Exam 1 review

COMS 4771 Fall 2018
Agenda

- Quickly review big ideas from the class
- Practice problem 8a
- Homework 2 problem 2
- Practice problem 3
- Practice problem 5c
Big ideas from the class

- Statistical model for predictions and prediction functions
- Optimal predictions and optimal prediction functions
- The plug-in principle
- Risk and empirical risk
- Decision boundaries
- Linear functions and feature expansions
- Inductive bias, regularization
- Aggregation (e.g., model averaging, online-to-batch)
- Using mathematical optimization to formulate learning methods
Practice problem 8a
Consider the following fixed design setting for linear regression, where $x_1, \ldots, x_n \in \mathbb{R}$ are fixed real numbers (not random variables), and $Y_1, \ldots, Y_n$ are random variables.
Consider the following fixed design setting for linear regression, where $x_1, \ldots, x_n \in \mathbb{R}$ are fixed real numbers (not random variables), and $Y_1, \ldots, Y_n$ are random variables.

Let $\beta^* \in \mathbb{R}$ be the minimizer of the following function on $\mathbb{R}$:

$$J(\beta) := \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (x_i \beta - Y_i)^2 \right].$$

Let $\hat{\beta}$ be the minimizer of the following function on $\mathbb{R}$:

$$\hat{J}(\beta) := \frac{1}{n} \sum_{i=1}^{n} (x_i \beta - Y_i)^2.$$
Problem statement

Consider the following fixed design setting for linear regression, where \( x_1, \ldots, x_n \in \mathbb{R} \) are fixed real numbers (not random variables), and \( Y_1, \ldots, Y_n \) are random variables.

Let \( \beta^\star \in \mathbb{R} \) be the minimizer of the following function on \( \mathbb{R} \):

\[
J(\beta) := \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (x_i\beta - Y_i)^2 \right].
\]

Let \( \hat{\beta} \) be the minimizer of the following function on \( \mathbb{R} \):

\[
\hat{J}(\beta) := \frac{1}{n} \sum_{i=1}^{n} (x_i\beta - Y_i)^2.
\]

**True or false:** \( \mathbb{E}[\hat{\beta}] = \beta^\star \) for all \( n \geq 1 \).
Fixed-design risk

By linearity of expectation:

\[ J(\beta) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (x_i \beta - Y_i)^2 \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} (x_i \beta - Y_i)^2 \]
Fixed-design risk

By linearity of expectation:

\[ J(\beta) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (x_i \beta - Y_i)^2 \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(x_i \beta - Y_i)^2 \]

Bias-variance decomposition:

\[ \mathbb{E}[(x_i \beta - Y_i)^2] = (x_i \beta - \mathbb{E}(Y_i))^2 + \text{var}(Y_i). \]
Fixed-design risk

By linearity of expectation:

\[ J(\beta) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (x_i \beta - Y_i)^2 \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(x_i \beta - Y_i)^2 \]

Bias-variance decomposition:

\[ \mathbb{E}[(x_i \beta - Y_i)^2] = (x_i \beta - \mathbb{E}(Y_i))^2 + \text{var}(Y_i). \]

Therefore

\[ J(\beta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(x_i \beta - Y_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i \beta - \mathbb{E}(Y_i))^2 + \frac{1}{n} \sum_{i=1}^{n} \text{var}(Y_i), \]

which is fixed-design risk of \( \beta \), plus term that doesn’t depend on \( \beta \).
(Empirical) risk minimizer

Minimizer $\beta^*$ of $J$ satisfies

$$
\left(\frac{1}{n} \sum_{i=1}^{n} x_i^2\right) \beta^* = \left(\frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{E}(Y_i)\right).
$$
Minimizer $\beta^*$ of $J$ satisfies

$$
\left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) \beta^* = \left(\frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{E}(Y_i) \right).
$$

Empirical risk $\hat{J}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i \beta - Y_i)^2$ has minimizer $\hat{\beta}$ that satisfies

$$
\left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) \hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i Y_i \right).
$$
(Empirical) risk minimizer

Minimizer $\beta^\star$ of $J$ satisfies

$$
\left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) \beta^\star = \left( \frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{E}(Y_i) \right).
$$

