Generalization theory

COMS 4771 Fall 2018
Review: Statistical model of learning
Basic goal of machine learning

**Goal**: Predict outcome $y$ from set of possible outcomes $\mathcal{Y}$, on the basis of observation $x$ from feature space $\mathcal{X}$.

▶ **Examples**:

1. $x =$ email message, $y =$ spam or ham
2. $x =$ image of handwritten digit, $y =$ digit
3. $x =$ medical test results, $y =$ disease status

**Learning algorithm**:

▶ Receives *training data* $\{(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}\}$

and returns a prediction function

$$\hat{f}: \mathcal{X} \rightarrow \mathcal{Y}.$$ 

▶ On (new) *test example* $(x, y)$, predict $\hat{f}(x)$.  


Assessing the quality of predictions

**Loss function**: $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$

- Prediction is $\hat{y}$, true outcome is $y$.
- Loss $\ell(\hat{y}, y)$ measures how bad $\hat{y}$ is as a prediction of $y$.

---

**Examples**:

1. Zero-one loss:

   $\ell(\hat{y}, y) = 1\{\hat{y} \neq y\} = \begin{cases} 
   0 & \text{if } \hat{y} = y, \\
   1 & \text{if } \hat{y} \neq y.
   \end{cases}$

2. Squared loss (for $\mathcal{Y} \subseteq \mathbb{R}$):

   $\ell(\hat{y}, y) = (\hat{y} - y)^2$. 
Why is this possible?

- Only input provided to learning algorithm is training data
  \[(x_1, y_1), \ldots, (x_n, y_n).\]
- To be useful, training data must be related to test example
  \[(x, y).\]

How can we formalize this?
 IID model of data

Regard training data and test example as *independent and identically distributed* $(\mathcal{X} \times \mathcal{Y})$-valued random variables:

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

Can use tools from probability to study behavior of learning algorithms under this model.
Loss $\ell(f(X), Y)$ is random, so study average-case performance.

**Risk** of a prediction function $f$, defined by

$$\mathcal{R}(f) = \mathbb{E}[\ell(f(X), Y)],$$

where expectation is taken with respect to test example $(X, Y)$.

**Examples:**

1. *Mean squared error:* $\ell = \text{squared loss},$

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

2. *Error rate:* $\ell = \text{zero-one loss},$

   $$\mathcal{R}(f) = \mathbb{P}(f(X) \neq Y).$$
Inductive bias
Is predictability enough?

Requirements for learning:

- Relationship between training data and test example
  - Formalized by iid model for data.
- Relationship between $Y$ and $X$.
  - Example: $X$ and $Y$ are non-trivially correlated.

Is this enough?
No free lunch

For any $n \leq \frac{|\mathcal{X}|}{2}$ and any learning algorithm, there is a distribution, from which the $n$ training data and test example are drawn iid, s.t.:

1. There is a function $f^*: \mathcal{X} \to \mathcal{Y}$ with

   $$\mathbb{P}(f^*(X) \neq Y) = 0.$$ 

2. The learning algorithm returns a function $\hat{f}: \mathcal{X} \to \mathcal{Y}$ with

   $$\mathbb{P}(\hat{f}(X) \neq Y) \geq \frac{1}{4}.$$
Must make *some* assumption about learning problem in order for
learning algorithm to work well.

- Called *inductive bias* of the learning algorithm.

Common approach:

- Assume there is a good prediction function in a restricted
  function class $\mathcal{F} \subset \mathcal{Y}^\mathcal{X}$ (e.g., $\mathcal{F} =$ linear classifiers).
- Goal: find $\hat{f} : \mathcal{X} \to \mathcal{Y}$ with small excess risk

$$\mathcal{R}(\hat{f}) - \min_{f \in \mathcal{F}} \mathcal{R}(f)$$

either in expectation or with high probability over random draw
of training data.
Over-fitting and generalization
Over-fitting:

Phenomenon where learning algorithm returns \( \hat{f} \) that “fits” training data well, but does not give accurate predictions on test examples.

- **Empirical risk** of \( f \) (on training data \( (X_1, Y_1), \ldots, (X_n, Y_n) \)):

\[
R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).
\]

- **Over-fitting**: \( R_n(\hat{f}) \) small, but \( R(\hat{f}) \) large.
How to avoid over-fitting

“Theorem”: $\mathcal{R}(\hat{f}) - \mathcal{R}_n(\hat{f})$ is likely to be small, if learning algorithm chooses $\hat{f}$ from $\mathcal{F}$ that is “not too rich” relative to $n$.

- $\Rightarrow$ Observed performance on training data (i.e., empirical risk) generalizes to expected performance on test example (i.e., risk).
- Justifies learning algorithms based on minimizing empirical risk.
Risk decomposition

\[ R(\hat{f}) = \min_{g: \mathcal{X} \to \mathcal{Y}} R(g) \]

\[ + \min_{f \in \mathcal{F}} R(f) - \min_{g: \mathcal{X} \to \mathcal{Y}} R(g) \]

\[ + \min_{f \in \mathcal{F}} R_n(f) - \min_{f \in \mathcal{F}} R(f) \]

\[ + R_n(\hat{f}) - \min_{f \in \mathcal{F}} R_n(f) \]

\[ + R(\hat{f}) - R_n(\hat{f}) \]

(inherent unpredictability)

(approximation gap)

(optimization gap)

(more estimation gap)

▶ **Approximation:**

▶ Which function classes \( \mathcal{F} \) are “rich enough” for a broad class of learning problems?

▶ E.g., neural networks, Reproducing Kernel Hilbert Spaces.

▶ **Optimization:**

▶ Often finding minimizer of \( R_n \) is computationally hard.

▶ What can we do instead?
Generalization theory: Basic issues
Case study: Support vector machines

\[ \mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{-1, +1\}. \]

- Return solution \( \hat{w} \in \mathbb{R}^d \) to following optimization problem:
  \[
  \min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} [1 - y_i w^T x_i]^+.
  \]

- Loss function is hinge loss
  \[
  \ell(\hat{y}, y) = [1 - y\hat{y}]_+ = \max\{1 - y\hat{y}, 0\}.
  \]
  (Here, we are okay with a real-valued prediction.)

