound rather restricted, but actually they can be quite powerful if you are creative about side-information.

Examples:

1. Non-linear transformations of existing variables: for \( x \in \mathbb{R} \),
   \[
   \phi(x) = \ln(1 + x).
   \]

2. Logical formula of binary variables: for \( \mathbf{x} = (x_1, \ldots, x_d) \in \{0, 1\}^d \),
   \[
   \phi(\mathbf{x}) = (x_1 \land x_5 \land \neg x_{10}) \lor (\neg x_2 \land x_7).
   \]

3. Trigonometric expansion: for \( x \in \mathbb{R} \),
   \[
   \phi(x) = (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots).
   \]

4. Polynomial expansion: for \( \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \),
   \[
   \phi(\mathbf{x}) = (1, x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, x_1 x_2, \ldots, x_1 x_d, \ldots, x_{d-1} x_d).
   \]
Example: Taking advantage of linearity

Suppose you are trying to predict some health outcome.
Example: Taking advantage of linearity

Suppose you are trying to predict some health outcome.

- Physician suggests that body temperature is relevant, specifically the (square) deviation from normal body temperature:

\[
\phi(x) = (x_{\text{temp}} - 98.6)^2.
\]
Example: Taking advantage of linearity

Suppose you are trying to predict some health outcome.

- Physician suggests that body temperature is relevant, specifically the (square) deviation from normal body temperature:

\[ \phi(x) = (x_{\text{temp}} - 98.6)^2. \]

- What if you didn’t know about this magic constant 98.6?
Example: Taking advantage of linearity

Suppose you are trying to predict some health outcome.

- Physician suggests that body temperature is relevant, specifically the (square) deviation from normal body temperature:

\[ \phi(x) = (x_{\text{temp}} - 98.6)^2. \]

- What if you didn’t know about this magic constant 98.6?
- Instead, use

\[ \phi(x) = (1, x_{\text{temp}}, x_{\text{temp}}^2). \]

Can learn coefficients \( \beta \) such that

\[ \beta^T \phi(x) = (x_{\text{temp}} - 98.6)^2, \]

or any other quadratic polynomial in \( x_{\text{temp}} \) (which may be better!).
Quadratic expansion

Quadratic function $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = ax^2 + bx + c, \quad x \in \mathbb{R},$$

for $a, b, c \in \mathbb{R}$. 
Quadratic expansion

Quadratic function $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = ax^2 + bx + c, \quad x \in \mathbb{R},$$

for $a, b, c \in \mathbb{R}$.

This can be written as a linear function of $\phi(x)$, where

$$\phi(x) := (1, x, x^2),$$

since

$$f(x) = \beta^T \phi(x)$$

where $\beta = (c, b, a)$. 
Quadratic function $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = ax^2 + bx + c, \quad x \in \mathbb{R},$$

for $a, b, c \in \mathbb{R}$.

This can be written as a linear function of $\phi(x)$, where

$$\phi(x) := (1, x, x^2),$$

since

$$f(x) = \beta^T \phi(x)$$

where $\beta = (c, b, a)$.

For multivariate quadratic function $f : \mathbb{R}^d \to \mathbb{R}$, use

$$\phi(x) := (1, x_1, \ldots, x_d, \underbrace{x_1^2, \ldots, x_d^2}_{\text{squared terms}}, \underbrace{x_1 x_2, \ldots, x_1 x_d, \ldots, x_{d-1} x_d}_{\text{cross terms}}).$$
Woodward needed an *affine expansion* for “Old Faithful” data:

\[ \phi(x) := (1, x). \]
Affine expansion and “Old Faithful”

Woodward needed an affine expansion for “Old Faithful” data:

\[ \phi(x) := (1, x). \]

Affine function \( f_{a,b} : \mathbb{R} \to \mathbb{R} \) for \( a, b \in \mathbb{R} \),

\[ f_{a,b}(x) = a + bx, \]

is a linear function \( f_\beta \) of \( \phi(x) \) for \( \beta = (a, b) \).

(This easily generalizes to multivariate affine functions.)
Why linear regression models?

1. Linear regression models benefit from good choice of features.
2. Structure of linear functions is very well-understood.
3. Many well-understood and efficient algorithms for learning linear functions from data, even when $n$ and $d$ are large.
1. Linear regression models benefit from good choice of features.
Why linear regression models?

1. Linear regression models benefit from good choice of features.
2. Structure of linear functions is very well-understood.
Why linear regression models?

1. Linear regression models benefit from good choice of features.
2. Structure of linear functions is very well-understood.
3. Many well-understood and efficient algorithms for learning linear functions from data, even when $n$ and $d$ are large.
3. From data to prediction functions
**Linear regression model with Gaussian noise:**

\((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, with

\[ Y \mid X = x \sim N(x^T \beta, \sigma^2), \quad x \in \mathbb{R}^d. \]

(Traditional to study linear regression in context of this model.)
Linear regression model with Gaussian noise:

\((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, with

\[ Y \mid X = x \sim N(x^T \beta, \sigma^2), \quad x \in \mathbb{R}^d. \]

(Traditional to study linear regression in context of this model.)

Log-likelihood of \((\beta, \sigma^2)\), given data \((X_i, Y_i) = (x_i, y_i)\) for \(i = 1, \ldots, n\):

\[
\sum_{i=1}^{n} \left\{ -\frac{1}{2\sigma^2} (x_i^T \beta - y_i)^2 + \frac{1}{2} \ln \frac{1}{2\pi\sigma^2} \right\} + \left\{ \text{terms not involving } (\beta, \sigma^2) \right\}.
\]
Linear regression model with Gaussian noise:

\((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, with

\[ Y \mid X = x \sim N(x^T \beta, \sigma^2), \quad x \in \mathbb{R}^d. \]

(Traditional to study linear regression in context of this model.)

Log-likelihood of \((\beta, \sigma^2)\), given data \((X_i, Y_i) = (x_i, y_i)\) for \(i = 1, \ldots, n\):

\[
\sum_{i=1}^{n} \left\{ -\frac{1}{2\sigma^2} (x_i^T \beta - y_i)^2 + \frac{1}{2} \ln \frac{1}{2\pi \sigma^2} \right\} + \left\{ \text{terms not involving } (\beta, \sigma^2) \right\}.
\]

The \(\beta\) that maximizes log-likelihood is also \(\beta\) that minimizes

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2.
\]
Linear regression model with Gaussian noise:

\((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, with

\[
Y \mid X = x \sim N(x^T \beta, \sigma^2), \quad x \in \mathbb{R}^d.
\]

(Traditional to study linear regression in context of this model.)

Log-likelihood of \((\beta, \sigma^2)\), given data \((X_i, Y_i) = (x_i, y_i)\) for \(i = 1, \ldots, n\):

\[
\sum_{i=1}^{n} \left\{ -\frac{1}{2\sigma^2} (x_i^T \beta - y_i)^2 + \frac{1}{2} \ln \frac{1}{2\pi\sigma^2} \right\} + \left\{ \text{terms not involving } (\beta, \sigma^2) \right\}.
\]

The \(\beta\) that maximizes log-likelihood is also \(\beta\) that minimizes

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2.
\]

This coincides with another approach, called empirical risk minimization, which is studied beyond the context of the linear regression model . . .
**Empirical distribution** $P_n$ on $(x_1, y_1), \ldots, (x_n, y_n)$ has probability mass function $p_n$ given by

$$p_n((x, y)) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \{(x, y) = (x_i, y_i)\}, \quad (x, y) \in \mathbb{R}^d \times \mathbb{R}.$$
**Empirical distribution** $P_n$ on $(x_1, y_1), \ldots, (x_n, y_n)$ has probability mass function $p_n$ given by

$$p_n((x, y)) := \frac{1}{n} \sum_{i=1}^{n} 1\{ (x, y) = (x_i, y_i) \}, \quad (x, y) \in \mathbb{R}^d \times \mathbb{R}.$$ 

**Plug-in principle:** Goal is to find function $f$ that minimizes (squared loss) risk

$$\mathcal{R}(f) = \mathbb{E}[(f(X) - Y)^2].$$

But we don’t know the distribution $P$ of $(X, Y)$. 

Empirical distribution $P_n$ on $(x_1, y_1), \ldots, (x_n, y_n)$ has probability mass function $p_n$ given by

$$p_n((x, y)) := \frac{1}{n} \sum_{i=1}^{n} 1\{(x, y) = (x_i, y_i)\}, \quad (x, y) \in \mathbb{R}^d \times \mathbb{R}.$$ 

Plug-in principle: Goal is to find function $f$ that minimizes (squared loss) risk

$$\mathcal{R}(f) = \mathbb{E}[(f(X) - Y)^2].$$

But we don’t know the distribution $P$ of $(X, Y)$.

Replace $P$ with $P_n \rightarrow \text{Empirical (squared loss) risk } \hat{\mathcal{R}}(f)$:

$$\hat{\mathcal{R}}(f) := \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2.$$
Empirical risk minimization (ERM) is the learning method that returns a function (from a specified function class) that minimizes the empirical risk.
Empirical risk minimization (ERM) is the learning method that returns a function (from a specified function class) that minimizes the empirical risk. For linear functions and squared loss: ERM returns

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^d} \hat{R}(\beta),$$

which coincides with MLE under the basic linear regression model.
Empirical risk minimization (ERM) is the learning method that returns a function (from a specified function class) that minimizes the empirical risk.

For linear functions and squared loss: ERM returns

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^d} \hat{R}(\beta),$$

which coincides with MLE under the basic linear regression model.

