Linear regression

COMS 4721 Spring 2018
1. Old Faithful and prediction functions
Prediction problem: Old Faithful geyser (Yellowstone)

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

Historical records of eruptions:

\[
\begin{array}{cccccccc}
  a_0 & b_0 & a_1 & b_1 & a_2 & b_2 & a_3 & b_3 & \ldots
\end{array}
\]

Data: \(Y_i := a_i - b_i - 1\), \(i = 1, \ldots, n\).

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 \(Y\).)
Statistical model for time between eruptions

Historical records of eruptions:

\[
\begin{array}{cccccccc}
  a_0 & b_0 & a_1 & b_1 & a_2 & b_2 & a_3 & b_3 \\
  \quad & \quad & \quad & \quad & \quad & \quad & \quad & \quad \\
\end{array}
\]

\[\xleftarrow{\text{Y_1}} \quad \xleftarrow{\text{Y_2}} \quad \xleftarrow{\text{Y_3}}\]

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

Data: \(Y_i := a_i - b_{i-1}, \ i = 1, \ldots, n\).
Statistical model for time between eruptions

Historical records of eruptions:

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

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

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

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:

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

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

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

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:

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

\[
\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \text{data}
\]

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

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

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 \( \hat{\mu} = 70.7941 \).
Statistical model for time between eruptions

Historical records of eruptions:

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

\[ \text{data} \]

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

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

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:

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

\[ \text{data} \]

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

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

**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 time until the next eruption seems to be related to duration of last eruption.
Using side-information

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

\[ \ldots a_{n-1} \ b_{n-1} \ a_n \ b_n \ \ldots \ \ | \ ]
\[ \text{data} \]

\[ \leftarrow X \rightarrow Y \]

1. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$, and choose a prediction function (a.k.a. predictor) $\hat{f} : X \rightarrow \mathbb{R}$.

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?
Using side-information

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

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

\[
\cdots \ X_n \leftrightarrow Y_n
\]

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

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

\[
\begin{array}{cccccccc}
\cdots & a_{n-1} & b_{n-1} & a_n & b_n & \cdots & t \\
\cdots & X_n & \dashv & Y_n & \dashv & \cdots & X & \dashv & Y
\end{array}
\]

**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.

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

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”.

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*.

$IID$ 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 $(\hat{f}(X) - Y)^2$. 

\[
\begin{array}{ccccccc}
\ldots & a_{n-1} & b_{n-1} & a_n & b_n & \ldots & t \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
X_n \quad \quad \quad \quad \quad \quad Y_n \\
\hline
X \quad \quad \quad \quad \quad \quad \quad \quad \quad Y \\
\end{array}
\]
Using side-information

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

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

\[
\begin{array}{c}
\cdots \\
X_n \rightarrow \ Y_n \leftarrow \\
\end{array}
\]

**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 $(\hat{f}(X) - Y)^2$.

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 (\textit{feature space}).

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

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

\( \mathcal{X} \): space of possible side-information (\textit{feature space}).
\( \mathcal{Y} \): space of possible outcomes (\textit{label space} or \textit{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. \textit{Marginal distribution} \( P_X \) of \( X \):
   \( P_X \) is a probability distribution on \( \mathcal{X} \).

2. \textit{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 \( \mathcal{Y} \).
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 \( \mathcal{Y} \).

**This lecture**: \( \mathcal{Y} = \mathbb{R} \) (regression problems).
What function $f : \mathcal{X} \to \mathbb{R}$ has smallest (squared loss) risk

$$\mathcal{R}(f) := \mathbb{E}[(f(X) - Y)^2]?$$
What function $f : \mathcal{X} \to \mathbb{R}$ has smallest (squared loss) \textit{risk}

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

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

$f^\star$ is called the \textit{regression function} or \textit{conditional mean function}. 

9 / 30
What function $f : \mathcal{X} \rightarrow \mathbb{R}$ has smallest (squared loss) risk

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

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

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

is the conditional mean $\mathbb{E}[Y \mid X = x]$.

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

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

has the smallest risk.

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

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

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

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

is the conditional mean $\mathbb{E}[Y \mid X = x]$.

- 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.