- The \( \frac{\lambda}{2} \|w\|_2^2 \) term is called Tikhonov regularization, which we’ll discuss later.
Basic statistical model for data

IID model of data

- Training data and test example are *independent and identically distributed* \((\mathcal{X} \times \mathcal{Y})\)-valued random variables:

\[
(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim \text{iid } P.
\]
Basic statistical model for data

**IID model of data**

- Training data and test example are *independent and identically distributed* \((\mathcal{X} \times \mathcal{Y})\)-valued random variables:

  \[(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{iid} P.\]

**SVM in the iid model**

- Return solution \(\hat{w}\) to following optimization problem:

  \[
  \min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} [1 - Y_i w^T X_i]^+. \]

- Therefore, \(\hat{w}\) is a random variable, depending on \((X_1, Y_1), \ldots, (X_n, Y_n)\).
For $w$ that does not depend on training data:

Empirical risk

$$
\mathcal{R}_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w^T X_i, Y_i)
$$

is a sum of iid random variables.
For $w$ that does not depend on training data:

Empirical risk

$$\mathcal{R}_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w^T X_i, Y_i)$$

is a sum of iid random variables.

Law of Large Numbers gives an asymptotic result:

$$\mathcal{R}_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w^T X_i, Y_i) \rightarrow \mathbb{E}[\ell(w^T X, Y)] = \mathcal{R}(w).$$

(“$\rightarrow$” denotes some kind of probabilistic asymptotic convergence; but this result can also be made non-asymptotic.)
However, $\hat{w}$ does depend on training data.

Empirical risk of $\hat{w}$ is *not* a sum of iid random variables:

$$R_n(\hat{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{w}^T X_i, Y_i).$$
However, $\hat{w}$ does depend on training data.

Empirical risk of $\hat{w}$ is *not* a sum of iid random variables:

$$R_n(\hat{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{w}^T X_i, Y_i).$$

**Idea**: $\hat{w}$ could conceivably take any value $w$, but if

$$\max_w |R_n(w) - R(w)| \to 0,$$

then $R_n(\hat{w}) \to R(\hat{w})$ as well.

(1) is called *uniform convergence*. 
Detour: Concentration inequalities
Symmetric random walk

Rademacher random variables

\( \varepsilon_1, \ldots, \varepsilon_n \) iid with \( \mathbb{P}(\varepsilon_i = -1) = \mathbb{P}(\varepsilon_i = 1) = 1/2 \).

Position of symmetric random walk: after \( n \) steps,

\[
S_n = \sum_{i=1}^{n} \varepsilon_i.
\]
Symmetric random walk

Rademacher random variables

\(\varepsilon_1, \ldots, \varepsilon_n\) iid with \(\mathbb{P}(\varepsilon_i = -1) = \mathbb{P}(\varepsilon_i = 1) = 1/2\).

Position of symmetric random walk: after \(n\) steps,

\[ S_n = \sum_{i=1}^{n} \varepsilon_i. \]

How far from origin?

- By independence, \(\text{var}(S_n) = \sum_{i=1}^{n} \text{var}(\varepsilon_i) = n.\)
- So expected distance from origin is

\[ \mathbb{E}|S_n| \leq \sqrt{\text{var}(S_n)} \leq \sqrt{n}. \]
Symmetric random walk

Rademacher random variables

\( \varepsilon_1, \ldots, \varepsilon_n \) iid with \( \mathbb{P}(\varepsilon_i = -1) = \mathbb{P}(\varepsilon_i = 1) = 1/2 \).

Position of symmetric random walk: after \( n \) steps,

\[
S_n = \sum_{i=1}^{n} \varepsilon_i.
\]

How far from origin?

- By independence, \( \text{var}(S_n) = \sum_{i=1}^{n} \text{var}(\varepsilon_i) = n \).
- So expected distance from origin is

\[
\mathbb{E}|S_n| \leq \sqrt{\text{var}(S_n)} \leq \sqrt{n}.
\]

How many realizations are \( \gg \sqrt{n} \) from origin?
Markov’s inequality

For any random variable $X$ and any $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq \frac{E|X|}{t}. $$

Proof:

$$t \cdot 1\{|X| \geq t\} \leq |X|. $$
Markov’s inequality

For any random variable $X$ and any $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t}.$$

Proof:

$$t \cdot 1\{|X| \geq t\} \leq |X|.$$

Application to symmetric random walk:

$$\mathbb{P}(|S_n| \geq c\sqrt{n}) \leq \frac{\mathbb{E}|S_n|}{c\sqrt{n}} \leq \frac{1}{c}.$$
Hoeffding’s inequality

If $X_1, \ldots, X_n$ are independent random variables, with $X_i$ taking values in $[a_i, b_i]$, then for any $t \geq 0$,

$$
P \left( \sum_{i=1}^{n} (X_i - \mathbb{E}(X_i)) \geq t \right) \leq \exp \left( -\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$
Hoeffding’s inequality

If $X_1, \ldots, X_n$ are independent random variables, with $X_i$ taking values in $[a_i, b_i]$, then for any $t \geq 0$,

$$
\mathbb{P} \left( \sum_{i=1}^{n} (X_i - \mathbb{E}(X_i)) \geq t \right) \leq \exp \left( -\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).
$$

E.g., Rademacher random variables have $[a_i, b_i] = [-1, +1]$, so

$$
\mathbb{P}(S_n \geq t) \leq \exp(-2t^2/(4n)).
$$
Applying Hoeffding’s inequality to symmetric random walk

Union bound: For any events $A$ and $B$,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$
Union bound: For any events $A$ and $B$,\[\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).\]

1. Apply Hoeffding to $\varepsilon_1, \ldots, \varepsilon_n$: \[\mathbb{P}(S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).\]
Applying Hoeffding’s inequality to symmetric random walk

Union bound: For any events $A$ and $B$,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

1. Apply Hoeffding to $\varepsilon_1, \ldots, \varepsilon_n$:

$$\mathbb{P}(S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).$$

2. Apply Hoeffding to $-\varepsilon_1, \ldots, -\varepsilon_n$:

$$\mathbb{P}(-S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).$$
Applying Hoeffding’s inequality to symmetric random walk

Union bound: For any events \( A \) and \( B \),

\[
P(A \cup B) \leq P(A) + P(B).
\]

1. Apply Hoeffding to \( \varepsilon_1, \ldots, \varepsilon_n \):

\[
P(S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).
\]

2. Apply Hoeffding to \( -\varepsilon_1, \ldots, -\varepsilon_n \):

\[
P(-S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).
\]