Empirical risk $\tilde{J}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i \beta - Y_i)^2$ has minimizer $\hat{\beta}$ that satisfies

$$
\left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) \hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^{n} x_i Y_i \right).
$$

As long as not all $x_i = 0$, can divide both sides by $\frac{1}{n} \sum_{i=1}^{n} x_i^2 \ldots$

$$
\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} x_i Y_i \right) \right] = \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{E}(Y_i) \right).
$$
Homework 2 problem 2
High-level problem

There are $n$ independent random variables $Y_1, \ldots, Y_n$. Let

$$\mu_i := \mathbb{E}(Y_i) \quad \text{for all } i = 1, \ldots, n.$$  

You observe $Y = (Y_1, \ldots, Y_n)$. Goal: estimate $\mu := (\mu_1, \ldots, \mu_n)$. 
There are $n$ independent random variables $Y_1, \ldots, Y_n$. Let

$$
\mu_i := \mathbb{E}(Y_i) \quad \text{for all } i = 1, \ldots, n.
$$

You observe $Y = (Y_1, \ldots, Y_n)$. Goal: estimate $\mu := (\mu_1, \ldots, \mu_n)$.

You have real numbers $z_1, \ldots, z_n \in [-\pi, \pi]$, and you are willing to assume that $\mu_i = h(z_i)$ for some “nice” function $h: \mathbb{R} \to \mathbb{R}$. 

High-level problem

There are $n$ independent random variables $Y_1, \ldots, Y_n$. Let

$$
\mu_i := \mathbb{E}(Y_i) \quad \text{for all } i = 1, \ldots, n.
$$

You observe $Y = (Y_1, \ldots, Y_n)$. Goal: estimate $\mu := (\mu_1, \ldots, \mu_n)$.

You have real numbers $z_1, \ldots, z_n \in [-\pi, \pi]$, and you are willing to assume that $\mu_i = h(z_i)$ for some “nice” function $h: \mathbb{R} \to \mathbb{R}$.

**Strategy.** Choose a feature expansion $x_i := \phi(z_i) \in \mathbb{R}^d$ and estimate $\mu$ via ordinary least squares:

$$
\hat{\mu}_i := \phi(z_i)^T \hat{\beta} \quad \text{for all } i = 1, \ldots, n
$$

where

$$
\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\phi(z_i)^T \beta - Y_i)^2.
$$
What is a good feature expansion?

Weierstrass approximation theorem. Any continuous function on a closed interval can be well-approximated by a polynomial.
What is a good feature expansion?

**Weierstrass approximation theorem.** Any continuous function on a closed interval can be well-approximated by a polynomial.

Ok, so let’s use polynomial feature expansion

\[
\phi(z_i) := (1, z_i, z_i^2, \ldots, z_i^k) \in \mathbb{R}^{k+1} \quad \text{for all } i = 1, \ldots, n.
\]
What is a good feature expansion?

**Weierstrass approximation theorem.** Any continuous function on a closed interval can be well-approximated by a polynomial.

Ok, so let's use polynomial feature expansion

$$\phi(z_i) := (1, z_i, z_i^2, \ldots, z_i^k) \in \mathbb{R}^{k+1} \quad \text{for all } i = 1, \ldots, n.$$  

**Question:** How well does this work?

Fixed-design analysis tells us

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2 \right] = \min_{\beta \in \mathbb{R}^{k+1}} \frac{1}{n} \sum_{i=1}^{n} (\phi(z_i)^\top \beta - \mu_i)^2 + \frac{k + 1}{n}.$$
What is a good feature expansion?

**Weierstrass approximation theorem.** Any continuous function on a closed interval can be well-approximated by a polynomial.