In general:

- MLE makes sense in context of statistical model for which it is derived.
- ERM makes sense in context of general iid model for supervised learning.
Empirical risk minimization in pictures

Red dots: data points.

Affine hyperplane: linear function $\beta$
(via affine expansion $(x_1, x_2) \mapsto (1, x_1, x_2)$).

ERM: minimize sum of squared vertical lengths from hyperplane to points.
Define $n \times d$ matrix $A$ and $n \times 1$ column vector $b$ by

$$
A := \frac{1}{\sqrt{n}} \begin{bmatrix}
\leftarrow & x_1^\top & \rightarrow \\
\vdots & \vdots & \vdots \\
\leftarrow & x_n^\top & \rightarrow 
\end{bmatrix}, \quad b := \frac{1}{\sqrt{n}} \begin{bmatrix}
y_1 \\
\vdots \\
y_n
\end{bmatrix}.
$$
Define $n \times d$ matrix $A$ and $n \times 1$ column vector $b$ by

$$A := \frac{1}{\sqrt{n}} \begin{bmatrix}
\leftarrow & x_1^\top & \rightarrow \\
\leftarrow & \vdots & \\
\leftarrow & x_n^\top & \rightarrow 
\end{bmatrix}, \quad b := \frac{1}{\sqrt{n}} \begin{bmatrix} y_1 \\
\vdots \\
y_n \end{bmatrix}.$$ 

Can write empirical risk as

$$\hat{R}(\beta) = \|A\beta - b\|_2^2.$$
Empirical risk minimization in matrix notation

Define $n \times d$ matrix $A$ and $n \times 1$ column vector $b$ by

$$A := \frac{1}{\sqrt{n}} \begin{bmatrix}
\leftarrow x_1^\top & \rightarrow \\
\vdots \\
\leftarrow x_n^\top & \rightarrow
\end{bmatrix}, \quad b := \frac{1}{\sqrt{n}} \begin{bmatrix}
y_1 \\
\vdots \\
y_n
\end{bmatrix}.$$ 

Can write empirical risk as

$$\hat{R}(\beta) = \|A\beta - b\|^2_2.$$ 

Necessary condition for $\beta$ to be a minimizer of $\hat{R}$:

$$\nabla \hat{R}(\beta) = 0,$$

i.e., $\beta$ is a critical point of $\hat{R}$. 

This translates to

$$(A^\top A)\beta = A^\top b,$$

a system of linear equations called the normal equations.

It can be proved that every critical point of $\hat{R}$ is a minimizer of $\hat{R}$. 

Empirical risk minimization in matrix notation

Define $n \times d$ matrix $A$ and $n \times 1$ column vector $b$ by

$$A := \frac{1}{\sqrt{n}} \begin{bmatrix} x_1^\top & \rightarrow \\ \vdots & \\ x_n^\top & \rightarrow \end{bmatrix}, \quad b := \frac{1}{\sqrt{n}} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$ 

Can write empirical risk as

$$\hat{R}(\beta) = \|A\beta - b\|_2^2.$$ 

Necessary condition for $\beta$ to be a minimizer of $\hat{R}$:

$$\nabla \hat{R}(\beta) = 0, \quad \text{i.e., } \beta \text{ is a critical point of } \hat{R}.$$ 

This translates to

$$(A^\top A)\beta = A^\top b,$$

a system of linear equations called the *normal equations*.
Define $n \times d$ matrix $A$ and $n \times 1$ column vector $b$ by

$$A := \frac{1}{\sqrt{n}} \begin{bmatrix} \leftarrow & x_1^\top & \rightarrow \newline \vdots & \vdots & \vdots \newline \leftarrow & x_n^\top & \rightarrow \end{bmatrix}, \quad b := \frac{1}{\sqrt{n}} \begin{bmatrix} y_1 \\
\vdots \\
y_n \end{bmatrix}.$$  

Can write empirical risk as

$$\hat{R}(\beta) = \| A\beta - b \|_2^2.$$  

Necessary condition for $\beta$ to be a minimizer of $\hat{R}$:

$$\nabla \hat{R}(\beta) = 0,$$

i.e., $\beta$ is a critical point of $\hat{R}$.  

This translates to

$$(A^\top A)\beta = A^\top b,$$

a system of linear equations called the normal equations.  

It can be proved that every critical point of $\hat{R}$ is a minimizer of $\hat{R}$:
Aside: Convexity

Let $f: \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.

Suppose we find $x \in \mathbb{R}^d$ such that $\nabla f(x) = 0$. Is $x$ a minimizer of $f$?
Aside: Convexity

Let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) be a differentiable function.

Suppose we find \( x \in \mathbb{R}^d \) such that \( \nabla f(x) = 0 \). Is \( x \) a minimizer of \( f \)?

Yes, if \( f \) is a convex function:

\[
 f((1 - t)x + tx') \leq (1 - t)f(x) + tf(x'),
\]

for any \( 0 \leq t \leq 1 \) and any \( x, x' \in \mathbb{R}^d \).
Convexity of empirical risk

Checking convexity of $g(x) = \|Ax - b\|_2^2$:

$$g((1 - t)x + tx')$$
Convexity of empirical risk

Checking convexity of \( g(x) = \|Ax - b\|^2 \):

\[
g((1 - t)x + tx') \\
= \|(1 - t)(Ax - b) + t(Ax' - b)\|^2
\]
Convexity of empirical risk

Checking convexity of $g(x) = \|Ax - b\|_2^2$:

\[
\begin{align*}
g((1 - t)x + tx') &= \|(1 - t)(Ax - b) + t(Ax' - b)\|_2^2 \\
&= (1 - t)^2 \|Ax - b\|_2^2 + t^2 \|Ax' - b\|_2^2 + 2(1 - t)t(Ax - b)^T(Ax' - b)
\end{align*}
\]
Checking convexity of $g(x) = \|Ax - b\|_2^2$:

$$g((1 - t)x + tx')$$

$$= \|(1 - t)(Ax - b) + t(Ax' - b)\|_2^2$$

$$= (1 - t)^2\|Ax - b\|_2^2 + t^2\|Ax' - b\|_2^2 + 2(1 - t)t(Ax - b)^T(Ax' - b)$$

$$= (1 - t)\|Ax - b\|_2^2 + t\|Ax' - b\|_2^2$$

$$- (1 - t)t[\|Ax - b\|_2^2 + \|Ax' - b\|_2^2] + 2(1 - t)t(Ax - b)^T(Ax' - b)$$

where last step uses Cauchy-Schwarz inequality and arithmetic mean/geometric mean (AM/GM) inequality.
Convexity of empirical risk

Checking convexity of \( g(x) = \|Ax - b\|^2 \):

\[
g((1 - t)x + tx') = \|(1 - t)(Ax - b) + t(Ax' - b)\|^2 \\
= (1 - t)^2\|Ax - b\|^2 + t^2\|Ax' - b\|^2 + 2(1 - t)t(Ax - b)^T(Ax' - b) \\
= (1 - t)\|Ax - b\|^2 + t\|Ax' - b\|^2 \\
- (1 - t)t[\|Ax - b\|^2 + \|Ax' - b\|^2] + 2(1 - t)t(Ax - b)^T(Ax' - b) \\
\leq (1 - t)\|Ax - b\|^2 + t\|Ax' - b\|^2
\]

where last step uses Cauchy-Schwarz inequality and arithmetic mean/geometric mean (AM/GM) inequality.
Convexity of empirical risk, another way

Preview of convex analysis

Recall \( \hat{\mathcal{R}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^\top \beta - y_i)^2 \).
Preview of convex analysis

Recall $\hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2$.

- Scalar function $g(z) = cz^2$ is convex for any $c \geq 0$. 

Convexity of empirical risk, another way

Convexity is a useful mathematical property to understand! (We'll study more convex analysis in a few weeks.)
Preview of convex analysis

Recall $\hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^\top \beta - y_i)^2$.

- Scalar function $g(z) = cz^2$ is convex for any $c \geq 0$.
- Composition $(g \circ a): \mathbb{R}^d \to \mathbb{R}$ of any convex function $g: \mathbb{R} \to \mathbb{R}$ and any affine function $a: \mathbb{R}^d \to \mathbb{R}$ is convex.
Convexity of empirical risk, another way

Preview of convex analysis

Recall $\hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2$.

- Scalar function $g(z) = cz^2$ is convex for any $c \geq 0$.

- Composition $(g \circ a): \mathbb{R}^d \rightarrow \mathbb{R}$ of any convex function $g: \mathbb{R} \rightarrow \mathbb{R}$ and any affine function $a: \mathbb{R}^d \rightarrow \mathbb{R}$ is convex.

- Therefore, function $\beta \mapsto \frac{1}{n} (x_i^T \beta - y_i)^2$ is convex.
Convexity of empirical risk, another way

Preview of convex analysis

Recall \( \hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2 \).

- Scalar function \( g(z) = cz^2 \) is convex for any \( c \geq 0 \).
- Composition \( (g \circ a): \mathbb{R}^d \rightarrow \mathbb{R} \) of any convex function \( g: \mathbb{R} \rightarrow \mathbb{R} \) and any affine function \( a: \mathbb{R}^d \rightarrow \mathbb{R} \) is convex.
- Therefore, function \( \beta \mapsto \frac{1}{n} (x_i^T \beta - y_i)^2 \) is convex.
- Sum of convex functions is convex.
Convexity of empirical risk, another way

Preview of convex analysis

Recall $\hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2$.