3. Therefore, by union bound,

\[
P(|S_n| \geq c\sqrt{n}) \leq 2 \exp(-c^2/2).
\]

(Compare to bound from Markov’s inequality: \( 1/c \).)
Equivalent form of Hoeffding’s inequality

Let $X_1, \ldots, X_n$ be independent random variables, with $X_i$ taking values in $[a_i, b_i]$, and let $S_n = \sum_{i=1}^{n} X_i$. For any $\delta \in (0, 1)$,

$$P \left( S_n - \mathbb{E}[S_n] < \sqrt{\frac{1}{2} \sum_{i=1}^{n} (b_i - a_i)^2 \ln(1/\delta)} \right) \geq 1 - \delta.$$ 

This is a “high probability” upper-bound on $S_n - \mathbb{E}[S_n]$. 
Uniform convergence: Finite classes
Back to statistical learning

Cast of characters:

- feature and outcome spaces: \( \mathcal{X}, \mathcal{Y} \)
- function class: \( \mathcal{F} \subset \mathcal{Y}^{\mathcal{X}} \)
- loss function: \( \ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \) (assume bounded by 1)
- training and test data: \( (X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P \)
Back to statistical learning

Cast of characters:

- feature and outcome spaces: $\mathcal{X}, \mathcal{Y}$
- function class: $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$
- loss function: $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ (assume bounded by 1)
- training and test data: $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P$

We let $\hat{f} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}_n(f)$ be minimizer of empirical risk

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$
Back to statistical learning

Cast of characters:

- feature and outcome spaces: \( \mathcal{X}, \mathcal{Y} \)
- function class: \( \mathcal{F} \subset \mathcal{Y}^{\mathcal{X}} \)
- loss function: \( \ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+ \) (assume bounded by 1)
- training and test data: \( (X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P \)

We let \( \hat{f} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}_n(f) \) be minimizer of empirical risk

\[
\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).
\]

Our worry: over-fitting \( \mathcal{R}(\hat{f}) \gg \mathcal{R}_n(\hat{f}) \).

Simple solution: ensure no function over-fits!
Convergence of empirical risk for fixed function

For any fixed function $f \in \mathcal{F}$,

$$
\mathbb{E} \left[ \mathcal{R}_n(f) \right] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \ell(f(X_i), Y_i) \right] = \mathcal{R}(f).
$$
Convergence of empirical risk for fixed function

For any fixed function $f \in \mathcal{F}$,

$$
\mathbb{E} \left[ R_n(f) \right] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \ell(f(X_i), Y_i) \right] = R(f).
$$

Since $R_n(f)$ is sum of $n$ independent $[0, \frac{1}{n}]$-valued random variables,

$$
\mathbb{P} \left( |R_n(f) - R(f)| \geq t \right) \leq 2 \exp \left( -\frac{2t^2}{\sum_{i=1}^{n}(\frac{1}{n})^2} \right) = 2 \exp(-2nt^2)
$$

for any $t > 0$, by Hoeffding’s inequality and union bound.
Convergence of empirical risk for fixed function

For any fixed function \( f \in \mathcal{F} \),

\[
\mathbb{E} [\mathcal{R}_n(f)] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} [\ell(f(X_i), Y_i)] = \mathcal{R}(f).
\]

Since \( \mathcal{R}_n(f) \) is sum of \( n \) independent \([0, \frac{1}{n}]\)-valued random variables,

\[
\mathbb{P} (|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t) \leq 2 \exp \left( -\frac{2t^2}{\sum_{i=1}^{n} \frac{1}{n}^2} \right) = 2 \exp(-2nt^2)
\]

for any \( t > 0 \), by Hoeffding’s inequality and union bound.

This argument does not apply to \( \hat{f} \), because \( \hat{f} \) depends on \((X_1, Y_1), \ldots, (X_n, Y_n)\).
Uniform convergence for all functions in a finite class

If $|\mathcal{F}| < \infty$, then by Hoeffding’s inequality and union bound,

$$
P \left( \exists f \in \mathcal{F} \text{ s.t. } |R_n(f) - R(f)| \geq t \right) = P \left( \bigcup_{f \in \mathcal{F}} \{|R_n(f) - R(f)| \geq t\} \right)
$$

$$
\leq \sum_{f \in \mathcal{F}} P \left( |R_n(f) - R(f)| \geq t \right)
$$

$$
\leq |\mathcal{F}| \cdot 2 \exp(-2nt^2).
$$
Uniform convergence for all functions in a finite class

If $|\mathcal{F}| < \infty$, then by Hoeffding’s inequality and union bound,

$$
P \left( \exists f \in \mathcal{F} \text{ s.t. } |R_n(f) - \mathcal{R}(f)| \geq t \right) = P \left( \bigcup_{f \in \mathcal{F}} \{ |R_n(f) - \mathcal{R}(f)| \geq t \} \right)
$$

$$
\leq \sum_{f \in \mathcal{F}} P (|R_n(f) - \mathcal{R}(f)| \geq t)

\leq |\mathcal{F}| \cdot 2 \exp(-2nt^2).
$$

Choose $t$ so that RHS is $\delta$, and “invert”.

**Theorem.** For any $\delta \in (0, 1)$,

$$
P \left( \forall f \in \mathcal{F} : |R_n(f) - \mathcal{R}(f)| < \sqrt{\frac{\ln(2|\mathcal{F}|/\delta)}{2n}} \right) \geq 1 - \delta.
$$
What we get from uniform convergence

If \( n \gg \log |\mathcal{F}| \), then with high probability, no function \( f \in \mathcal{F} \) will over-fit the training data.
What we get from uniform convergence

If \( n \gg \log |\mathcal{F}| \), then with high probability, no function \( f \in \mathcal{F} \) will over-fit the training data.

**Also:** An **empirical risk minimizer (ERM)**, like \( \hat{f} \), is near optimal!

**Theorem.** With probability at least \( 1 - \delta \),

\[
\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) = \mathcal{R}(\hat{f}) - \mathcal{R}_n(\hat{f}) \leq \epsilon \\
+ \mathcal{R}_n(\hat{f}) - \mathcal{R}_n(f^*) \leq 0 \\
+ \mathcal{R}_n(f^*) - \mathcal{R}(f^*) \leq \epsilon \\
\leq 2\epsilon
\]

where \( f^* \in \arg\min_{f \in \mathcal{F}} \mathcal{R}(f) \) and \( \epsilon = \sqrt{\frac{\ln(2|\mathcal{F}|/\delta)}{2n}} \).
Uniform convergence: General case
Uniform convergence: General case

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a class of real-valued functions, and let $P$ be a probability distribution on $\mathcal{X}$. 
Uniform convergence: General case

Let $\mathcal{F} \subset \mathbb{R}^\mathcal{X}$ be a class of real-valued functions, and let $P$ be a probability distribution on $\mathcal{X}$.

