From Statistical to Online Learning

Sasha Rakhlin

Department of Statistics, The Wharton School
University of Pennsylvania

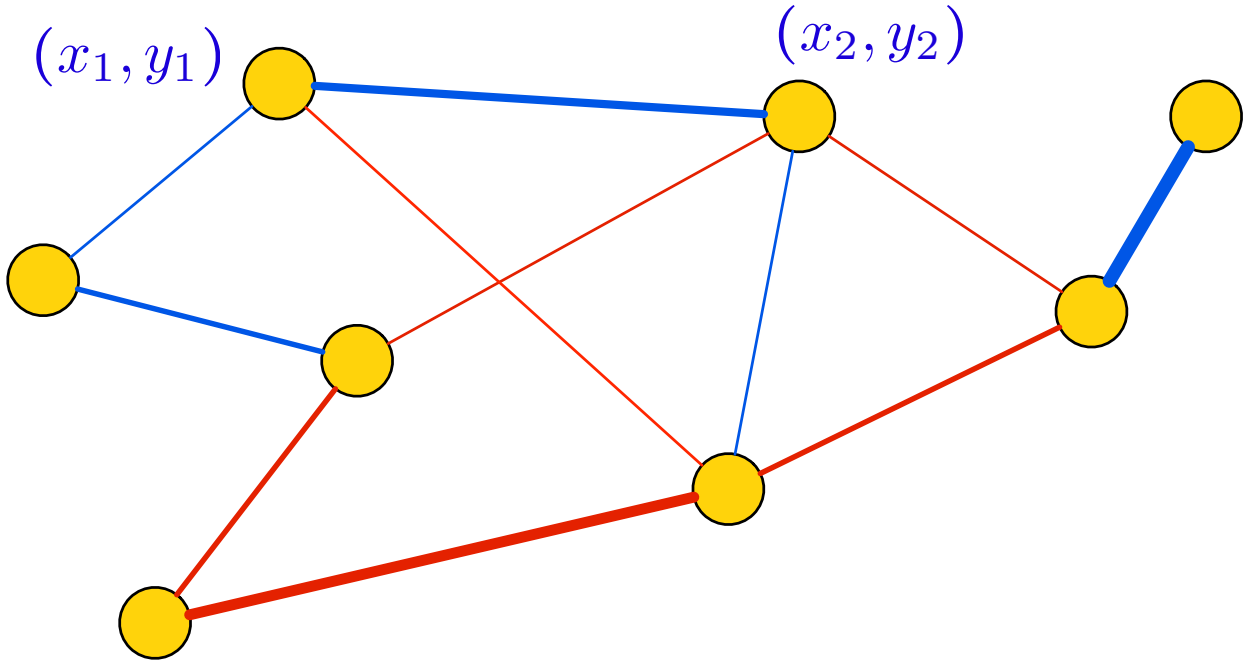
March 17-20, 2015

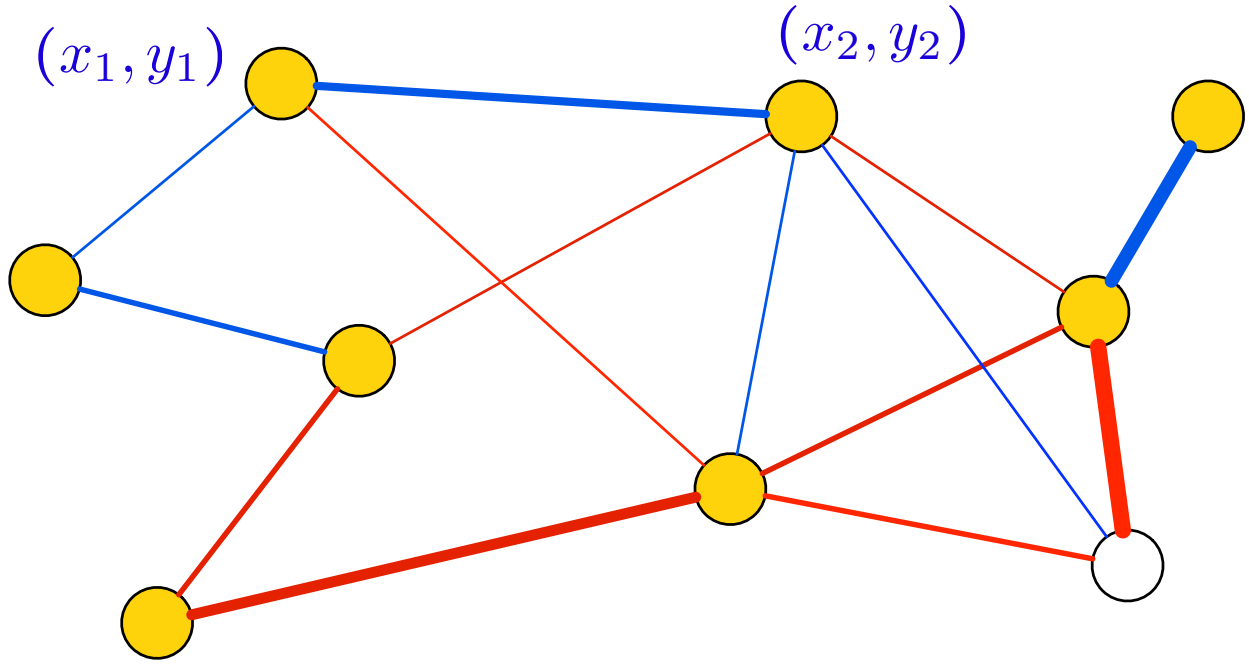
Spring School "Structural Inference"
Sylt, Germany