- Scalar function $g(z) = cz^2$ is convex for any $c \geq 0$.
- Composition $(g \circ a): \mathbb{R}^d \to \mathbb{R}$ of any convex function $g: \mathbb{R} \to \mathbb{R}$ and any affine function $a: \mathbb{R}^d \to \mathbb{R}$ is convex.
- Therefore, function $\beta \mapsto \frac{1}{n} (x_i^T \beta - y_i)^2$ is convex.
- Sum of convex functions is convex.
- Therefore $\hat{R}$ is convex.
Convexity of empirical risk, another way

Preview of convex analysis

Recall \( \hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2. \)

- Scalar function \( g(z) = cz^2 \) is convex for any \( c \geq 0. \)
- Composition \( (g \circ a): \mathbb{R}^d \to \mathbb{R} \) of any convex function \( g: \mathbb{R} \to \mathbb{R} \) and any affine function \( a: \mathbb{R}^d \to \mathbb{R} \) is convex.
- Therefore, function \( \beta \mapsto \frac{1}{n} (x_i^T \beta - y_i)^2 \) is convex.
- Sum of convex functions is convex.
- Therefore \( \hat{R} \) is convex.

Convexity is a useful mathematical property to understand!
(We’ll study more convex analysis in a few weeks.)
Algorithm for ERM

Algorithm for ERM with linear functions and squared loss†

input  Data \((x_1, y_1), \ldots, (x_n, y_n)\) from \(\mathbb{R}^d \times \mathbb{R}\).
output Linear function \(\hat{\beta} \in \mathbb{R}^d\).
1: Find solution \(\hat{\beta}\) to the normal equations defined by the data (using, e.g., Gaussian elimination).
2: return \(\hat{\beta}\).

†Also called “ordinary least squares” in this context.
Algorithm for ERM

Algorithm for ERM with linear functions and squared loss†

**input** Data \((x_1, y_1), \ldots, (x_n, y_n)\) from \(\mathbb{R}^d \times \mathbb{R}\).

**output** Linear function \(\hat{\beta} \in \mathbb{R}^d\).

1: Find solution \(\hat{\beta}\) to the normal equations defined by the data (using, e.g., Gaussian elimination).
2: return \(\hat{\beta}\).

† Also called “ordinary least squares” in this context.

**Running time** (dominated by Gaussian elimination): \(O(nd^2)\).

Note: there are many approximate solvers that run in nearly linear time!
Geometric interpretation of ERM

Let \( a_j \in \mathbb{R}^n \) be the \( j \)-th column of matrix \( A \in \mathbb{R}^{n \times d} \), so

\[
A = \begin{bmatrix}
\uparrow & \uparrow \\
\downarrow & \downarrow \\
a_1 & \cdots & a_d
\end{bmatrix}.
\]

Minimizing \( \| A\beta - b \|_2 \) is the same as finding vector \( \hat{b} \in \text{range}(A) \) closest to \( b \).

Solution \( \hat{b} \) is orthogonal projection of \( b \) onto \( \text{range}(A) = \{ A\beta : \beta \in \mathbb{R}^d \} \).

\( \hat{b} \) is uniquely determined.

If \( \text{rank}(A) < d \), then \( \hat{b} \) is not unique.

To get \( \beta \) from \( \hat{b} \): solve system of linear equations \( A\beta = \hat{b} \).
Geometric interpretation of ERM

Let $a_j \in \mathbb{R}^n$ be the $j$-th column of matrix $A \in \mathbb{R}^{n \times d}$, so

$$A = \begin{bmatrix}
\uparrow & \cdots & \uparrow \\
a_1 & \cdots & a_d \\
\downarrow & \cdots & \downarrow
\end{bmatrix}.$$ 

Minimizing $\|A\beta - b\|_2^2$ is the same as finding vector $\hat{b} \in \text{range}(A)$ closest to $b$. 

$\hat{b}$ is uniquely determined.

If $\text{rank}(A) < d$, then

1. way to write $\hat{b}$ as linear combination of $a_1, \ldots, a_d$.

If $\text{rank}(A) < d$, then ERM solution is not unique.

To get $\beta$ from $\hat{b}$: solve system of linear equations $A\beta = \hat{b}$. 

Geometric interpretation of ERM

Let $a_j \in \mathbb{R}^n$ be the $j$-th column of matrix $A \in \mathbb{R}^{n \times d}$, so

$$A = \begin{bmatrix}
\uparrow & \uparrow \\
\downarrow & \downarrow \\
a_1 & \cdots & a_d \\
\end{bmatrix}.$$ 

Minimizing $\|A\beta - b\|_2^2$ is the same as finding vector $\hat{b} \in \text{range}(A)$ closest to $b$.

Solution $\hat{b}$ is \textit{orthogonal projection} of $b$ onto $\text{range}(A) = \{A\beta : \beta \in \mathbb{R}^d\}$.
Geometric interpretation of ERM

Let \( a_j \in \mathbb{R}^n \) be the \( j \)-th column of matrix \( A \in \mathbb{R}^{n \times d} \), so

\[
A = \begin{bmatrix}
\uparrow & \cdots & \uparrow \\
\downarrow & \ddots & \downarrow \\
a_1 & \cdots & a_d \\
\end{bmatrix}.
\]

Minimizing \( \| A\beta - b \|_2^2 \) is the same as finding vector \( \hat{b} \in \text{range}(A) \) closest to \( b \).

Solution \( \hat{b} \) is orthogonal projection of \( b \) onto \( \text{range}(A) = \{ A\beta : \beta \in \mathbb{R}^d \} \).

\( \hat{b} \) is uniquely determined.
Geometric interpretation of ERM

Let $a_j \in \mathbb{R}^n$ be the $j$-th column of matrix $A \in \mathbb{R}^{n \times d}$, so

$$A = \begin{bmatrix}
\uparrow & \cdots & \uparrow \\
a_1 & \cdots & a_d \\
\downarrow & \cdots & \downarrow
\end{bmatrix}.$$

Minimizing $\|A\beta - b\|_2^2$ is the same as finding vector $\hat{b} \in \text{range}(A)$ closest to $b$.

Solution $\hat{b}$ is orthogonal projection of $b$ onto $\text{range}(A) = \{A\beta : \beta \in \mathbb{R}^d\}$.

- $\hat{b}$ is uniquely determined.
- If $\text{rank}(A) < d$, then $\geq 1$ way to write $\hat{b}$ as linear combination of $a_1, \ldots, a_d$. 
Geometric interpretation of ERM

Let $a_j \in \mathbb{R}^n$ be the $j$-th column of matrix $A \in \mathbb{R}^{n \times d}$, so

$$A = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ a_1 & \cdots & a_d \\ \downarrow & \cdots & \downarrow \end{bmatrix}.$$ 

Minimizing $\|A\beta - b\|_2^2$ is the same as finding vector $\hat{b} \in \text{range}(A)$ closest to $b$.

Solution $\hat{b}$ is **orthogonal projection** of $b$ onto $\text{range}(A) = \{A\beta : \beta \in \mathbb{R}^d\}$.

- $\hat{b}$ is uniquely determined.
- If $\text{rank}(A) < d$, then $>1$ way to write $\hat{b}$ as linear combination of $a_1, \ldots, a_d$.

If $\text{rank}(A) < d$, then **ERM solution is not unique.**
Geometric interpretation of ERM

Let \( a_j \in \mathbb{R}^n \) be the \( j \)-th column of matrix \( A \in \mathbb{R}^{n \times d} \), so

\[
A = \begin{bmatrix}
\uparrow & \cdots & \uparrow \\
a_1 & \cdots & a_d \\
\downarrow & \cdots & \downarrow
\end{bmatrix}.
\]

Minimizing \( \| A\beta - b \|_2^2 \) is the same as finding vector \( \hat{b} \in \text{range}(A) \) closest to \( b \).

Solution \( \hat{b} \) is orthogonal projection of \( b \) onto \( \text{range}(A) = \{ A\beta : \beta \in \mathbb{R}^d \} \).

- \( \hat{b} \) is uniquely determined.
- If \( \text{rank}(A) < d \), then \( \geq 1 \) way to write \( \hat{b} \) as linear combination of \( a_1, \ldots, a_d \).

If \( \text{rank}(A) < d \), then ERM solution is not unique.

To get \( \beta \) from \( \hat{b} \):

solve system of linear equations \( A\beta = \hat{b} \).
Let \((X, Y) \sim P\), where \(P\) is some distribution on \(\mathbb{R}^d \times \mathbb{R}\). Which \(\beta\) have smallest risk \(R(\beta) = \mathbb{E}[(X^T\beta - Y)^2]\)?
Statistical interpretation of ERM

Let \((X, Y) \sim P\), where \(P\) is some distribution on \(\mathbb{R}^d \times \mathbb{R}\). Which \(\beta\) have smallest risk \(\mathcal{R}(\beta) = \mathbb{E}[(X^T\beta - Y)^2]\)?

Necessary condition for \(\beta\) to be a minimizer of \(\mathcal{R}\):

\[
\nabla \mathcal{R}(\beta) = 0, \quad \text{i.e., } \beta \text{ is a critical point of } \mathcal{R}.
\]
Let \((X, Y) \sim P\), where \(P\) is some distribution on \(\mathbb{R}^d \times \mathbb{R}\). Which \(\beta\) have smallest risk \(R(\beta) = \mathbb{E}[(X^T \beta - Y)^2]\)?

Necessary condition for \(\beta\) to be a minimizer of \(R\):

\[
\nabla R(\beta) = 0, \quad \text{i.e., } \beta \text{ is a critical point of } R.
\]

This translates to

\[
\mathbb{E}[XX^T]\beta = \mathbb{E}[YX],
\]

a system of linear equations called the population normal equations.