Notation:

- Let $Pf = \mathbb{E}[f(X)]$ for $X \sim P$.
- Let $P_n$ be the empirical distribution on $X_1, \ldots, X_n \sim_{iid} P$, which assigns probability mass $1/n$ to each $X_i$.
- So $P_nf = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$.

We are interested in the maximum deviation:

$$\max_{f \in \mathcal{F}} |P_nf - Pf|.$$
Uniform convergence: General case

Let $\mathcal{F} \subset \mathbb{R}^X$ be a class of real-valued functions, and let $P$ be a probability distribution on $X$.

**Notation:**

- Let $Pf = \mathbb{E}[f(X)]$ for $X \sim P$.
- Let $P_n$ be the *empirical distribution* on $X_1, \ldots, X_n \sim_{iid} P$, which assigns probability mass $1/n$ to each $X_i$.
- So $P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$.

We are interested in the maximum deviation:

$$\max_{f \in \mathcal{F}} |P_n f - Pf|.$$

The arguments from before show that for any *finite class* of bounded functions $\mathcal{F}$,

$$\max_{f \in \mathcal{F}} |P_n f - Pf| \to 0.$$
Infinite classes

For which classes $\mathcal{F} \subset \mathbb{R}^X$ does uniform convergence hold?

Example:

$F = \{ f_S(x) = 1 \mid x \in S \subset \mathbb{R}, |S| < \infty \}$, i.e., $\{0, 1\}$-valued functions that take value 1 on a finite set.

$\Rightarrow$ If $P$ is continuous, then $Pf = 0$ for all $f \in \mathcal{F}$.

$\Rightarrow$ But $\max f \in \mathcal{F} P_n f = 1$ for all $n$.

$\Rightarrow$ So $\max f \in \mathcal{F} |P_n f - Pf| = 1$ for all $n$. 

What is the appropriate "complexity" measure of a function class?
Infinite classes

For which classes $\mathcal{F} \subset \mathbb{R}^X$ does uniform convergence hold?

**Example:**

$$\mathcal{F} = \{ f_S(x) = 1_{\{ x \in S \}} : S \subset \mathbb{R}, |S| < \infty \},$$

i.e., $\{0, 1\}$-valued functions that take value 1 on a finite set.

- If $P$ is continuous, then $Pf = 0$ for all $f \in \mathcal{F}$.
- But $\max_{f \in \mathcal{F}} P_n f = 1$ for all $n$.
- So $\max_{f \in \mathcal{F}} |P_n f - Pf| = 1$ for all $n$. 

What is the appropriate "complexity" measure of a function class?
For which classes \( \mathcal{F} \subseteq \mathbb{R}^X \) does uniform convergence hold?

**Example:**

\[
\mathcal{F} = \{ f_S(x) = 1\{x \in S\} : S \subseteq \mathbb{R}, |S| < \infty \},
\]

i.e., \( \{0, 1\} \)-valued functions that take value 1 on a finite set.

- If \( P \) is continuous, then \( Pf = 0 \) for all \( f \in \mathcal{F} \).
- But \( \max_{f \in \mathcal{F}} P_nf = 1 \) for all \( n \).
- So \( \max_{f \in \mathcal{F}} |P_nf - Pf| = 1 \) for all \( n \).

What is the appropriate “complexity” measure of a function class?
Rademacher complexity

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables.

*Rademacher complexity* of $\mathcal{F}$:

$$\text{Rad}_n(\mathcal{F}) = \mathbb{E}_\varepsilon \mathbb{E}_\varepsilon \left[ \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right]$$

where $\mathbb{E}_\varepsilon$ is expectation with respect to $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. 
Rademacher complexity

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables.

**Rademacher complexity of $\mathcal{F}$:**

$$\text{Rad}_n(\mathcal{F}) = \mathbb{E} \mathbb{E}_{\varepsilon} \left[ \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right]$$

where $\mathbb{E}_{\varepsilon}$ is expectation with respect to $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$.

$\text{Rad}_n(\mathcal{F})$ measures how well vectors in (random) set

$$\mathcal{F}(X_{1:n}) = \{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$$

can correlate with uniformly random signs $\varepsilon_1, \ldots, \varepsilon_n$. 

Theorem. Uniform convergence with $\mathcal{F}$ holds iff $\lim_{n \to \infty} \text{Rad}_n(\mathcal{F}) = 0$. 

Rademacher complexity

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables.

**Rademacher complexity of $F$:**

$$\text{Rad}_n(F) = \mathbb{E}\mathbb{E}_{\varepsilon} \left[ \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right]$$

where $\mathbb{E}_\varepsilon$ is expectation with respect to $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$.

$\text{Rad}_n(F)$ measures how well vectors in (random) set

$$\mathcal{F}(X_{1:n}) = \{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$$

can correlate with uniformly random signs $\varepsilon_1, \ldots, \varepsilon_n$.

**Theorem.** Uniform convergence with $\mathcal{F}$ holds iff

$$\lim_{n \to \infty} \text{Rad}_n(\mathcal{F}) = 0.$$
Extreme cases of Rademacher complexity

For simplicity, assume $X_1, \ldots, X_n$ are distinct (e.g., $P$ continuous).