Ok, so let’s use polynomial feature expansion

\[
\phi(z_i) := (1, z_i, z_i^2, \ldots, z_i^k) \in \mathbb{R}^{k+1} \quad \text{for all } i = 1, \ldots, n.
\]

**Question:** How well does this work?

Fixed-design analysis tells us

\[
\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2 \right] = \min_{\beta \in \mathbb{R}^{k+1}} \frac{1}{n} \sum_{i=1}^{n} (\phi(z_i)^T \beta - \mu_i)^2 + \frac{k+1}{n}.
\]

Second term \(\frac{k+1}{n}\) is easy to understand; what about the first term?
Taylor’s theorem. For any \( k + 1 \)-times differentiable function \( h: \mathbb{R} \rightarrow \mathbb{R} \) and any \( z \in \mathbb{R} \), there exists \( \xi \) between 0 and \( z \) such that

\[
h(z) = h(0) + \sum_{\ell=1}^{k} \frac{1}{\ell!} h^{(\ell)}(0) z^\ell + \frac{1}{(k + 1)!} h^{(k+1)}(\xi) z^{k+1}.
\]

- **k-th order Taylor approx.**
- **remainder term**
**Taylor’s theorem.** For any $k + 1$-times differentiable function $h: \mathbb{R} \to \mathbb{R}$ and any $z \in \mathbb{R}$, there exists $\xi$ between $0$ and $z$ such that

$$h(z) = h(0) + \sum_{\ell=1}^{k} \frac{1}{\ell!} h^{(\ell)}(0)z^\ell + \frac{1}{(k + 1)!} h^{(k+1)}(\xi)z^{k+1}.$$  

$k$-th order Taylor approx.  

remainder term

Our $h = \cos$ satisfies preconditions, so there exists $\xi_1, \ldots, \xi_n \in \mathbb{R}$ such that

$$h(z_i) - \left( h(0) + \sum_{\ell=1}^{k} \frac{1}{\ell!} h^{(\ell)}(0)z_i^\ell \right) = \frac{1}{(k + 1)!} h^{(k+1)}(\xi_i)z_i^{k+1}$$

for each $i = 1, \ldots, n.$
Taylor’s theorem. For any $k + 1$-times differentiable function $h: \mathbb{R} \to \mathbb{R}$ and any $z \in \mathbb{R}$, there exists $\xi$ between 0 and $z$ such that

$$h(z) = h(0) + \sum_{\ell=1}^{k} \frac{1}{\ell!} h^{(\ell)}(0) z^{\ell} + \frac{1}{(k + 1)!} h^{(k+1)}(\xi) z^{k+1}.$$ 

$k$-th order Taylor approx. 
rem

Our $h = \cos$ satisfies preconditions, so there exists $\xi_1, \ldots, \xi_n \in \mathbb{R}$ such that

$$h(z_i) - \left( h(0) + \sum_{\ell=1}^{k} \frac{1}{\ell!} h^{(\ell)}(0) z_i^{\ell} \right) = \frac{1}{(k + 1)!} h^{(k+1)}(\xi_i) z_i^{k+1}$$

for each $i = 1, \ldots, n$.

Parenthesized terms are $\phi(z_i)^T \tilde{\beta}$ for

$$\tilde{\beta} := (h(0), h^{(1)}(0), \frac{1}{2} h^{(2)}(0), \ldots, \frac{1}{k!} h^{(k)}(0))$$.
There exists $\xi_1, \ldots, \xi_n \in \mathbb{R}$ such that

$$\min_{\beta \in \mathbb{R}^{k+1}} \frac{1}{n} \sum_{i=1}^{n} (\phi(z_i)^T \beta - \mu_i)^2 \leq \frac{1}{n} \sum_{i=1}^{n} (\phi(z_i)^T \tilde{\beta} - \mu_i)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{(k + 1)!} h^{(k+1)}(\xi_i) z_i^{k+1} \right)^2.$$

How large is RHS?
There exists $\xi_1, \ldots, \xi_n \in \mathbb{R}$ such that

\[
\min_{\beta \in \mathbb{R}^{k+1}} \frac{1}{n} \sum_{i=1}^{n} (\phi(z_i)^T \beta - \mu_i)^2 \leq \frac{1}{n} \sum_{i=1}^{n} (\phi(z_i)^T \tilde{\beta} - \mu_i)^2
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{(k+1)!} h^{(k+1)}(\xi_i) z_i^{k+1} \right)^2.
\]

How large is RHS?