$f^*$ is called the regression function or conditional mean function.
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:

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

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

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

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

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

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

- **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.
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 linear regression model, such as the following:

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

- **Parameters:** \( \beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d, \sigma^2 > 0. \)
- \( \mathbf{X} = (X_1, \ldots, X_d) \), a \textit{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.
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:

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

- 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).
- 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 $f_\beta : \mathbb{R}^d \to \mathbb{R}$,

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

(We'll often refer to $f_\beta$ just by $\beta$.)
Enhancing linear regression models

Linear functions might sound rather restricted, but actually they can be quite powerful if you are creative about side-information.
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 $x \in \mathbb{R}$,

$$
\phi(x) = 2\sqrt{x + \frac{3}{8}}.
$$

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 $x \in \mathbb{R}$,

$$
\phi(x) = (1, \sin(2\pi x), \cos(2\pi x), \sin(4\pi x), \cos(4\pi x), \ldots).
$$

4. Polynomial expansion: for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$
\phi(x) = (1, x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, x_1 x_2, \ldots, x_1 x_d, \ldots, x_{d-1} x_d).
$$
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 \( x \in \mathbb{R} \),

\[
\phi(x) = 2\sqrt{x + \frac{3}{8}}.
\]

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

\[
\phi(x) = (x_1 \land x_5 \land \neg x_{10}) \lor (\neg x_2 \land x_7).
\]
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 $x \in \mathbb{R}$,

   \[ \phi(x) = 2\sqrt{x + \frac{3}{8}}. \]

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 $x \in \mathbb{R}$,

   \[ \phi(x) = (1, \sin(2\pi x), \cos(2\pi x), \sin(4\pi x), \cos(4\pi x), \ldots). \]
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 \( x \in \mathbb{R} \),

   \[
   \phi(x) = 2\sqrt{x + \frac{3}{8}}.
   \]

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 \( x \in \mathbb{R} \),

   \[
   \phi(x) = (1, \sin(2\pi x), \cos(2\pi x), \sin(4\pi x), \cos(4\pi x), \ldots).
   \]

4. Polynomial expansion: for \( x = (x_1, \ldots, x_d) \in \mathbb{R}^d \),

   \[
   \phi(x) = (1, x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, x_1 x_2, \ldots, x_1 x_d, \ldots, x_{d-1} x_d).
   \]
Example: quadratic expansion

Quadratic function \( f: \mathbb{R} \to \mathbb{R} \)

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

for \( a, b, c \in \mathbb{R} \).
Example: 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 $\beta = (c, b, a)$. 

Example: quadratic expansion

Quadratic function \( f : \mathbb{R} \rightarrow \mathbb{R} \)

\[
f(x) = ax^2 + bx + c, \quad x \in \mathbb{R},
\]
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 \( f : \mathbb{R}^d \rightarrow \mathbb{R} \), 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:

\[ \phi(x) := (1, x). \]
Example: 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
Basic linear regression model: \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, with

\[ Y \mid X = x \sim N(x^T \beta, \sigma^2), \quad x \in \mathbb{R}^d. \]
Basic linear regression model: \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, with

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

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

\[
\sum_{i=1}^{n} \left\{ -\frac{1}{2\sigma^2} (x_i^T \beta - y_i)^2 + \frac{1}{2} \ln \frac{1}{2\pi\sigma^2} \right\} + \left\{ \text{terms not involving } (\beta, \sigma^2) \right\} .
\]
Basic linear regression model: \( (X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \) are iid, with

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

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

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

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

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2.
\]
MLE for basic linear regression model

Basic linear regression model: \((X_1, Y_1), \ldots,(X_n, Y_n), (X, Y)\) are iid, with

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

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

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

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

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

This coincides with another approach, called empirical risk minimization . . .
Empirical distribution $P_n$ on $(x_1, y_1), \ldots, (x_n, y_n)$ has probability mass function $p_n$ given by

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

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

Empirical (squared loss) risk $\hat{R}(f)$ of $f: \mathbb{R}^d \to \mathbb{R}$ is, for $(X', Y') \sim P_n$,

$$\hat{R}(f) := \mathbb{E}[(f(X') - Y')^2] = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2.$$
Empirical distribution $P_n$ on $(x_1, y_1), \ldots, (x_n, y_n)$ has probability mass function $p_n$ given by

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

Empirical (squared loss) risk $\hat{R}(f)$ of $f: \mathbb{R}^d \to \mathbb{R}$ is, for $(X', Y') \sim P_n$,

$$\hat{R}(f) := \mathbb{E}[(f(X') - Y')^2] = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2.$$ 

Empirical risk minimization (ERM) is the learning method that returns a function (from a specified function class) that minimizes the empirical risk.
Empirical distribution $P_n$ on $(x_1, y_1), \ldots, (x_n, y_n)$ has probability mass function $p_n$ given by

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

Empirical (squared loss) risk $\hat{\mathcal{R}}(f)$ of $f : \mathbb{R}^d \to \mathbb{R}$ is, for $(X', Y') \sim P_n$,

$$\hat{\mathcal{R}}(f) := \mathbb{E}[(f(X') - Y')^2] = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2.$$

Empirical risk minimization (ERM) is the learning method that returns a function (from a specified function class) that minimizes the empirical risk.