It can be proved that every critical point of \(R\) is a minimizer of \(R\).
Statistical interpretation of ERM

Let \((X, Y) \sim P\), where \(P\) is some distribution on \(\mathbb{R}^d \times \mathbb{R}\). Which \(\beta\) have smallest risk \(R(\beta) = E[(X^T\beta - Y)^2]\)?

Necessary condition for \(\beta\) to be a minimizer of \(R\):

\[
\nabla R(\beta) = 0, \text{ i.e., } \beta \text{ is a critical point of } R.
\]

This translates to

\[
E[XX^T]\beta = E[YX],
\]

a system of linear equations called the \textit{population normal equations}.

It can be proved that every critical point of \(R\) is a minimizer of \(R\).

Looks familiar?
Statistical interpretation of ERM

Let \((X, Y) \sim P\), where \(P\) is some distribution on \(\mathbb{R}^d \times \mathbb{R}\).
Which \(\beta\) have smallest risk \(\mathcal{R}(\beta) = \mathbb{E}[(X^T \beta - Y)^2]\)?

Necessary condition for \(\beta\) to be a minimizer of \(\mathcal{R}\):
\[
\nabla \mathcal{R}(\beta) = 0, \quad \text{i.e., } \beta \text{ is a critical point of } \mathcal{R}.
\]

This translates to
\[
\mathbb{E}[XX^T] \beta = \mathbb{E}[YX],
\]
a system of linear equations called the population normal equations.
It can be proved that every critical point of \(\mathcal{R}\) is a minimizer of \(\mathcal{R}\).

Looks familiar?

If \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, then
\[
\mathbb{E}[A^T A] = \mathbb{E}[XX^T] \quad \text{and} \quad \mathbb{E}[A^T b] = \mathbb{E}[YX],
\]
so ERM can be regarded as a plug-in estimator for a minimizer of \(\mathcal{R}\).
4. Risk, empirical risk, and estimating risk
IID model: \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, taking values in \(\mathbb{R}^d \times \mathbb{R}\).

Let \(\beta^*\) be a minimizer of \(R\) over all \(\beta \in \mathbb{R}^d\), i.e., \(\beta^*\) satisfies population normal equations

\[
\mathbb{E}[XX^\top] \beta^* = \mathbb{E}[YX].
\]
**IID model:** \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, taking values in \(\mathbb{R}^d \times \mathbb{R}\).

Let \(\beta^*\) be a minimizer of \(R\) over all \(\beta \in \mathbb{R}^d\), i.e., \(\beta^*\) satisfies population normal equations

\[
\mathbb{E}[XX^\top] \beta^* = \mathbb{E}[YX].
\]

▶ If ERM solution \(\hat{\beta}\) is not unique (e.g., if \(n < d\)), then \(R(\hat{\beta})\) can be arbitrarily worse than \(R(\beta^*)\).
IID model: \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, taking values in \(\mathbb{R}^d \times \mathbb{R}\).

Let \(\beta^*\) be a minimizer of \(\mathcal{R}\) over all \(\beta \in \mathbb{R}^d\), i.e., \(\beta^*\) satisfies population normal equations
\[
\mathbb{E}[XX^\top]\beta^* = \mathbb{E}[YX].
\]

- If ERM solution \(\hat{\beta}\) is not unique (e.g., if \(n < d\)), then \(\mathcal{R}(\hat{\beta})\) can be arbitrarily worse than \(\mathcal{R}(\beta^*)\).
- What about when ERM solution is unique?
Risk of ERM

**IID model:** \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, taking values in \(\mathbb{R}^d \times \mathbb{R}\).

Let \(\beta^*\) be a minimizer of \(\mathcal{R}\) over all \(\beta \in \mathbb{R}^d\), i.e., \(\beta^*\) satisfies population normal equations

\[
\mathbb{E}[XX^T]\beta^* = \mathbb{E}[YX].
\]

- If ERM solution \(\hat{\beta}\) is not unique (e.g., if \(n < d\)), then \(\mathcal{R}(\hat{\beta})\) can be arbitrarily worse than \(\mathcal{R}(\beta^*)\).

- What about when ERM solution is unique?

**Theorem.** Under mild assumptions on distribution of \(X\),

\[
\mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta^*) = O \left( \frac{\text{tr}(\text{cov}(\varepsilon W))}{n} \right)
\]

“asymptotically”, where \(W := \mathbb{E}[XX^T]^{-\frac{1}{2}} X\) and \(\varepsilon := Y - X^T\beta^*\).
Let $\varepsilon_i := Y_i - X_i^T \beta^*$ for each $i = 1, \ldots, n$, so

$$E[\varepsilon_i X_i] = E[Y_i X_i] - E[X_i X_i^T] \beta^* = 0$$

and

$$\sqrt{n}(\hat{\beta} - \beta^*) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i X_i.$$
Risk of ERM analysis (rough sketch)

Let $\varepsilon_i := Y_i - \mathbf{X}_i^T \beta^*$ for each $i = 1, \ldots, n$, so

$$
\mathbb{E}[\varepsilon_i \mathbf{X}_i] = \mathbb{E}[Y_i \mathbf{X}_i] - \mathbb{E}[\mathbf{X}_i \mathbf{X}_i^T] \beta^* = 0
$$

and

$$
\sqrt{n}(\hat{\beta} - \beta^*) = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \mathbf{X}_i.
$$

1. By LLN: $\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \xrightarrow{p} \mathbb{E}[\mathbf{X} \mathbf{X}^T]$
Risk of ERM analysis (rough sketch)

Let \( \varepsilon_i := Y_i - X_i^T \beta^* \) for each \( i = 1, \ldots, n \), so

\[
\mathbb{E}[\varepsilon_i X_i] = \mathbb{E}[Y_i X_i] - \mathbb{E}[X_i X_i^T] \beta^* = 0
\]

and

\[
\sqrt{n}(\hat{\beta} - \beta^*) = \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i X_i.
\]

1. By LLN: \( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \xrightarrow{p} \mathbb{E}[X X^T] \)

2. By CLT: \( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i X_i \xrightarrow{d} \text{cov}(\varepsilon X)^{\frac{1}{2}} Z \), where \( Z \sim \text{N}(0, I) \).
Risk of ERM analysis (rough sketch)

Let \( \varepsilon_i := Y_i - X_i^T \beta^* \) for each \( i = 1, \ldots, n \), so

\[
\mathbb{E}[\varepsilon_i X_i] = \mathbb{E}[Y_i X_i] - \mathbb{E}[X_i X_i^T] \beta^* = 0
\]

and

\[
\sqrt{n} (\hat{\beta} - \beta^*) = \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i X_i.
\]

1. By LLN: \( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \xrightarrow{p} \mathbb{E}[XX^T] \)

2. By CLT: \( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i X_i \xrightarrow{d} \text{cov}(\varepsilon X)^{\frac{1}{2}} Z, \) where \( Z \sim N(0, I) \).

Therefore, asymptotic distribution of \( \sqrt{n}(\hat{\beta} - \beta^*) \) is

\[
\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} \mathbb{E}[XX^T]^{-1} \text{cov}(\varepsilon X)^{\frac{1}{2}} Z.
\]
Risk of ERM analysis (rough sketch)

Let \( \varepsilon_i := Y_i - X_i^T \beta^* \) for each \( i = 1, \ldots, n \), so

\[
\mathbb{E}[\varepsilon_i \mathbf{X}_i] = \mathbb{E}[Y_i \mathbf{X}_i] - \mathbb{E}[\mathbf{X}_i \mathbf{X}_i^T] \beta^* = 0
\]

and

\[
\sqrt{n}(\hat{\beta} - \beta^*) = \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \mathbf{X}_i.
\]

1. By LLN: \( \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \xrightarrow{p} \mathbb{E}[\mathbf{X} \mathbf{X}^T] \)

2. By CLT: \( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \mathbf{X}_i \xrightarrow{d} \text{cov}(\varepsilon \mathbf{X})^\frac{1}{2} \mathbf{Z}, \) where \( \mathbf{Z} \sim \mathcal{N}(0, I) \).

Therefore, asymptotic distribution of \( \sqrt{n}(\hat{\beta} - \beta^*) \) is

\[
\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} \mathbb{E}[\mathbf{X} \mathbf{X}^T]^{-1} \text{cov}(\varepsilon \mathbf{X})^\frac{1}{2} \mathbf{Z}.
\]

A few more steps gives

\[
n \left( \mathbb{E}[(\mathbf{X}^T \hat{\beta} - Y)^2] - \mathbb{E}[(\mathbf{X}^T \beta^* - Y)^2] \right) \xrightarrow{d} \| \mathbb{E}[\mathbf{X} \mathbf{X}^T]^{-\frac{1}{2}} \text{cov}(\varepsilon \mathbf{X})^\frac{1}{2} \mathbf{Z} \|^2.
\]

Random variable on RHS is “concentrated” around its mean tr(cov(\( \varepsilon \mathbf{W} \))).
Analysis does not assume that the linear regression model is “correct”; the data distribution need not be from normal linear regression model.
Risk of ERM: postscript

- Analysis does not assume that the linear regression model is “correct”; the data distribution need not be from normal linear regression model.

- Only assumptions are those needed for LLN and CLT to hold.

\[ Y \mid X = x \sim N(x^T \beta^\star, \sigma^2), \]

which is familiar to those who have taken introductory statistics.