- $\mathcal{F}$ contains a single function $f_0: \mathcal{X} \to \{-1, +1\}$:

  $$\text{Rad}_n(\mathcal{F}) = \mathbb{E} \mathbb{E}_\epsilon \left[ \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_0(X_i) \right| \right] \leq \frac{1}{\sqrt{n}}.$$

- $\mathcal{F}$ contains all functions $\mathcal{X} \to \{-1, +1\}$:

  $$\text{Rad}_n(\mathcal{F}) = \mathbb{E} \mathbb{E}_\epsilon \left[ \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right] = 1.$$
**Theorem.** For any $\mathcal{F} \subset \mathbb{R}^X$, 

$$
\mathbb{E} \left[ \max_{f \in \mathcal{F}} |P_n f - Pf| \right] \leq 2 \text{Rad}_n(\mathcal{F}).
$$

For any $\mathcal{F} \subset [-1, +1]^X$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$,

$$
\max_{f \in \mathcal{F}} |P_n f - Pf| \leq 2 \text{Rad}_n(\mathcal{F}) + \sqrt{\frac{2 \ln(1/\delta)}{n}}.
$$
Step 1: Symmetrization by “ghost sample”

Let $P'_n$ be empirical distribution on independent copies $X'_1, \ldots, X'_n$ of $X_1, \ldots, X_n$. Write $E'$ for expectation with respect to $X'_1:n$. 

\[
\max_{f \in F} |P_n f - Pf| \leq \mathbb{E} \left( \max_{f \in F} \left| \mathbb{E}' \sum_{i=1}^{n} f(X_i) - f(X'_i) \right| \right).
\]

Inequality follows from Jensen’s inequality: if $h$ is convex, then $h \left( \mathbb{E}(Z) \right) \leq \mathbb{E}(h(Z))$. 

Step 1: Symmetrization by “ghost sample”

Let $P'_n$ be empirical distribution on independent copies $X'_1, \ldots, X'_n$ of $X_1, \ldots, X_n$. Write $E'$ for expectation with respect to $X'_1:n$.

Then

$$
\mathbb{E} \left[ \max_{f \in \mathcal{F}} |P_n f - P f| \right] = \mathbb{E} \max_{f \in \mathcal{F}} \left| \mathbb{E}' \left[ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X'_i) \right] \right|
$$

$$
\leq \mathbb{E} \mathbb{E}' \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X'_i) \right|
$$

$$
= \mathbb{E} \mathbb{E}' \max_{f \in \mathcal{F}} |P_n f - P'_n f|.
$$

Inequality follows from Jensen’s inequality: if $h$ is convex, then $h(\mathbb{E}(Z)) \leq \mathbb{E}(h(Z))$. 

Step 2: Symmetrization by random signs

Consider any \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, +1\}^n \). Distribution of

\[
P_n f - P'_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X'_i)
\]

is the same distribution of

\[
P_n f - P'_n f = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left( f(X_i) - f(X'_i) \right).
\]
Step 2: Symmetrization by random signs

Consider any \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, +1\}^n \). Distribution of

\[
P_n f - P'_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X'_i)
\]

is the same distribution of

\[
P_n f - P'_n f = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left( f(X_i) - f(X'_i) \right).
\]

Thus, this is also true for uniform average over all \( \varepsilon \in \{-1, +1\}^n \) (i.e., expectation over Rademacher \( \varepsilon \)):

\[
\mathbb{E} \mathbb{E}' \max_{f \in \mathcal{F}} |P_n f - P'_n f| = \mathbb{E} \mathbb{E}' \mathbb{E}_\varepsilon \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left( f(X_i) - f(X'_i) \right) \right|
\]
Step 3: Back to a single sample

By triangle inequality,

\[
\max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left( f(X_i) - f(X'_i) \right) \right|
\leq \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| + \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X'_i) \right|
\]

The two terms on the RHS have the same distribution.
Step 3: Back to a single sample

By triangle inequality,

\[
\max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left( f(X_i) - f(X'_i) \right) \right| \\
\leq \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| + \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X'_i) \right|
\]

The two terms on the RHS have the same distribution.

So

\[
\mathbb{E} \mathbb{E}' \mathbb{E}_\varepsilon \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left( f(X_i) - f(X'_i) \right) \right| \leq 2 \mathbb{E} \mathbb{E}_\varepsilon \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \\
= 2 \text{Rad}_n(\mathcal{F}).
\]
For any $\mathcal{F} \subset \mathbb{R}^X$, 

$$
\mathbb{E} \left[ \max_{f \in \mathcal{F}} |P_n f - Pf| \right] \leq 2 \text{Rad}_n(\mathcal{F}).
$$

For any $\mathcal{F} \subset [-1, +1]^X$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$, 

$$
\max_{f \in \mathcal{F}} |P_n f - Pf| \leq 2 \text{Rad}_n(\mathcal{F}) + \sqrt{\frac{2 \ln(1/\delta)}{n}}.
$$
Recap

For any $\mathcal{F} \subset \mathbb{R}^X$,

$$
\mathbb{E} \left( \max_{f \in \mathcal{F}} |P_n f - P f| \right) \leq 2 \text{Rad}_n(\mathcal{F}).
$$

For any $\mathcal{F} \subset [-1, +1]^X$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$,

$$
\max_{f \in \mathcal{F}} |P_n f - P f| \leq 2 \text{Rad}_n(\mathcal{F}) + \sqrt{\frac{2 \ln(1/\delta)}{n}}.
$$

Conclusion

If $\text{Rad}_n(\mathcal{F}) \to 0$, then uniform convergence holds.
For any $\mathcal{F} \subset \mathbb{R}^X$,

$$\mathbb{E} \left[ \max_{f \in \mathcal{F}} |P_n f - P f| \right] \leq 2 \text{Rad}_n(\mathcal{F}).$$

For any $\mathcal{F} \subset [-1, +1]^X$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$,

$$\max_{f \in \mathcal{F}} |P_n f - P f| \leq 2 \text{Rad}_n(\mathcal{F}) + \sqrt{\frac{2 \ln(1/\delta)}{n}}.$$

**Conclusion**

If $\text{Rad}_n(\mathcal{F}) \to 0$, then uniform convergence holds.

(Can also show: If uniform convergence holds, then $\text{Rad}_n(\mathcal{F}) \to 0$.)
Analysis of SVM
Loss class

Back to classes of prediction functions $\mathcal{F} \subset \mathbb{R}^X$. 

Consider a loss function $\ell : \mathbb{R} \times Y \to \mathbb{R}^+$ that satisfies $\ell(0,y) \leq 1$ for all $y \in Y$, and is 1-Lipschitz in first argument: for all $\hat{y}, \hat{y}' \in \mathbb{R}$,

$$|\ell(\hat{y},y) - \ell(\hat{y}',y)| \leq |\hat{y} - \hat{y}'|.$$ 

(Example: hinge loss $\ell(\hat{y},y) = [1 - \hat{y}y]^+$.)

Define the associated loss class by $\ell\mathcal{F} = \{(x,y) \mapsto \ell(f(x),y) : f \in \mathcal{F}\}$.