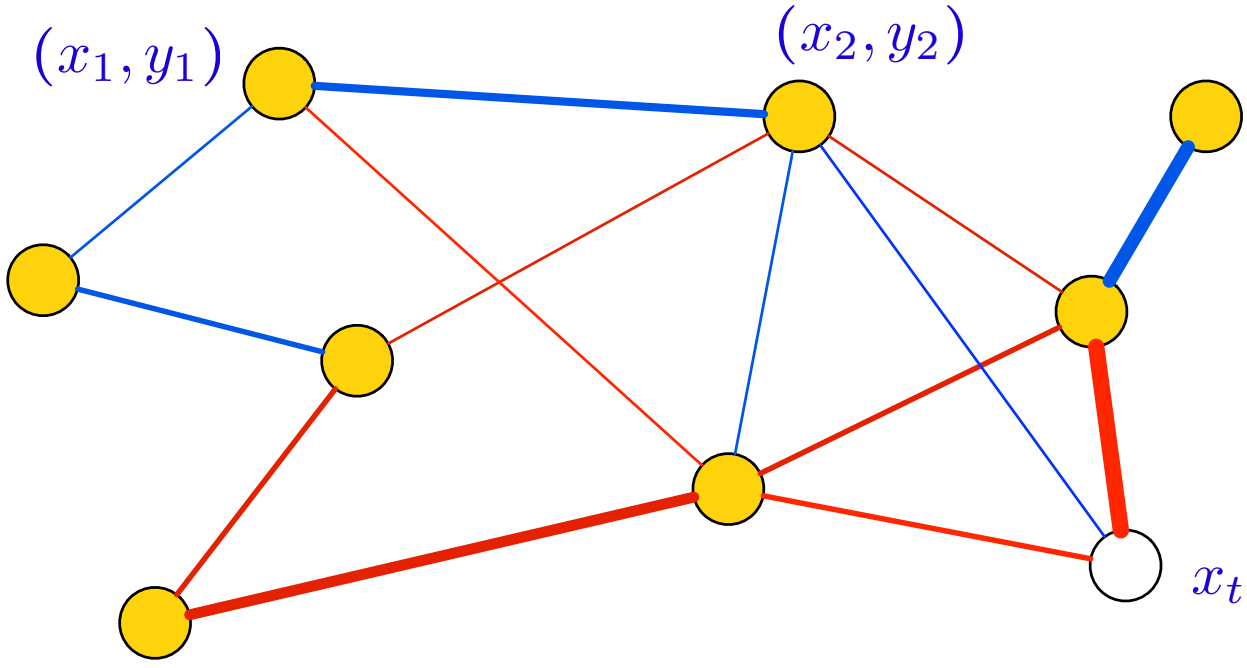
Draft of notes with K. Sridharan:
http://stat.wharton.upenn.edu/~rakhlin/book_draft.pdf