With more work, can also prove non-asymptotic risk bound of similar form.

In homework/reading, we look at a simpler setting for studying ERM for linear regression, called “fixed design.”
Analysis does not assume that the linear regression model is “correct”; the data distribution need not be from normal linear regression model.

Only assumptions are those needed for LLN and CLT to hold.

However, if normal linear regression model holds, i.e.,

\[ Y \mid X = x \sim N(x^T\beta^*, \sigma^2), \]

then the bound from the theorem becomes

\[ \mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta^*) = O\left(\frac{\sigma^2 d}{n}\right), \]

which is familiar to those who have taken introductory statistics.
Analysis does not assume that the linear regression model is “correct”; the data distribution need not be from normal linear regression model.

Only assumptions are those needed for LLN and CLT to hold.

However, if normal linear regression model holds, i.e.,

\[ Y \mid X = x \sim N(x^T \beta^*, \sigma^2), \]

then the bound from the theorem becomes

\[ \mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta^*) = O \left( \frac{\sigma^2 d}{n} \right), \]

which is familiar to those who have taken introductory statistics.

With more work, can also prove non-asymptotic risk bound of similar form.
Analysis does not assume that the linear regression model is “correct”; the data distribution need not be from normal linear regression model.

Only assumptions are those needed for LLN and CLT to hold.

However, if normal linear regression model holds, i.e.,

\[ Y \mid X = x \sim N(x^T \beta^*, \sigma^2), \]

then the bound from the theorem becomes

\[ \mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta^*) = O \left( \frac{\sigma^2 d}{n} \right), \]

which is familiar to those who have taken introductory statistics.

With more work, can also prove non-asymptotic risk bound of similar form.

In homework/reading, we look at a simpler setting for studying ERM for linear regression, called “fixed design”.

\[ \]
Let $\hat{\beta}$ be ERM solution.
Risk vs empirical risk

Let $\hat{\beta}$ be ERM solution.

1. Empirical risk of ERM: $\hat{R}(\hat{\beta})$
Let $\hat{\beta}$ be ERM solution.

1. Empirical risk of ERM: $\hat{R}(\hat{\beta})$
2. True risk of ERM: $R(\hat{\beta})$
Let $\hat{\beta}$ be ERM solution.

1. Empirical risk of ERM: $\hat{R}(\hat{\beta})$
2. True risk of ERM: $R(\hat{\beta})$

**Theorem.**

$$\mathbb{E}[\hat{R}(\hat{\beta})] \leq \mathbb{E}[R(\hat{\beta})].$$

(Empirical risk can sometimes be larger than true risk, but not on average.)
Let $\hat{\beta}$ be ERM solution.

1. Empirical risk of ERM: $\hat{R}(\hat{\beta})$

2. True risk of ERM: $R(\hat{\beta})$

**Theorem.**

$$\mathbb{E}[\hat{R}(\hat{\beta})] \leq \mathbb{E}[R(\hat{\beta})].$$

(Empirical risk can sometimes be larger than true risk, but not on average.)

**Overfitting:** empirical risk is “small”, but true risk is “much higher”.
(\(X_1, Y_1\), \ldots, \(X_n, Y_n\), \((X, Y)\)) are iid; \(X\) is continuous random variable in \(\mathbb{R}\). Suppose we use degree-\(k\) polynomial expansion

\[
\phi(x) = (1, x^1, \ldots, x^k), \quad x \in \mathbb{R},
\]

so dimension is \(d = k + 1\).
Overfitting example

\((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid; \(X\) is continuous random variable in \(\mathbb{R}\).

Suppose we use degree-\(k\) polynomial expansion

\[ \phi(x) = (1, x^1, \ldots, x^k), \quad x \in \mathbb{R}, \]

so dimension is \(d = k + 1\).

**Fact**: Any function on \(\leq k + 1\) points can be interpolated by a polynomial of degree at most \(k\).

Conclusion: If \(n \leq k + 1 = d\), ERM solution \(\hat{\beta}\) with this feature expansion has \(\hat{\mathcal{R}}(\hat{\beta}) = 0\) always, regardless of its true risk (which can be \(\gg 0\)).
Overfitting example

\((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid; \(X\) is continuous random variable in \(\mathbb{R}\).

Suppose we use degree-\(k\) polynomial expansion

\[
\phi(x) = (1, x^1, \ldots, x^k), \quad x \in \mathbb{R},
\]

so dimension is \(d = k + 1\).

Fact: Any function on \(\leq k + 1\) points can be interpolated by a polynomial of degree at most \(k\).

Conclusion: If \(n \leq k + 1 = d\), ERM solution \(\hat{\beta}\) with this feature expansion has \(\hat{R}(\hat{\beta}) = 0\) always, regardless of its true risk (which can be \(\gg 0\)).
Estimating risk

**IID model:** $(X_1, Y_1), \ldots, (X_n, Y_n), (\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m) \sim_{iid} P$. 
Estimating risk

IID model: \((X_1, Y_1), \ldots, (X_n, Y_n), (\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m) \sim_{\text{iid}} P\).

- **training data** \((X_1, Y_1), \ldots, (X_n, Y_n)\) used to learn \(\hat{f}\).
Estimating risk

IID model: \((X_1, Y_1), \ldots, (X_n, Y_n), (\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m) \sim_{iid} P\).

- **training data** \((X_1, Y_1), \ldots, (X_n, Y_n)\) used to learn \(\hat{f}\).
- **test data** \((\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m)\) used to estimate risk, via **test risk**

\[
\hat{R}_{test}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(\tilde{X}_i) - \tilde{Y}_i)^2.
\]
Estimating risk

IID model: \((X_1, Y_1), \ldots, (X_n, Y_n), (\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m) \sim \text{iid } P.\)

- **training data** \((X_1, Y_1), \ldots, (X_n, Y_n)\) used to learn \(\hat{f}\).
- **test data** \((\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m)\) used to estimate risk, via **test risk**

\[
\hat{R}_{\text{test}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(\tilde{X}_i) - \tilde{Y}_i)^2.
\]

- Training data is independent of test data, so \(\hat{f}\) is independent of test data.
Estimating risk

IID model: \((X_1, Y_1), \ldots, (X_n, Y_n), (\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m) \sim_{\text{iid}} P.\)

- **training data** \((X_1, Y_1), \ldots, (X_n, Y_n)\) used to learn \(\hat{f}.\)
- **test data** \((\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m)\) used to estimate risk, via **test risk**

\[
\hat{R}_{\text{test}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(\tilde{X}_i) - \tilde{Y}_i)^2.
\]

- Training data is independent of test data, so \(\hat{f}\) is independent of test data.
- Let \(L_i := (\hat{f}(\tilde{X}_i) - \tilde{Y}_i)^2\) for each \(i = 1, \ldots, m\), so

\[
\mathbb{E} \left[ \hat{R}_{\text{test}}(\hat{f}) \mid \hat{f} \right] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} \left[ L_i \mid \hat{f} \right] = \mathcal{R}(\hat{f}).
\]
Estimating risk

**IID model:** \((X_1, Y_1), \ldots, (X_n, Y_n), (\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m) \sim_{\text{iid}} P.\)

- **training data** \((X_1, Y_1), \ldots, (X_n, Y_n)\) used to learn \(\hat{f}\).
- **test data** \((\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m)\) used to estimate risk, via test risk

\[
\hat{R}_{\text{test}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(\tilde{X}_i) - \tilde{Y}_i)^2.
\]

- Training data is independent of test data, so \(\hat{f}\) is independent of test data.
- Let \(L_i := (\hat{f}(\tilde{X}_i) - \tilde{Y}_i)^2\) for each \(i = 1, \ldots, m\), so

\[
\mathbb{E} \left[ \hat{R}_{\text{test}}(\hat{f}) \mid \hat{f} \right] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} \left[ L_i \mid \hat{f} \right] = \mathcal{R}(\hat{f}).
\]

- Moreover, \(L_1, \ldots, L_m\) are conditionally iid given \(\hat{f}\), and hence by Law of Large Numbers,

\[
\hat{R}_{\text{test}}(\hat{f}) \xrightarrow{p} \mathcal{R}(\hat{f}) \quad \text{as} \ m \to \infty.
\]
Estimating risk

IID model: \((X_1, Y_1), \ldots, (X_n, Y_n), (\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m) \sim_{\text{iid}} P.\)

- **training data** \((X_1, Y_1), \ldots, (X_n, Y_n)\) used to learn \(\hat{f}\).
- **test data** \((\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m)\) used to estimate risk, via test risk

\[
\hat{R}_{\text{test}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(\tilde{X}_i) - \tilde{Y}_i)^2.
\]

- Training data is independent of test data, so \(\hat{f}\) is independent of test data.
- Let \(L_i := (\hat{f}(\tilde{X}_i) - \tilde{Y}_i)^2\) for each \(i = 1, \ldots, m\), so

\[
\mathbb{E} \left[ \hat{R}_{\text{test}}(\hat{f}) \mid \hat{f} \right] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} \left[ L_i \mid \hat{f} \right] = \mathcal{R}(\hat{f}).
\]

- Moreover, \(L_1, \ldots, L_m\) are conditionally iid given \(\hat{f}\), and hence by Law of Large Numbers,

\[
\hat{R}_{\text{test}}(\hat{f}) \xrightarrow{p} \mathcal{R}(\hat{f}) \quad \text{as} \quad m \to \infty.
\]

- By CLT, the rate of convergence is \(m^{-1/2}\).
Rates for risk minimization vs. rates for risk estimation

One may think that ERM “works” because, somehow, training risk is a good “plug-in” estimate of true risk.
Rates for risk minimization vs. rates for risk estimation

One may think that ERM “works” because, somehow, training risk is a good “plug-in” estimate of true risk.