- Derivatives of $\cos$ are $-\sin$, $-\cos$, $\sin$, etc., which are all bounded in absolute value by 1.
- $z_i \in [-\pi, \pi]$, so $|z_i|^{k+1} \leq \pi^{k+1}$.
- Therefore, RHS is at most

\[
\left( \frac{\pi^{k+1}}{(k+1)!} \right)^2.
\]
For $h = \cos$, 

$$
\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2 \right] \leq \left( \frac{\pi^{k+1}}{(k + 1)!} \right)^2 + \frac{k + 1}{n}.
$$
For $h = \cos$,

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2 \right] \leq \left( \frac{\pi^{k+1}}{(k+1)!} \right)^2 + \frac{k + 1}{n}.$$

**Question:** For what $k$ is this bound as small as possible?

- First term gets smaller with $k$.
  - Not entirely obvious, since numerator is exponential in $k$. But it is true because denominator grows *even faster* with $k$.

- Second term gets larger with $k$. 
Conclusion

For $h = \cos$,

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2 \right] \leq \left( \frac{\pi^{k+1}}{(k + 1)!} \right)^2 + \frac{k + 1}{n}.$$

**Question:** For what $k$ is this bound as small as possible?

- First term gets smaller with $k$.
  - Not entirely obvious, since numerator is exponential in $k$. But it is true because denominator grows even faster with $k$.
- Second term gets larger with $k$.

**Easier:** How to set $k$ as a function of $n$ so that resulting bound decreases as fast as possible with $n$ (ignoring constant factors)?
Replace $k + 1$ with $k(n)$, and let

$$g(n) := \underbrace{\pi^{2k(n)}}_{a(n)} + \underbrace{\frac{k(n)}{n}}_{b(n)}.$$
Optimizing the bound (1)

Replace $k + 1$ with $k(n)$, and let

$$g(n) := \frac{\pi^{2k(n)}}{k(n)!^2} + \frac{k(n)}{n}.$$

Observation 1: $k(n)$ should be increasing with $n$. 
Optimizing the bound (1)

Replace $k + 1$ with $k(n)$, and let

$$g(n) := \frac{\pi^{2k(n)}}{k(n)!^2} + \frac{k(n)}{n}. \underline{a(n)} \underline{b(n)}$$

Observation 1: $k(n)$ should be increasing with $n$.

Observation 2:

- If $a(n) = \Omega(b(n))$, then $g(n) = \Theta(a(n))$.
- If $a(n) = O(b(n))$, then $g(n) = \Theta(b(n))$. 
Optimizing the bound (1)

Replace $k + 1$ with $k(n)$, and let

$$g(n) := \frac{\pi^{2k(n)}}{k(n)!^2} + \frac{k(n)}{n a(n) b(n)}.$$  

Observation 1: $k(n)$ should be increasing with $n$.

Observation 2:

- If $a(n) = \Omega(b(n))$, then $g(n) = \Theta(a(n))$.
- If $a(n) = O(b(n))$, then $g(n) = \Theta(b(n))$.

Therefore, choose $k(n)$, growing with $n$, so that $a(n) = \Theta(b(n))$.  

For convenience, we'll solve

$$-\frac{1}{2} \ln a(n) = -\frac{1}{2} \ln b(n).$$
Replace $k + 1$ with $k(n)$, and let

$$g(n) := \frac{\pi^{2k(n)}}{k(n)!^2} + \frac{k(n)}{n}.$$

Observation 1: $k(n)$ should be increasing with $n$.

Observation 2:

- If $a(n) = \Omega(b(n))$, then $g(n) = \Theta(a(n))$.
- If $a(n) = O(b(n))$, then $g(n) = \Theta(b(n))$.

Therefore, choose $k(n)$, growing with $n$, so that $a(n) = \Theta(b(n))$.

For convenience, we’ll solve $-\frac{1}{2} \ln a(n) = -\frac{1}{2} \ln b(n)$. 
Optimizing the bound (2)

Target equation \(-\frac{1}{2} \ln a(n) = -\frac{1}{2} \ln b(n)\) is equivalent to

\[\ln k(n)! - k(n) \ln \pi = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n).\]
Optimizing the bound (2)

Target equation $-\frac{1}{2} \ln a(n) = -\frac{1}{2} \ln b(n)$ is equivalent to

$$\ln k(n)! - k(n) \ln \pi = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n).$$

Stirling’s approximation:

$$\ln k(n)! = k(n) \ln k(n) - k(n) + O(\log k(n)).$$
Optimizing the bound (2)