For linear functions: ERM returns

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

which coincides with MLE under the basic linear regression model.
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

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

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

\[
A := \frac{1}{\sqrt{n}} \begin{bmatrix}
    ← x_1^\top → \\
    ⋮ \\
    ← x_n^\top →
\end{bmatrix}, \quad b := \frac{1}{\sqrt{n}} \begin{bmatrix}
y_1 \\
    ⋮ \\
y_n
\end{bmatrix}.
\]

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

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

Can write empirical risk as

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

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

$$\nabla \hat{R}(\beta) = 0, \quad \text{i.e., } \beta \text{ is a critical point of } \hat{R}. $$
Empirical risk minimization in matrix notation

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

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

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

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

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

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

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

\[
\mathbf{A} = \begin{bmatrix}
\mathbf{a}_1 & \cdots & \mathbf{a}_d \\
\uparrow & \cdots & \uparrow \\
\downarrow & \cdots & \downarrow
\end{bmatrix}.
\]
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

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

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

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

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

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

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

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

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.
- But \( >1 \) way to write \( \hat{b} \) as linear combination of \( a_1, \ldots, a_d \) iff \( \text{rank}(A) < d \).
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

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

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

If $\text{rank}(A) < d$, then ERM solution is not unique.
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 \(R(\beta) = \mathbb{E}[(X^T\beta - Y)^2]\)?
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 \(R(\beta) = \mathbb{E}[(X^T \beta - Y)^2]\)?

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

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

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

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?
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?

\( A^T A = \mathbb{E}[\tilde{X} \tilde{X}^T] \) and \( A^T b = \mathbb{E}[\tilde{Y} \tilde{X}] \) for \( (\tilde{X}, \tilde{Y}) \sim P_n \).
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 \(R(\beta) = \mathbb{E}[(X^T \beta - Y)^2]\)?

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 \(R\).

Looks familiar?

\[
A^T A = \mathbb{E}[\tilde{X} \tilde{X}^T] \text{ and } A^T b = \mathbb{E}[\tilde{Y} \tilde{X}] \text{ for } (\tilde{X}, \tilde{Y}) \sim P_n.
\]

If \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, then as \(n \to \infty\),

\[
P_n \to P,
A^T A \to \mathbb{E}[XX^T],
A^T b \to \mathbb{E}[YX].
\]
4. Risk, empirical risk, and estimating risk
Statistical theory explains “optimality” of MLE in basic linear regression model.
Statistical theory explains “optimality” of MLE in basic linear regression model. What about in iid model where regression function is not necessarily linear?
Risk of ERM

Statistical theory explains “optimality” of MLE in basic linear regression model. What about in iid model where regression function is not necessarily linear?

Let $\beta^*$ be a minimizer of $R$ over all $\beta \in \mathbb{R}^d$.

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

$\Pr$actical implication: Be cautious when $n$ is not much larger than $d$. 
Risk of ERM

Statistical theory explains “optimality” of MLE in basic linear regression model. What about in iid model where regression function is not necessarily linear?

Let $\beta^*$ be a minimizer of $R$ over all $\beta \in \mathbb{R}^d$.

- If ERM solution $\hat{\beta}$ is not unique (e.g., if $n < d$), then $R(\hat{\beta})$ can be arbitrarily worse than $R(\beta^*)$.
- What about when ERM solution is unique?
Statistical theory explains “optimality” of MLE in basic linear regression model.

What about in iid model where regression function is not necessarily linear?

Let $\beta^*$ be a minimizer of $\mathcal{R}$ over all $\beta \in \mathbb{R}^d$.

- 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?

$$
\mathbb{E} \mathcal{R}(\hat{\beta}) = \mathcal{R}(\beta^*) + \mathbb{E}[(X^T(\hat{\beta} - \beta^*))^2].
$$

Under some assumptions, second term $\to 0$ as $n \to \infty$ at rate $\frac{d}{n}$. 

Practical implication: Be cautious when $n$ is not much larger than $d$. 
Statistical theory explains “optimality” of MLE in basic linear regression model.

What about in iid model where regression function is not necessarily linear? Let $\beta^*$ be a minimizer of $\mathcal{R}$ over all $\beta \in \mathbb{R}^d$.

- 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?

\[
\mathbb{E} \mathcal{R}(\hat{\beta}) = \mathcal{R}(\beta^*) + \mathbb{E}[X^T(\hat{\beta} - \beta^*))^2].
\]

Under some assumptions, second term $\to 0$ as $n \to \infty$ at rate $\frac{d}{n}$.