Then $\text{Rad}_n(\ell\mathcal{F}) \leq 2 \text{Rad}_n(\mathcal{F}) + \sqrt{2 \ln 2 \over n}$. So uniform convergence holds for $\ell\mathcal{F}$ if it holds for $\mathcal{F}$. 

34 / 46
Loss class

Back to classes of prediction functions $\mathcal{F} \subset \mathbb{R}^X$.

Consider a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ that satisfies $\ell(0, y) \leq 1$ for all $y \in \mathcal{Y}$, and is 1-Lipschitz in first argument: for all $\hat{y}, \hat{y}' \in \mathbb{R},$

$$|\ell(\hat{y}, y) - \ell(\hat{y}', y)| \leq |\hat{y} - \hat{y}'|.$$  

(Example: hinge loss $\ell(\hat{y}, y) = [1 - \hat{y}y]_+$.)

Back to classes of prediction functions $\mathcal{F} \subset \mathbb{R}^X$.

Consider a loss function $\ell: \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ that satisfies $\ell(0, y) \leq 1$ for all $y \in \mathcal{Y}$, and is 1-Lipschitz in first argument: for all $\hat{y}, \hat{y}' \in \mathbb{R}$,

$$|\ell(\hat{y}, y) - \ell(\hat{y}', y)| \leq |\hat{y} - \hat{y}'|.$$  

(Example: hinge loss $\ell(\hat{y}, y) = [1 - \hat{y} y]_+$.)

Define the associated loss class by

$$\ell_\mathcal{F} = \{(x, y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}.$$

Then

$$\text{Rad}_n(\ell_\mathcal{F}) \leq 2 \text{Rad}_n(\mathcal{F}) + \sqrt{\frac{2 \ln 2}{n}}.$$  

So uniform convergence holds for $\ell_\mathcal{F}$ if it holds for $\mathcal{F}$. 

Linear functions $\mathcal{F}_{\text{lin}} = \{ w \in \mathbb{R}^d \}$.

What is the Rademacher complexity of $\mathcal{F}_{\text{lin}}$?

$$\text{Rad}_n(\mathcal{F}_{\text{lin}}) = \mathbb{E} \mathbb{E}_\varepsilon \left[ \max_{w \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i w^\top X_i \right| \right].$$
Rademacher complexity of linear predictors

Linear functions $\mathcal{F}_{\text{lin}} = \{ w \in \mathbb{R}^d \}$.

What is the Rademacher complexity of $\mathcal{F}_{\text{lin}}$?

$$\text{Rad}_n(\mathcal{F}_{\text{lin}}) = \mathbb{E}_{\mathbb{E}_\varepsilon} \left[ \max_{w \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i w^\top X_i \right| \right].$$

Inside the $\mathbb{E}_{\mathbb{E}_\varepsilon}$:

$$\max_{w \in \mathbb{R}^d} \left| w^\top \left( \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right) \right| = \max_{w \in \mathbb{R}^d} \| w \|_2 \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_2.$$

As long as $\sum_{i=1}^{n} \varepsilon_i X_i \neq 0$, this is unbounded!
Regularization

Recall SVM optimization problem:

$$
\min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} [1 - y_i w^T x_i]_+.
$$
Recall SVM optimization problem:

\[
\min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} [1 - y_i w^\top x_i]^+.
\]

Objective value at \( w = 0 \) is 1, so objective value at minimizer \( \hat{w} \) is no worse than this:

\[
\frac{\lambda}{2} \|\hat{w}\|^2 + \frac{1}{n} \sum_{i=1}^{n} [1 - y_i \hat{w}^\top x_i]^+ \leq 1.
\]
Recall SVM optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} [1 - y_i w^T x_i]^+.$$ 

Objective value at $w = 0$ is 1, so objective value at minimizer $\hat{w}$ is no worse than this:

$$\frac{\lambda}{2} \|\hat{w}\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} [1 - y_i \hat{w}^T x_i]^+ \leq 1.$$ 

Therefore

$$\|\hat{w}\|_2^2 \leq \frac{2}{\lambda}.$$
Rademacher complexity of bounded linear predictors

**Bounded** linear functions $\mathcal{F}_{\ell_2,B} = \{ w \in \mathbb{R}^d : \|w\|_2 \leq B \}$.  

This is a $d$-dimensional random walk, where the $i$-th step is $\pm X_i$.  

Rademacher complexity of bounded linear predictors

**Bounded** linear functions $\mathcal{F}_{\ell_2,B} = \{ w \in \mathbb{R}^d : \|w\|_2 \leq B \}$.

What is the Rademacher complexity of $\mathcal{F}_{\ell_2,B}$?

$$\text{Rad}_n(\mathcal{F}_{\ell_2,B}) = \mathbb{E}\mathbb{E}_{\varepsilon} \left[ \max_{\|w\|_2 \leq B} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i w^T X_i \right| \right]$$

$$= B \mathbb{E}\mathbb{E}_{\varepsilon} \left[ \max_{\|u\|_2 \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i u^T X_i \right| \right]$$

$$= B \mathbb{E}\mathbb{E}_{\varepsilon} \left( \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_2 \right) \leq B \sqrt{\mathbb{E}\mathbb{E}_{\varepsilon} \left( \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_2^2 \right)}.$$
Rademacher complexity of bounded linear predictors

**Bounded** linear functions $\mathcal{F}_{\ell_2,B} = \{w \in \mathbb{R}^d : \|w\|_2 \leq B\}$.

What is the Rademacher complexity of $\mathcal{F}_{\ell_2,B}$?

$$\text{Rad}_n(\mathcal{F}_{\ell_2,B}) = \mathbb{E}\mathbb{E}_\varepsilon \left[ \max_{\|w\|_2 \leq B} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i w^\top X_i \right| \right]$$

$$= B\mathbb{E}\mathbb{E}_\varepsilon \left[ \max_{\|u\|_2 \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i u^\top X_i \right| \right]$$

$$= B\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_2 \leq B \sqrt{\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_2^2}.$$ 