Target equation \(-\frac{1}{2} \ln a(n) = -\frac{1}{2} \ln b(n)\) is equivalent to

\[\ln k(n)! - k(n) \ln \pi = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n).\]

Stirling’s approximation:

\[\ln k(n)! = k(n) \ln k(n) - k(n) + O(\log k(n)).\]

Plugging-in:

\[k(n) \ln k(n) - \left(1 + \ln \pi - O \left(\frac{\log k(n)}{k(n)}\right)\right) k(n) = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n).\]
Optimizing the bound (2)

Target equation $-\frac{1}{2} \ln a(n) = -\frac{1}{2} \ln b(n)$ is equivalent to

$$\ln k(n)! - k(n) \ln \pi = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n).$$

Stirling’s approximation:

$$\ln k(n)! = k(n) \ln k(n) - k(n) + O(\log k(n)).$$

Plugging-in:

$$k(n) \ln k(n) - \left(1 + \ln \pi - O\left(\frac{\log k(n)}{k(n)}\right)\right) k(n) = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n).$$

Call parenthesized term $C(n)$, and note that it converges to a constant $C := 1 + \ln \pi$ as $n \to \infty$ (since $k(n) \to \infty$ as $n \to \infty$).
New target equation:

\[ k(n) \ln k(n) - Ck(n) = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n). \]
Optimizing the bound (3)

New target equation:

\[ k(n) \ln k(n) - Ck(n) = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n). \]

Try \( k(n) := \frac{1}{2} \ln(n). \)
Optimizing the bound (3)

New target equation:

\[ k(n) \ln k(n) - Ck(n) = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n). \]

Try \( k(n) := \frac{1}{2} \ln(n) \).

Then

- LHS is \( \Omega(\log(n) \cdot \log \log(n)) \).
- RHS is \( O(\log n) \).

Too big!
Optimizing the bound (4)

New target equation:

\[ k(n) \ln k(n) - Ck(n) = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n). \]
Optimizing the bound (4)

New target equation:

\[ k(n) \ln k(n) - Ck(n) = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n). \]

Try \( k(n) := \frac{\frac{1}{2} \ln(n)}{\ln(\frac{1}{2} \ln n)}. \)
Optimizing the bound (4)

New target equation:

\[ k(n) \ln k(n) - Ck(n) = \frac{1}{2} \ln n - \frac{1}{2} \ln k(n). \]

Try \( k(n) := \frac{\frac{1}{2} \ln(n)}{\ln(\frac{1}{2} \ln n)} \).

Then

- LHS is \( \frac{1}{2} \ln(n) \cdot \left(1 - O\left(\frac{\log \log \log n}{\log \log n}\right)\right) \).
- RHS is \( \frac{1}{2} \ln(n) \cdot \left(1 - O\left(\frac{\log \log n}{\log n}\right)\right) \).

Just right!
Conclusion

Using $k(n) = \frac{1}{2} \frac{\ln(n)}{\ln(\frac{1}{2} \ln n)}$ gives the bound

$$O \left( \frac{\log n}{n \log \log n} \right).$$
Practice problem 3
Let \((X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)\) be iid random examples from a probability distribution \(P\) over \(\mathbb{R}^d \times \mathbb{R}\). Let \(\mathcal{R}\) denote the true (squared loss) risk with respect to \(P\), and let \(\mathcal{R}_n\) denote the empirical (squared loss) risk based on \((X_1, Y_1), \ldots, (X_n, Y_n)\).

Let \(\hat{\beta}\) be a minimizer of \(\mathcal{R}_n\). Prove that

\[
\mathbb{E} \left[ \mathcal{R}_n(\hat{\beta}) \right] \leq \mathbb{E} \left[ \mathcal{R}(\hat{\beta}) \right].
\]
Main idea

\( \hat{\beta} \) depends on \((X_1, Y_1), \ldots, (X_n, Y_n)\), and hence is random.
Main idea

\( \hat{\beta} \) depends on \((X_1, Y_1), \ldots, (X_n, Y_n)\), and hence is random.