- **Practical implication**: Be cautious when $n$ is not much larger than $d$. 
For ERM $\hat{\beta}$, we generally have

$$\mathbb{E} \hat{\mathcal{R}}(\hat{\beta}) \leq \mathbb{E} \mathcal{R}(\hat{\beta}).$$
For ERM $\hat{\beta}$, we generally have

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

*Overfitting*: empirical risk is “small”, but (true) risk is “much higher”. (Hard to make this very precise.)
Overfitting example

\((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) are iid, with \(X\) a continuous random variable.
(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) are iid, with X a continuous random variable. 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, with \(X\) a continuous random variable. 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 \(k\).
Overfitting example

\[(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\] are iid, with \(X\) a continuous random variable. 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 true risk of any function.
Estimating risk

If $\hat{R}(\hat{f})$ is not (necessarily) a good estimate of $R(\hat{f})$, then what is?

Use empirical risk on observations not used to learn $\hat{f}$.

IID model:

$\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \sim \text{iid} P$.

$\rightarrow$ training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ used to learn $\hat{f}$.

$\rightarrow$ test data $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m)$ used to estimate risk, via test risk $\hat{R}_{\text{test}}(\hat{f}) := \frac{1}{m} \sum_{i=1}^{m} (\hat{f}(\tilde{X}_i) - \tilde{Y}_i)^2$.

$\rightarrow$ Training data is independent of test data, so $\hat{f}$ is independent of test data.

$\rightarrow$ Let $L_i := (\hat{f}(\tilde{X}_i) - \tilde{Y}_i)^2$ for each $i = 1, \ldots, m$, so $E[\hat{R}_{\text{test}}(\hat{f}) | \hat{f}] = \frac{1}{m} \sum_{i=1}^{m} E[L_i | \hat{f}] = R(\hat{f})$.

Moreover, $L_1, \ldots, L_m$ are conditionally iid given $\hat{f}$, and hence by Law of Large Numbers, $\hat{R}_{\text{test}}(\hat{f}) \to R(\hat{f})$ as $m \to \infty$. 
Estimating risk

If $\hat{\mathcal{R}}(\hat{f})$ is not (necessarily) a good estimate of $\mathcal{R}(\hat{f})$, then what is?

Use empirical risk on observations not used to learn $\hat{f}$. 

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

If $\hat{R}(\hat{f})$ is not (necessarily) a good estimate of $R(\hat{f})$, then what is?

Use empirical risk on observations not used to learn $\hat{f}$.

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

If $\hat{R}(\hat{f})$ is not (necessarily) a good estimate of $\mathcal{R}(f)$, then what is?

Use empirical risk on observations not used to learn $\hat{f}$.

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 $\hat{f}$.
Estimating risk

If $\hat{R}(\hat{f})$ is not (necessarily) a good estimate of $R(\hat{f})$, then what is?

Use empirical risk on observations not used to learn $\hat{f}$.

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

If $\hat{\mathcal{R}}(\hat{f})$ is not (necessarily) a good estimate of $\mathcal{R}(\hat{f})$, then what is?

Use empirical risk on observations not used to learn $\hat{f}$.

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

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

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

If $\hat{R}(\hat{f})$ is not (necessarily) a good estimate of $R(\hat{f})$, then what is?

Use empirical risk on observations not used to learn $\hat{f}$.

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 $L_i := (f(\tilde{X}_i) - \tilde{Y}_i)^2$ for each $i = 1, \ldots, m$, so

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

If \( \hat{R}(\hat{f}) \) is not (necessarily) a good estimate of \( R(\hat{f}) \), then what is?

Use empirical risk on observations not used to learn \( \hat{f} \).

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**

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

- 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

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

- Moreover, \( L_1, \ldots, L_m \) are conditionally iid given \( \hat{f} \), and hence by Law of Large Numbers, \( \hat{R}_{test}(\hat{f}) \to R(\hat{f}) \) as \( m \to \infty \).
Why doesn’t argument about $\hat{R}_{\text{test}}$ also apply to $\hat{R}$ (based on training data)?
Training risk vs test risk

Why doesn’t argument about $\hat{R}_{\text{test}}$ also apply to $\hat{R}$ (based on training data)?

$(\hat{f}(X_i) - Y_i)^2$ for $i = 1, \ldots, n$ are not (conditionally) independent given $\hat{f}$, since $\hat{f}$ depends on $((X_i, Y_i))_{i=1}^n$. 
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.)

\[ \sqrt{35.9} \approx 3.5 \]

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

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

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

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, $\sqrt{35.9} > \text{mean duration} \approx 3.5$.)
Key takeaways

1. IID model for supervised learning.
2. Optimal predictors, linear regression models, and optimal linear predictors.
4. Estimating risk from test data.