This is $d$-dimensional random walk, where $i$-th step is $\pm X_i$. 
Rademacher complexity of bounded linear predictors (2)

$$\mathbb{E} \mathbb{E}_{\varepsilon} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_2^2 = \frac{1}{n^2} \mathbb{E} \mathbb{E}_{\varepsilon} \left[ \sum_{i=1}^{n} \| \varepsilon_i X_i \|_2^2 + \sum_{i \neq j} \varepsilon_i \varepsilon_j X_i^T X_j \right]$$

$$= \frac{1}{n^2} \mathbb{E} \left[ \sum_{i=1}^{n} \| X_i \|_2^2 \right]$$

$$= \frac{1}{n} \mathbb{E} \| X \|_2^2.$$

Conclusion

Rademacher complexity of $F_{\ell^2, B} = \{ w \in \mathbb{R}^d : \| w \|_2 \leq B \}$:

$$\text{Rad}_n(F_{\ell^2, B}) \leq B \sqrt{\frac{2}{n}} \mathbb{E} \| X \|_2^2.$$
Rademacher complexity of bounded linear predictors (2)

\[
\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_2^2 = \frac{1}{n^2} \mathbb{E}\mathbb{E}_\varepsilon \left[ \sum_{i=1}^{n} \| \varepsilon_i X_i \|_2^2 + \sum_{i \neq j} \varepsilon_i \varepsilon_j X_i^T X_j \right]
\]
\[
= \frac{1}{n^2} \mathbb{E} \left[ \sum_{i=1}^{n} \| X_i \|_2^2 \right]
\]
\[
= \frac{1}{n} \mathbb{E} \| X \|_2^2.
\]

Conclusion

Rademacher complexity of \( \mathcal{F}_{\ell_2,B} = \{ w \in \mathbb{R}^d : \| w \|_2 \leq B \} \):

\[
\text{Rad}_n(\mathcal{F}_{\ell_2,B}) \leq B \sqrt{\frac{\mathbb{E} \| X \|_2^2}{n}}.
\]
Risk bound for SVM

\[
\mathbb{E} [\mathcal{R}(\hat{w}) - \mathcal{R}(w^*)] \\
= \mathbb{E} [\mathcal{R}(\hat{w}) - \mathcal{R}_n(\hat{w})] \\
+ \mathbb{E} \left[ \frac{\lambda}{2} \|\hat{w}\|^2 + \mathcal{R}_n(\hat{w}) - \frac{\lambda}{2} \|w^*\|^2 - \mathcal{R}_n(w^*) \right] \leq \epsilon \quad (\leq \epsilon)
\]

\[
+ \mathbb{E} [\mathcal{R}_n(w^*) - \mathcal{R}(w^*)] \quad (= 0)
\]

\[
+ \mathbb{E} \left[ \frac{\lambda}{2} \|w^*\|^2 - \frac{\lambda}{2} \|\hat{w}\|^2 \right] \\
\leq \epsilon + \frac{\lambda}{2} \|w^*\|^2
\]

where

\[
w^* \in \arg\min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|^2 + \mathcal{R}(w), \quad \epsilon = O \left( \sqrt{\frac{\mathbb{E}\|X\|_2^2}{\lambda n}} + \frac{1}{\sqrt{n}} \right).
\]
Risk bound for SVM

\[ \mathbb{E} \left[ \mathcal{R}(\hat{w}) - \mathcal{R}(w^*) \right] \]
\[ = \mathbb{E} \left[ \mathcal{R}(\hat{w}) - \mathcal{R}_n(\hat{w}) \right] \tag{\leq \epsilon} \]
\[ + \mathbb{E} \left[ \frac{\lambda}{2} \| \hat{w} \|^2_2 + \mathcal{R}_n(\hat{w}) - \frac{\lambda}{2} \| w^* \|^2_2 - \mathcal{R}_n(w^*) \right] \tag{\leq 0} \]
\[ + \mathbb{E} \left[ \mathcal{R}_n(w^*) - \mathcal{R}(w^*) \right] \tag{= 0} \]
\[ + \mathbb{E} \left[ \frac{\lambda}{2} \| \hat{w} \|^2_2 - \frac{\lambda}{2} \| \hat{w} \|^2_2 \right] \]
\[ \leq \epsilon + \frac{\lambda}{2} \| w^* \|^2_2 \]

where

\[ w^* \in \arg\min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \| w \|^2_2 + \mathcal{R}(w), \quad \epsilon = O \left( \sqrt{\frac{\mathbb{E} \| X \|^2_2}{\lambda n}} + \frac{1}{\sqrt{n}} \right). \]

This suggests we should use \( \lambda \to 0 \) such that \( \lambda n \to \infty \) as \( n \to \infty \).
Excess risk bound has no *explicit* dependence on the dimension $d$. In particular, it holds in infinite dimensional inner product spaces.

- SVM can be applied in such spaces as long as there is an algorithm for computing inner products.
- This is the *kernel trick*, and these corresponding spaces are called *Reproducing Kernel Hilbert Spaces* (RKHS).
Excess risk bound has no *explicit* dependence on the dimension $d$. In particular, it holds in infinite dimensional inner product spaces.

- SVM can be applied in such spaces as long as there is an algorithm for computing inner products.
- This is the *kernel trick*, and these corresponding spaces are called *Reproducing Kernel Hilbert Spaces (RKHS)*.

**Universal approximation**

With some RKHS, can approximate any function arbitrarily well:

$$\lim_{\lambda \to 0} \left\{ \min_{w \in \mathcal{F}} \frac{\lambda}{2} \|w\|^2 + \mathcal{R}(w) \right\} = \min_{g: \mathcal{X} \to \mathbb{R}} \mathcal{R}(g).$$
Other regularizers

Instead of SVM, suppose $\hat{w}$ is solution to

$$\min_{w \in \mathbb{R}^d} \lambda \|w\|_1 + \mathcal{R}_n(w).$$

So $\hat{w} \in \mathcal{F}_{\ell_1,B} = \{w \in \mathbb{R}^d : \|w\|_1 \leq B\}$ for $B = 1/\lambda$. 
Other regularizers

Instead of SVM, suppose \( \hat{w} \) is solution to

\[
\min_{w \in \mathbb{R}^d} \lambda \|w\|_1 + \mathcal{R}_n(w).
\]

So \( \hat{w} \in \mathcal{F}_{\ell_1,B} = \{ w \in \mathbb{R}^d : \|w\|_1 \leq B \} \) for \( B = 1/\lambda \).