So \( \mathcal{R}(\hat{\beta}) = \mathbb{E}[(X^T \hat{\beta} - Y)^2 | \hat{\beta}] \) is a random variable, with expectation

\[
\mathbb{E} \left[ \mathbb{E}[(X^T \hat{\beta} - Y)^2 | \hat{\beta}] \right] = \mathbb{E} \left[ \mathcal{R}(\hat{\beta}) \right].
\]

(Here, we explicitly show conditioning on \( \hat{\beta} \).)
\( \hat{\beta} \) depends on \((X_1, Y_1), \ldots, (X_n, Y_n)\), and hence is random.

So \( \mathcal{R}(\hat{\beta}) = \mathbb{E}[(X^T \hat{\beta} - Y)^2 \mid \hat{\beta}] \) is a random variable, with expectation

\[
\mathbb{E} \left[ \mathbb{E}[(X^T \hat{\beta} - Y)^2 \mid \hat{\beta}] \right] = \mathbb{E} \left[ \mathcal{R}(\hat{\beta}) \right].
\]

(Here, we explicitly show conditioning on \( \hat{\beta} \).)

**Key observation:** Inside the (conditional) expectation is “test risk” on a single example \((X, Y)\), which is indepedendent of \( \hat{\beta} \).
Main idea

\( \hat{\beta} \) depends on \((X_1, Y_1), \ldots, (X_n, Y_n)\), and hence is random.

So \( \mathcal{R}(\hat{\beta}) = \mathbb{E}[(X^T \hat{\beta} - Y)^2 \mid \hat{\beta}] \) is a random variable, with expectation

\[
\mathbb{E} \left[ \mathbb{E}[(X^T \hat{\beta} - Y)^2 \mid \hat{\beta}] \right] = \mathbb{E} \left[ \mathcal{R}(\hat{\beta}) \right].
\]

(Here, we explicitly show conditioning on \( \hat{\beta} \).)

**Key observation**: Inside the (conditional) expectation is “test risk” on a single example \((X, Y)\), which is independent of \( \hat{\beta} \).

**Main idea**: Instead of one test example, use \( n \) iid test examples and average. *Same (conditional) expectation!*
**Ghost sample**: Let \((\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_n, \tilde{Y}_n) \sim \text{iid } P\) be independent of \((X_1, Y_1), \ldots, (X_n, Y_n)\).
Ghost sample: Let $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_n, \tilde{Y}_n) \sim_{iid} P$ be independent of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then

$$\mathbb{E}[(X^T \hat{\beta} - Y)^2 | \hat{\beta}] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \hat{\beta} - \tilde{Y}_i)^2 | \hat{\beta} \right].$$

Here we use linearity of expectation and the fact that $(\tilde{X}_i, \tilde{Y}_i) \sim P.$
The ghost sample has an empirical risk minimizer $\tilde{\beta}$:

$$\frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \tilde{\beta} - \tilde{Y}_i)^2 = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \beta - \tilde{Y}_i)^2$$
The ghost sample has an empirical risk minimizer $\tilde{\beta}$:

$$\frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \tilde{\beta} - \tilde{Y}_i)^2 = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \beta - \tilde{Y}_i)^2$$

Therefore

$$\frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \hat{\beta} - \tilde{Y}_i)^2 \leq \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \hat{\beta} - \tilde{Y}_i)^2$$

since $\hat{\beta}$ cannot have smaller (ghost sample) empirical risk than the ghost sample ERM $\tilde{\beta}$.
The ghost sample has an empirical risk minimizer $\tilde{\beta}$:

$$\frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \tilde{\beta} - \tilde{Y}_i)^2 = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \beta - \tilde{Y}_i)^2$$

Therefore

$$\frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \tilde{\beta} - \tilde{Y}_i)^2 \leq \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \hat{\beta} - \tilde{Y}_i)^2$$

since $\hat{\beta}$ cannot have smaller (ghost sample) empirical risk than the ghost sample ERM $\tilde{\beta}$.

Now take expectations of both sides.
Conclusion

\[ \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i^T \hat{\beta} - \hat{Y}_i)^2 \right] \leq \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i^T \hat{\beta} - \hat{Y}_i)^2 \right]. \]
Conclusion

\[ \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \tilde{\beta} - \tilde{Y}_i)^2 \right] \leq \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \hat{\beta} - \tilde{Y}_i)^2 \right]. \]

- LHS: expectation of empirical risk of ERM, i.e., \( \mathbb{E}[\mathcal{R}_n(\hat{\beta})] \).
- RHS: expectation of risk of ERM, i.e., \( \mathbb{E}[\mathcal{R}(\hat{\beta})] \).