- Sometimes this is partially true—we’ll revisit this when we discuss generalization theory.

Roughly speaking, under some assumptions, can expect that

$$|\hat{R}(\beta) - R(\beta)| \leq O\left(\sqrt{\frac{d}{n}}\right)$$

for all $\beta \in \mathbb{R}^d$.

Implication: Selecting a good predictor can be “easier” than estimating how good predictors are!
Rates for risk minimization vs. rates for risk estimation

One may think that ERM “works” because, somehow, training risk is a good “plug-in” estimate of true risk.

- Sometimes this is partially true—we’ll revisit this when we discuss generalization theory.

Roughly speaking, under some assumptions, can expect that

\[ |\hat{R}(\beta) - R(\beta)| \leq O\left(\sqrt{\frac{d}{n}}\right) \text{ for all } \beta \in \mathbb{R}^d. \]

However . . .
One may think that ERM “works” because, somehow, training risk is a good “plug-in” estimate of true risk.

- Sometimes this is partially true—we’ll revisit this when we discuss generalization theory.

  Roughly speaking, under some assumptions, can expect that

  \[ |\hat{R}(\beta) - R(\beta)| \leq O\left(\sqrt{\frac{d}{n}}\right) \text{ for all } \beta \in \mathbb{R}^d. \]

However . . .

- By CLT, we know the following holds for a fixed \( \beta \):

  \[ \hat{R}(\beta) \overset{p}{\longrightarrow} R(\beta) \text{ at } n^{-1/2} \text{ rate.} \]

  (Here, we ignore the dependence on \( d \).)
Rates for risk minimization vs. rates for risk estimation

One may think that ERM “works” because, somehow, training risk is a good “plug-in” estimate of true risk.

- Sometimes this is partially true—we’ll revisit this when we discuss generalization theory.
  Roughly speaking, under some assumptions, can expect that
  \[ |\hat{R}(\beta) - R(\beta)| \leq O\left(\sqrt{\frac{d}{n}}\right) \quad \text{for all } \beta \in \mathbb{R}^d. \]

However . . .

- By CLT, we know the following holds for a fixed \( \beta \):
  \[ \hat{R}(\beta) \xrightarrow{p} R(\beta) \quad \text{at } n^{-1/2} \text{ rate.} \]
  (Here, we ignore the dependence on \( d \).)

- Yet, for ERM \( \hat{\beta} \),
  \[ R(\hat{\beta}) \xrightarrow{p} R(\beta^*) \quad \text{at } n^{-1} \text{ rate.} \]
  (Also ignoring dependence on \( d \).)
Rates for risk minimization vs. rates for risk estimation

One may think that ERM “works” because, somehow, training risk is a good “plug-in” estimate of true risk.

- Sometimes this is partially true—we’ll revisit this when we discuss generalization theory.
  Roughly speaking, under some assumptions, can expect that

  \[ |\hat{R}(\beta) - R(\beta)| \leq O\left(\sqrt{\frac{d}{n}}\right) \]

  for all \( \beta \in \mathbb{R}^d \).

  However . . .

- By CLT, we know the following holds for a fixed \( \beta \):

  \[ \hat{R}(\beta) \xrightarrow{p} R(\beta) \] at \( n^{-1/2} \) rate.

  (Here, we ignore the dependence on \( d \).)

- Yet, for ERM \( \hat{\beta} \),

  \[ R(\hat{\beta}) \xrightarrow{p} R(\beta^*) \]

  at \( n^{-1} \) rate.

  (Also ignoring dependence on \( d \).)

**Implication:** Selecting a good predictor can be “easier” than estimating how good predictors are!
Old Faithful example

Linear regression model + affine expansion on "duration of last eruption".

Learn $\hat{\beta} = (35.0929, 10.3258)$ from 136 past observations.

Mean squared loss of $\hat{\beta}$ on next 136 observations is 35.9404.

(Recall: mean squared loss of $\hat{\mu} = 70.7941$ was 187.1894.)

Unfortunately, $\sqrt{35.9} > \text{mean duration} \approx 3.5$. 

0 1 2 3 4 5 6

duration of last eruption

0
20
40
60
80
100
time until next eruption

linear model
constant prediction

\begin{center}
\begin{tabular}{c}
\hline
0 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
\end{tabular}
\end{center}
Old Faithful example

- Linear regression model + affine expansion on “duration of last eruption”.

\[
\hat{\beta} = (35.0929, 10.3258)
\]

Learn \( \hat{\beta} \) from 136 past observations.

Mean squared loss of \( \hat{\beta} \) on next 136 observations is 35.9404.

(Recall: mean squared loss of \( \hat{\mu} = 70.7941 \) was 187.1894.)

\( \sqrt{35.9} > \text{mean duration} \approx 3.5 \).
Old Faithful example

- Linear regression model + affine expansion on “duration of last eruption”.
- Learn $\hat{\beta} = (35.0929, 10.3258)$ from 136 past observations.
Old Faithful example

- Linear regression model + affine expansion on “duration of last eruption”.
- Learn $\hat{\beta} = (35.0929, 10.3258)$ from 136 past observations.
- Mean squared loss of $\hat{\beta}$ on next 136 observations is 35.9404. (Recall: mean squared loss of $\hat{\mu} = 70.7941$ was 187.1894.)
Old Faithful example

- Linear regression model + affine expansion on “duration of last eruption”.
- Learn $\hat{\beta} = (35.0929, 10.3258)$ from 136 past observations.
- Mean squared loss of $\hat{\beta}$ on next 136 observations is 35.9404. (Recall: mean squared loss of $\hat{\mu} = 70.7941$ was 187.1894.)

(Unfortunately, $\sqrt{35.9} > \text{mean duration} \approx 3.5.$)
5. Regularization
Inductive bias

Suppose ERM solution is not unique. What should we do?

One possible answer: Pick the $\hat{\beta}$ of shortest length.

Fact: The shortest solution $\hat{\beta}$ to $(A^T A) \beta = A^T b$ is always unique.

Obtain $\hat{\beta}$ via $\hat{\beta} = A^\dagger b$ where $A^\dagger$ is the (Moore-Penrose) pseudoinverse of $A$.

Why should this be a good idea?

Data does not give reason to choose a shorter $\beta$ over a longer $\beta$.

The preference for shorter $\beta$ is an inductive bias: it will work well for some problems (e.g., when "true" $\beta^\star$ is short), not for others.

All learning algorithms encode some kind of inductive bias.
Suppose ERM solution is **not unique**. What should we do?

**One possible answer:** Pick the $\beta$ of shortest length.
Suppose ERM solution is not unique. What should we do?

**One possible answer:** Pick the $\beta$ of shortest length.

- **Fact:** The shortest solution $\hat{\beta}$ to $(A^T A)\beta = A^T b$ is always unique.
Suppose ERM solution is not unique. What should we do?

**One possible answer**: Pick the \( \beta \) of shortest length.

- **Fact**: The shortest solution \( \hat{\beta} \) to \( (A^T A) \beta = A^T b \) is always unique.
- Obtain \( \hat{\beta} \) via
  \[
  \hat{\beta} = A^\dagger b
  \]
  where \( A^\dagger \) is the *(Moore-Penrose) pseudoinverse* of \( A \).
Inductive bias

Suppose ERM solution is not unique. What should we do?

**One possible answer:** Pick the $\beta$ of shortest length.

- **Fact:** The shortest solution $\hat{\beta}$ to $(A^T A)\beta = A^T b$ is always unique.

- Obtain $\hat{\beta}$ via
  \[
  \hat{\beta} = A^\dagger b
  \]
  where $A^\dagger$ is the *(Moore-Penrose) pseudoinverse* of $A$.

Why should this be a good idea?
Suppose ERM solution is not unique. What should we do?

One possible answer: Pick the $\beta$ of shortest length.

- Fact: The shortest solution $\hat{\beta}$ to $(A^T A)\beta = A^T b$ is always unique.
- Obtain $\hat{\beta}$ via

$$\hat{\beta} = A^+ b$$

where $A^+$ is the (Moore-Penrose) pseudoinverse of $A$.

Why should this be a good idea?

- Data does not give reason to choose a shorter $\beta$ over a longer $\beta$. 

Inductive bias
Suppose ERM solution is **not unique**. What should we do?

**One possible answer**: Pick the $\beta$ of shortest length.

- **Fact**: The shortest solution $\hat{\beta}$ to $(A^T A) \beta = A^T b$ is always unique.
- Obtain $\hat{\beta}$ via

  $$\hat{\beta} = A^\dagger b$$

  where $A^\dagger$ is the *(Moore-Penrose) pseudoinverse* of $A$.

Why should this be a good idea?

- Data does not give reason to choose a shorter $\beta$ over a longer $\beta$.
- The preference for shorter $\beta$ is an **inductive bias**: it will work well for some problems (e.g., when “true” $\beta^*$ is short), not for others.
Suppose ERM solution is not unique. What should we do?

**One possible answer**: Pick the $\beta$ of shortest length.

- **Fact**: The shortest solution $\hat{\beta}$ to $(A^T A)\beta = A^T b$ is always unique.
- Obtain $\hat{\beta}$ via
  $$\hat{\beta} = A^\dagger b$$
  where $A^\dagger$ is the *(Moore-Penrose) pseudoinverse* of $A$.

Why should this be a good idea?