What is Rademacher complexity of \( \mathcal{F}_{\ell_1,B} \)?

\[
\mathbf{Rad}_n(\mathcal{F}_{\ell_1,B}) = \mathbb{E} \mathbb{E}_\varepsilon \left[ \max_{\|w\|_1 \leq B} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i w^T X_i \right| \right]
\]

\[
= B \mathbb{E} \mathbb{E}_\varepsilon \left[ \max_{\|u\|_1 \leq 1} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i u^T X_i \right| \right]
\]

\[
= B \mathbb{E} \mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_\infty.
\]
Rademacher complexity of $\ell_1$-bounded linear predictors

Can show, using martingale argument,

$$\mathbb{E} \mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_\infty \leq \sqrt{O(\log d) \cdot \mathbb{E} \| X \|_2^2} \cdot \frac{\mathbb{E} \| X \|_\infty^2}{n}.$$
Rademacher complexity of $\ell_1$-bounded linear predictors

Can show, using martingale argument,

$$\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_\infty \leq \sqrt{\frac{O(\log d) \cdot \mathbb{E} \| X \|_2^2}{n}}.$$ 

Rademacher complexity of $\mathcal{F}_{\ell_1,B} = \{ w \in \mathbb{R}^d : \| w \|_1 \leq B \}$:

$$\text{Rad}_n(\mathcal{F}_{\ell_1,B}) \leq B \sqrt{\frac{O(\log d) \cdot \mathbb{E} \| X \|_2^2}{n}}.$$
Can show, using martingale argument,

\[
\mathbb{E} \mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_\infty \leq \sqrt{\frac{O(\log d) \cdot \mathbb{E} \|X\|_2^2}{n}}.
\]

Rademacher complexity of \( F_{\ell_1, B} = \{ w \in \mathbb{R}^d : \|w\|_1 \leq B \} \):

\[
\text{Rad}_n(F_{\ell_1, B}) \leq B \sqrt{\frac{O(\log d) \cdot \mathbb{E} \|X\|_2^2}{n}}.
\]

Let \( \mathcal{X} = \{-1, +1\}^d \). Then \( \|x\|_2^2 = d \) but \( \|x\|_\infty^2 = 1 \) for all \( x \in \mathcal{X} \).

Dependence on \( d \) much better than using bound for \( \ell_2 \)-bounded linear predictors, which would have looked like \( B \sqrt{d/n} \).
Beyond uniform convergence
Deficiencies of uniform convergence analysis

- For certain loss functions, if $\mathcal{R}(f)$ is small, then variance of $\mathcal{R}_n(f)$ is also small, and bound should reflect this.
  - Instead of Hoeffding’s / McDiarmid’s inequalities, use concentration inequalities that involve variance information (e.g., Bernstein’s / Talagrand’s inequalities).

- Overkill to require *all* functions in $\mathcal{F}$ to not over-fit.
  - Should just need to worry about those with small empirical risk.
  - Solution: Use *local* Rademacher complexity.
Example: Occam’s razor bound

Suppose $\mathcal{F}$ is countable and we fix (a priori) a probability distribution $\pi = (\pi_f : f \in \mathcal{F})$ on $\mathcal{F}$.

- Think of $\pi$ as placing bets on which functions are likely to be the one to be picked by your learning algorithm.
Example: Occam’s razor bound

Suppose $\mathcal{F}$ is countable and we fix (a priori) a probability distribution $\pi = (\pi_f : f \in \mathcal{F})$ on $\mathcal{F}$.

Think of $\pi$ as placing bets on which functions are likely to be the one to be picked by your learning algorithm.

For any fixed $f \in \mathcal{F}$,

$$\mathbb{P}\left(|R_n(f) - R(f)| \geq t_f \right) \leq 2 \exp(-2nt_f^2)$$

for any $t_f > 0$, by Hoeffding’s inequality and union bound.

**Note:** We can choose the $t_f$'s non-uniformly.
Occam’s razor bound (continued)

Let $t_f = \sqrt{\frac{\ln(1/\pi_f) + \ln(2/\delta)}{2n}}$.

By union bound,

$$
\mathbb{P}\left( \exists f \in \mathcal{F} \text{ s.t. } |\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f \right)
\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left( |\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f \right)
\leq \sum_{f \in \mathcal{F}} 2 \exp(-2nt_f^2) = \sum_{f \in \mathcal{F}} \pi_f \delta = \delta.
$$
Occam’s razor bound (continued)

Let \( t_f = \sqrt{\frac{\ln(1/\pi_f) + \ln(2/\delta)}{2n}} \).

By union bound,

\[
\mathbb{P} \left( \exists f \in F \text{ s.t. } |R_n(f) - R(f)| \geq t_f \right) \\
\leq \sum_{f \in F} \mathbb{P} \left( |R_n(f) - R(f)| \geq t_f \right) \\
\leq \sum_{f \in F} 2 \exp(-2nt_f^2) = \sum_{f \in F} \pi_f \delta = \delta.
\]

**Theorem.** For any \( \delta \in (0, 1) \),

\[
\mathbb{P} \left( \forall f \in F : |R_n(f) - R(f)| < \sqrt{\frac{\ln(1/\pi_f) + \ln(2/\delta)}{2n}} \right) \geq 1 - \delta.
\]
Let $t_f = \sqrt{\frac{\ln(1/\pi_f) + \ln(2/\delta)}{2n}}$.

By union bound,

$$
\mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } |R_n(f) - R(f)| \geq t_f\right)
\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(|R_n(f) - R(f)| \geq t_f\right)
\leq \sum_{f \in \mathcal{F}} 2 \exp(-2nt_f^2) = \sum_{f \in \mathcal{F}} \pi_f \delta = \delta.
$$

**Theorem.** For any $\delta \in (0, 1)$,

$$
\mathbb{P}\left(\forall f \in \mathcal{F} : |R_n(f) - R(f)| < \sqrt{\frac{\ln(1/\pi_f) + \ln(2/\delta)}{2n}}\right) \geq 1 - \delta.
$$

Better bound for functions $f$ with higher “prior probability” $\pi_f$!
Other forms of generalization analysis

- **Stability**
  - If a learning algorithm’s output does not change much if a single data point is changed, then its output will generalize.
  - Connections to differential privacy and regularization.

- **Compression bounds**
  - If a learning algorithm’s output is invariant to all but a small number $k \ll n$ of training data (e.g., # support vectors in SVM), then get bound of the form $\sqrt{k/(n-k)}$.

- **Direct analyses**
  - Some well-known learning algorithms do not fit the mold of typical (regularized) ERM algorithm, and seem to require a direct analysis.
  - E.g., nearest neighbor rule.

- Etc.