BTW: this is also true for any other loss function and function class, not just squared loss and linear functions.
\[
\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \beta - \tilde{Y}_i)^2 \right] \leq \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\tilde{X}_i^T \hat{\beta} - \tilde{Y}_i)^2 \right].
\]

- LHS: expectation of empirical risk of ERM, i.e., \(\mathbb{E}[R_n(\hat{\beta})]\).
- RHS: expectation of risk of ERM, i.e., \(\mathbb{E}[\mathcal{R}(\hat{\beta})]\).

BTW: this is also true for any other loss function and function class, not just squared loss and linear functions.
Practice problem 5c
Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ be given, and let $\mathcal{R}_n$ be defined by
\[
\mathcal{R}_n(\beta) := \| A\beta - b \|_2^2.
\]

Suppose you only have $A$ and not $b$, but you are given an orthogonal projection $\hat{b}$ of $b$ onto the range of $A$. Explain how to find a minimizer of $\mathcal{R}_n$ using only $A$ and $\hat{b}$.
**Premise:** You have $A$ and $\hat{b}$; you don’t have $b$. 
**Solution**

**Premise:** You have $A$ and $\hat{b}$; you don’t have $b$.

Since $\hat{b}$ is orthogonal projection of $b$ onto range of $A$, the residual vector $r := b - \hat{b}$ is orthogonal to every vector in range of $A$. Therefore

$$\|A\beta - b\|^2 = \|A\beta - \hat{b} - r\|^2 = \|A\beta - \hat{b}\|^2 - 2r^T(A\beta - \hat{b}) + \|r\|^2 = \|A\beta - \hat{b}\|^2 + \|r\|^2.$$

So just find minimizer of $J(\beta) := \|A\beta - \hat{b}\|^2$.

In fact: You can just solve system of equations $A\beta = \hat{b}$ for $\beta$.

Guaranteed that solution exists because $\hat{b}$ is in range of $A$. 


Solution

**Premise**: You have $A$ and $\hat{b}$; you don’t have $b$.

Since $\hat{b}$ is orthogonal projection of $b$ onto range of $A$, the residual vector $r := b - \hat{b}$ is orthogonal to every vector in range of $A$.

Therefore

$$\|A\beta - b\|_2^2 = \|A\beta - \hat{b} - r\|_2^2$$

$$= \|A\beta - \hat{b}\|_2^2 - 2r^\top(A\beta - \hat{b}) + \|r\|_2^2$$

$$= \|A\beta - \hat{b}\|_2^2 + \|r\|_2^2.$$
Solution

**Premise:** You have $A$ and $\hat{b}$; you don’t have $b$.

Since $\hat{b}$ is orthogonal projection of $b$ onto range of $A$, the residual vector $r := b - \hat{b}$ is orthogonal to every vector in range of $A$.

Therefore

$$\|A\beta - b\|_2^2 = \|A\beta - \hat{b} - r\|_2^2$$

$$= \|A\beta - \hat{b}\|_2^2 - 2r^T(A\beta - \hat{b}) + \|r\|_2^2$$

$$= \|A\beta - \hat{b}\|_2^2 + \|r\|_2^2.$$

So just find minimizer of $J(\beta) := \|A\beta - \hat{b}\|_2^2$. 
Solution

Premise: You have $A$ and $\hat{b}$; you don’t have $b$.

Since $\hat{b}$ is orthogonal projection of $b$ onto range of $A$, the residual vector $r := b - \hat{b}$ is orthogonal to every vector in range of $A$.

Therefore

$$\|A\beta - b\|_2^2 = \|A\beta - \hat{b} - r\|_2^2$$

$$= \|A\beta - \hat{b}\|_2^2 - 2r^T(A\beta - \hat{b}) + \|r\|_2^2$$

$$= \|A\beta - \hat{b}\|_2^2 + \|r\|_2^2.$$

So just find minimizer of $J(\beta) := \|A\beta - \hat{b}\|_2^2$.

In fact: You can just solve system of equations $A\beta = \hat{b}$ for $\beta$. Guaranteed that solution exists because $\hat{b}$ is in range of $A$.  