- Data does not give reason to choose a shorter $\beta$ over a longer $\beta$.
- The preference for shorter $\beta$ is an *inductive bias*: it will work well for some problems (e.g., when “true” $\beta^*$ is short), not for others.

All learning algorithms encode some kind of inductive bias.
Example

ERM with scaled trigonometric feature expansion:

$$\phi(x) = (1, \sin(x), \cos(x), \frac{1}{2} \sin(2x), \frac{1}{2} \cos(2x), \frac{1}{3} \sin(3x), \frac{1}{3} \cos(3x), \ldots).$$
ERM with scaled trigonometric feature expansion:

\[ \phi(x) = (1, \sin(x), \cos(x), \frac{1}{2} \sin(2x), \frac{1}{2} \cos(2x), \frac{1}{3} \sin(3x), \frac{1}{3} \cos(3x), \ldots). \]

Training data:
ERM with scaled trigonometric feature expansion:

\[ \phi(x) = (1, \sin(x), \cos(x), \frac{1}{2} \sin(2x), \frac{1}{2} \cos(2x), \frac{1}{3} \sin(3x), \frac{1}{3} \cos(3x), \ldots). \]
Example

ERM with scaled trigonometric feature expansion:

\[ \phi(x) = (1, \sin(x), \cos(x), \frac{1}{2} \sin(2x), \frac{1}{2} \cos(2x), \frac{1}{3} \sin(3x), \frac{1}{3} \cos(3x), \ldots). \]

Training data and least $\ell_2$ norm ERM:

It is *not* a given that the least norm ERM is better than the other ERM!
Regularized ERM

**Combine the two concerns:** For a given $\lambda \geq 0$, find minimizer of

$$\hat{R}(\beta) + \lambda \|\beta\|^2_2$$

over $\beta \in \mathbb{R}^d$. 

Fact: If $\lambda > 0$, then the solution is always unique (even if $n < d$)!

$\lambda = 0$ is ERM / Ordinary Least Squares.

Parameter $\lambda$ controls how much attention is paid to the regularizer $\|\beta\|^2_2$ relative to the data fitting term $\hat{R}(\beta)$. Choose $\lambda$ using cross-validation.
Combine the two concerns: For a given $\lambda \geq 0$, find minimizer of

$$\hat{R}(\beta) + \lambda \|\beta\|_2^2$$

over $\beta \in \mathbb{R}^d$.

**Fact:** If $\lambda > 0$, then the solution is always unique (even if $n < d$)!
Regularized ERM

Combine the two concerns: For a given $\lambda \geq 0$, find minimizer of

$$\hat{\mathcal{R}}(\beta) + \lambda \|\beta\|^2_2$$

over $\beta \in \mathbb{R}^d$.

Fact: If $\lambda > 0$, then the solution is always unique (even if $n < d$)!

- This is called *ridge regression*.
  
  ($\lambda = 0$ is ERM / Ordinary Least Squares.)
Combine the two concerns: For a given $\lambda \geq 0$, find minimizer of
\[
\hat{R}(\beta) + \lambda \|\beta\|_2^2
\]
over $\beta \in \mathbb{R}^d$.

**Fact:** If $\lambda > 0$, then the solution is always unique (even if $n < d$)!

- This is called *ridge regression*.
  ($\lambda = 0$ is ERM / Ordinary Least Squares.)
- Parameter $\lambda$ controls how much attention is paid to the regularizer $\|\beta\|_2^2$ relative to the data fitting term $\hat{R}(\beta)$. 
Combine the two concerns: For a given $\lambda \geq 0$, find minimizer of

$$\hat{\mathcal{R}}(\beta) + \lambda \|\beta\|_2^2$$

over $\beta \in \mathbb{R}^d$.

**Fact:** If $\lambda > 0$, then the solution is always unique (even if $n < d$)!

- This is called *ridge regression*.
  ($\lambda = 0$ is ERM / Ordinary Least Squares.)

- Parameter $\lambda$ controls how much attention is paid to the *regularizer* $\|\beta\|_2^2$ relative to the *data fitting term* $\hat{\mathcal{R}}(\beta)$.

- Choose $\lambda$ using cross-validation.
Another interpretation of ridge regression

Define \((n + d) \times d\) matrix \(\tilde{A}\) and \((n + d) \times 1\) column vector \(\tilde{b}\) by

\[
\tilde{A} := \frac{1}{\sqrt{n}} \left[ \begin{array}{c}
\leftarrow x_1^T \\
\vdots \\
\leftarrow x_n^T \\
\sqrt{n\lambda} \\
\vdots \\
\sqrt{n\lambda}
\end{array} \right], \quad \tilde{b} := \frac{1}{\sqrt{n}} \left[ \begin{array}{c}
y_1 \\
\vdots \\
y_n \\
0 \\
\vdots \\
0
\end{array} \right].
\]
Another interpretation of ridge regression

Define \((n + d) \times d\) matrix \(\tilde{A}\) and \((n + d) \times 1\) column vector \(\tilde{b}\) by

\[
\tilde{A} := \frac{1}{\sqrt{n}} \begin{bmatrix}
\sqrt{n\lambda} & \ldots & \sqrt{n\lambda} \\
\vdots & \ddots & \vdots \\
\sqrt{n\lambda} & \ldots & \sqrt{n\lambda}
\end{bmatrix}, \quad \tilde{b} := \frac{1}{\sqrt{n}} \begin{bmatrix}
y_1 \\
\vdots \\
y_n \\
0
\end{bmatrix}.
\]

Then

\[
\|\tilde{A}\beta - \tilde{b}\|_2^2 = \hat{R}(\beta) + \lambda \|\beta\|_2^2.
\]
Another interpretation of ridge regression

Define \((n + d) \times d\) matrix \(\tilde{A}\) and \((n + d) \times 1\) column vector \(\tilde{b}\) by

\[
\tilde{A} := \frac{1}{\sqrt{n}} \left[ \begin{array}{c}
\leftarrow x_1^T \rightarrow \\
\vdots \\
\leftarrow x_n^T \rightarrow \\
\sqrt{n} \lambda \\
\sqrt{n} \lambda \\
\end{array} \right], \quad \tilde{b} := \frac{1}{\sqrt{n}} \left[ \begin{array}{c}
y_1 \\
\vdots \\
y_n \\
0 \\
0 \\
\end{array} \right].
\]

Then

\[
\|\tilde{A} \beta - \tilde{b}\|_2^2 = \hat{R}(\beta) + \lambda \|\beta\|_2^2.
\]

Interpretation:

- \(d\) “fake” data points; ensure that augmented data matrix \(\tilde{A}\) has rank \(d\).
Another interpretation of ridge regression

Define \((n + d) \times d\) matrix \(\tilde{A}\) and \((n + d) \times 1\) column vector \(\tilde{b}\) by

\[
\tilde{A} := \frac{1}{\sqrt{n}} \begin{bmatrix}
\sqrt{n\lambda} & \rightarrow & \sqrt{n\lambda} \\
\vdots & \rightarrow & \vdots \\
\sqrt{n\lambda} & \rightarrow & \sqrt{n\lambda}
\end{bmatrix}, \quad \tilde{b} := \frac{1}{\sqrt{n}} \begin{bmatrix}
y_1 \\
\vdots \\
y_n \\
0
\end{bmatrix}.
\]

Then

\[
\|\tilde{A}\beta - \tilde{b}\|_2^2 = \hat{R}(\beta) + \lambda \|\beta\|_2^2.
\]

**Interpretation:**

- \(d\) “fake” data points; ensure that augmented data matrix \(\tilde{A}\) has rank \(d\).
- Squared length of each “fake” feature vector is \(n\lambda\).
  - All corresponding labels are 0.
Another interpretation of ridge regression

Define \((n + d) \times d\) matrix \(\tilde{A}\) and \((n + d) \times 1\) column vector \(\tilde{b}\) by

\[
\tilde{A} := \frac{1}{\sqrt{n}} \begin{bmatrix}
\leftarrow x_1^T \rightarrow \\
\leftarrow x_n^T \rightarrow \\
\vdots \\
\sqrt{n\lambda}
\end{bmatrix}, \quad \tilde{b} := \frac{1}{\sqrt{n}} \begin{bmatrix}
y_1 \\
\vdots \\
y_n \\
0
\end{bmatrix}.
\]

Then

\[
\|\tilde{A}\beta - \tilde{b}\|_2^2 = \hat{R}(\beta) + \lambda\|\beta\|_2^2.
\]

Interpretation:

- \(d\) “fake” data points; ensure that augmented data matrix \(\tilde{A}\) has rank \(d\).
- Squared length of each “fake” feature vector is \(n\lambda\).
  All corresponding labels are 0.
- Prediction of \(\beta\) on \(i\)-th fake feature vector is \(\sqrt{n\lambda}\beta_i\).
**Lasso:** For a given $\lambda \geq 0$, find minimizer of

$$\hat{R}(\beta) + \lambda \|\beta\|_1$$

over $\beta \in \mathbb{R}^d$. Here, $\|v\|_1 = \sum_{i=1}^d |v_i|$ is the $\ell_1$-norm.
Regularization with a different norm

Lasso: For a given \( \lambda \geq 0 \), find minimizer of

\[
\hat{R}(\beta) + \lambda \|\beta\|_1
\]

over \( \beta \in \mathbb{R}^d \). Here, \( \|v\|_1 = \sum_{i=1}^d |v_i| \) is the \( \ell_1 \)-norm.

- Prefers shorter \( \beta \), but using a different notion of length than ridge.
**Lasso:** For a given $\lambda \geq 0$, find minimizer of

$$\hat{\mathcal{R}}(\beta) + \lambda \|\beta\|_1$$

over $\beta \in \mathbb{R}^d$. Here, $\|v\|_1 = \sum_{i=1}^d |v_i|$ is the $\ell_1$-norm.

- Prefers shorter $\beta$, but using a different notion of length than ridge.
- Tends to produce $\beta$ that are **sparse**—i.e., have few non-zero coordinates—or at least well-approximated by sparse vectors.
Regularization with a different norm

**Lasso**: For a given $\lambda \geq 0$, find minimizer of

$$
\widehat{R}(\beta) + \lambda \|\beta\|_1
$$

over $\beta \in \mathbb{R}^d$. Here, $\|v\|_1 = \sum_{i=1}^d |v_i|$ is the $\ell_1$-norm.

- Prefers shorter $\beta$, but using a different notion of length than ridge.
- Tends to produce $\beta$ that are *sparse*—i.e., have few non-zero coordinates—or at least well-approximated by sparse vectors.

**Fact**: Vectors with small $\ell_1$-norm are well-approximated by sparse vectors. If $\tilde{\beta}$ contains just the $1/\varepsilon^2$-largest coefficients (by magnitude) of $\beta$, then

$$
\|\beta - \tilde{\beta}\|_2 \leq \varepsilon \|\beta\|_1.
$$
Claim: If $\tilde{\beta}$ contains just the $T$-largest coefficients (by magnitude) of $\beta$, then

$$
\|\beta - \tilde{\beta}\|_2 \leq \frac{\|\beta\|_1}{\sqrt{T} + 1}.
$$
Claim: If $\tilde{\beta}$ contains just the $T$-largest coefficients (by magnitude) of $\beta$, then

$$\|\beta - \tilde{\beta}\|_2 \leq \frac{\|\beta\|_1}{\sqrt{T + 1}}.$$

WLOG $|\beta_1| \geq |\beta_2| \geq \cdots,$
Claim: If $\tilde{\beta}$ contains just the $T$-largest coefficients (by magnitude) of $\beta$, then

$$
\|\beta - \tilde{\beta}\|_2 \leq \frac{\|\beta\|_1}{\sqrt{T + 1}}.
$$

WLOG $|\beta_1| \geq |\beta_2| \geq \cdots$, so $\tilde{\beta} = (\beta_1, \ldots, \beta_T, 0, \ldots, 0)$. 

![Bar chart showing decreasing magnitudes of coefficients $|\beta_i|$ with index $i$.](chart.png)
Sparse approximations

**Claim:** If $\tilde{\beta}$ contains just the $T$-largest coefficients (by magnitude) of $\beta$, then

$$
\|\beta - \tilde{\beta}\|_2 \leq \frac{\|\beta\|_1}{\sqrt{T + 1}}.
$$

WLOG $|\beta_1| \geq |\beta_2| \geq \cdots$, so $\tilde{\beta} = (\beta_1, \ldots, \beta_T, 0, \ldots, 0)$.

$$
\|\beta - \tilde{\beta}\|_2^2 = \sum_{i \geq T+1} \beta_i^2
$$
**Claim:** If $\tilde{\beta}$ contains just the $T$-largest coefficients (by magnitude) of $\beta$, then

$$
\|\beta - \tilde{\beta}\|_2 \leq \frac{\|\beta\|_1}{\sqrt{T} + 1}.
$$

WLOG $|\beta_1| \geq |\beta_2| \geq \cdots$, so $\tilde{\beta} = (\beta_1, \ldots, \beta_T, 0, \ldots, 0)$.

$$
\|\beta - \tilde{\beta}\|_2^2 = \sum_{i \geq T+1} \beta_i^2 \\
\leq \sum_{i \geq T+1} |\beta_i| \cdot |\beta_{T+1}|
$$

This is a consequence of "mismatch" between $\ell_1$- and $\ell_2$-norms.

Can get similar results for other $\ell_p$ norms.
**Claim:** If $\tilde{\beta}$ contains just the $T$-largest coefficients (by magnitude) of $\beta$, then

$$\|\beta - \tilde{\beta}\|_2 \leq \frac{\|\beta\|_1}{\sqrt{T + 1}}.$$ 

WLOG $|\beta_1| \geq |\beta_2| \geq \cdots$, so $\tilde{\beta} = (\beta_1, \ldots, \beta_T, 0, \ldots, 0)$. 

$$\|\beta - \tilde{\beta}\|_2^2 = \sum_{i \geq T + 1} \beta_i^2 
\leq \sum_{i \geq T + 1} |\beta_i| \cdot |\beta_{T+1}| 
\leq \|\beta\|_1 \cdot |\beta_{T+1}|.$$ 

This is a consequence of “mismatch” between $\ell_1$- and $\ell_2$-norms. Can get similar results for other $\ell_p$ norms.
**Claim**: If $\tilde{\beta}$ contains just the $T$-largest coefficients (by magnitude) of $\beta$, then

$$
\|\beta - \tilde{\beta}\|_2 \leq \frac{\|\beta\|_1}{\sqrt{T + 1}}.
$$

WLOG $|\beta_1| \geq |\beta_2| \geq \cdots$, so $\tilde{\beta} = (\beta_1, \ldots, \beta_T, 0, \ldots, 0)$.

$$
\|\beta - \tilde{\beta}\|_2^2 = \sum_{i \geq T+1} \beta_i^2 \\
\leq \sum_{i \geq T+1} |\beta_i| \cdot |\beta_{T+1}| \\
\leq \|\beta\|_1 \cdot |\beta_{T+1}| \\
\leq \|\beta\|_1 \cdot \frac{\|\beta\|_1}{T + 1}.
$$
**Claim:** If $\tilde{\beta}$ contains just the $T$-largest coefficients (by magnitude) of $\beta$, then

$$\|\beta - \tilde{\beta}\|_2 \leq \frac{\|\beta\|_1}{\sqrt{T + 1}}.$$

WLOG $|\beta_1| \geq |\beta_2| \geq \cdots$, so $\tilde{\beta} = (\beta_1, \ldots, \beta_T, 0, \ldots, 0)$.

$$\|\beta - \tilde{\beta}\|_2^2 = \sum_{i \geq T + 1} \beta_i^2 \leq \sum_{i \geq T + 1} |\beta_i| \cdot |\beta_{T + 1}| \leq \|\beta\|_1 \cdot |\beta_{T + 1}| \leq \|\beta\|_1 \cdot \frac{\|\beta\|_1}{T + 1}.$$  

This is a consequence of “mismatch” between $\ell_1$- and $\ell_2$-norms.

Can get similar results for other $\ell_p$ norms.
Example: Coefficient profile ($\ell_2$ vs. $\ell_1$)

$Y =$ levels of prostate cancer antigen, $X =$ clinical measurements

**Horizontal axis:** varying $\lambda$ (large $\lambda$ to left, small $\lambda$ to right).

**Vertical axis:** coefficient value in $\ell_2$-regularized ERM and $\ell_1$-regularized ERM, for eight different variables.
Other approaches to sparse regression

- **Subset selection:**
  Find $\hat{\beta}$ that minimizes empirical risk among all vectors with at most $k$ non-zero entries.
Other approaches to sparse regression

- **Subset selection**: Find $\hat{\beta}$ that minimizes empirical risk among all vectors with at most $k$ non-zero entries.

  Unfortunately, this seems to require time exponential in $k$. 

Other approaches to sparse regression

- **Subset selection:**
  Find $\hat{\beta}$ that minimizes empirical risk among all vectors with at most $k$ non-zero entries.

  Unfortunately, this seems to require time exponential in $k$.

- **Greedy algorithms:**
  Repeatedly choose new variable to “include” in support of $\hat{\beta}$ until $k$ variables are included.
Other approaches to sparse regression

- **Subset selection:**
  Find $\hat{\beta}$ that minimizes empirical risk among all vectors with at most $k$ non-zero entries.

  Unfortunately, this seems to require time exponential in $k$.

- **Greedy algorithms:**
  Repeatedly choose new variable to “include” in support of $\hat{\beta}$ until $k$ variables are included.

  Forward stepwise regression / Orthogonal matching pursuit
Other approaches to sparse regression

- **Subset selection:**
  Find $\hat{\beta}$ that minimizes empirical risk among all vectors with at most $k$ non-zero entries.

  Unfortunately, this seems to require time exponential in $k$.

- **Greedy algorithms:**
  Repeatedly choose new variable to “include” in support of $\hat{\beta}$ until $k$ variables are included.

  **Forward stepwise regression / Orthogonal matching pursuit**

  Often works as well as $\ell_1$-regularized ERM.
Other approaches to sparse regression

- Subset selection:
  Find $\hat{\beta}$ that minimizes empirical risk among all vectors with at most $k$ non-zero entries.

  Unfortunately, this seems to require time exponential in $k$.

- Greedy algorithms:
  Repeatedly choose new variable to “include” in support of $\hat{\beta}$ until $k$ variables are included.

  Forward stepwise regression / Orthogonal matching pursuit

  Often works as well as $\ell_1$-regularized ERM.

Why do we care about sparsity?
Key takeaways

1. IID model for supervised learning.
2. Optimal predictors, linear regression models, and optimal linear predictors.
4. Risk of ERM; training risk vs. test risk; risk minimization vs. risk estimation.
5. Inductive bias, $\ell_1$- and $\ell_2$-regularization, sparsity.

Make sure you do the assigned reading, especially from the handouts!