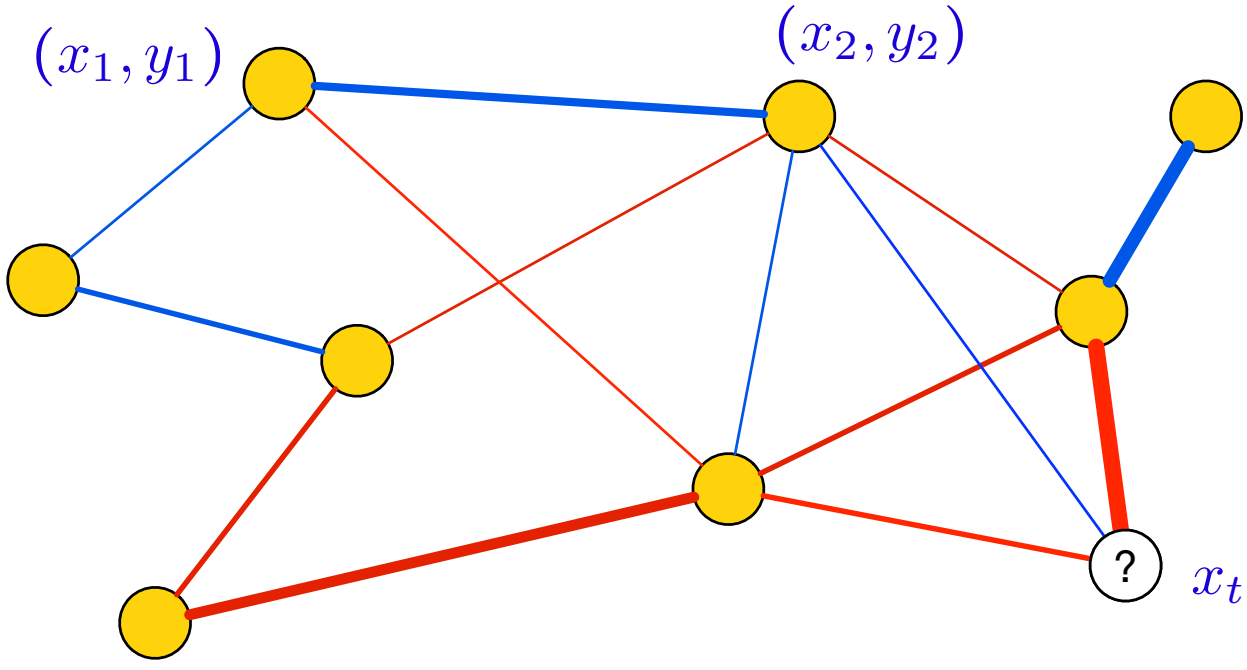
Outline

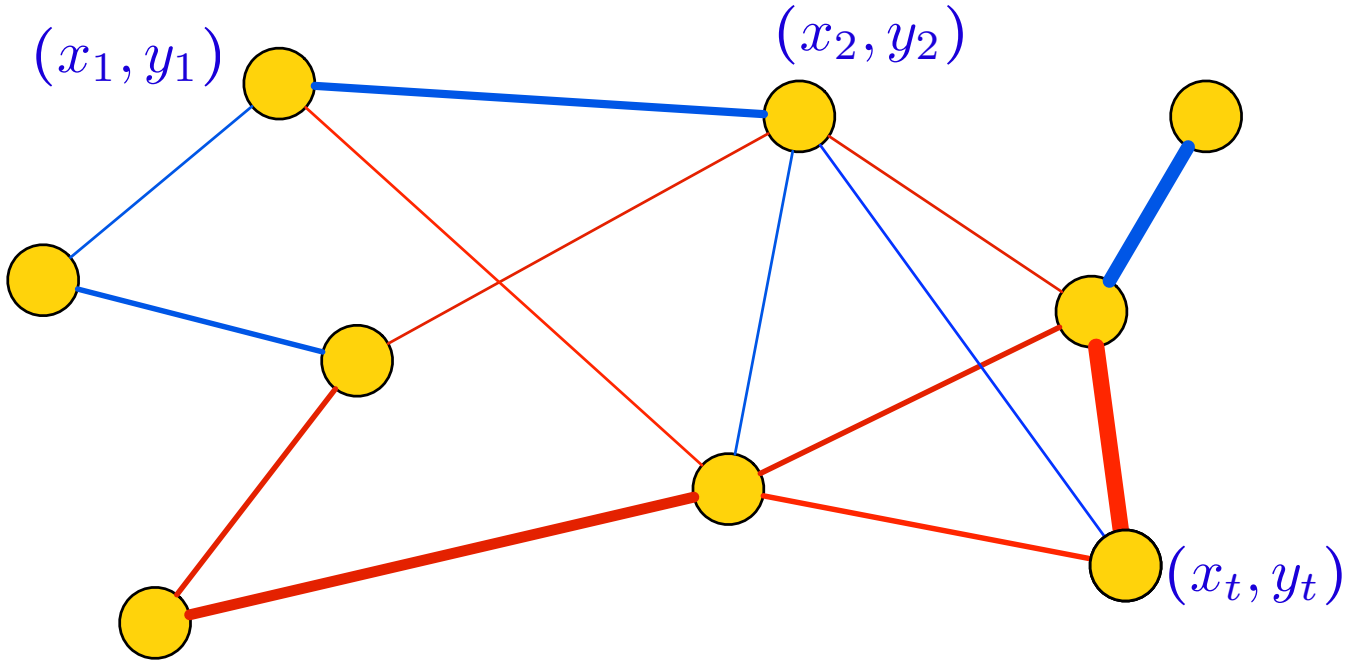
Outline

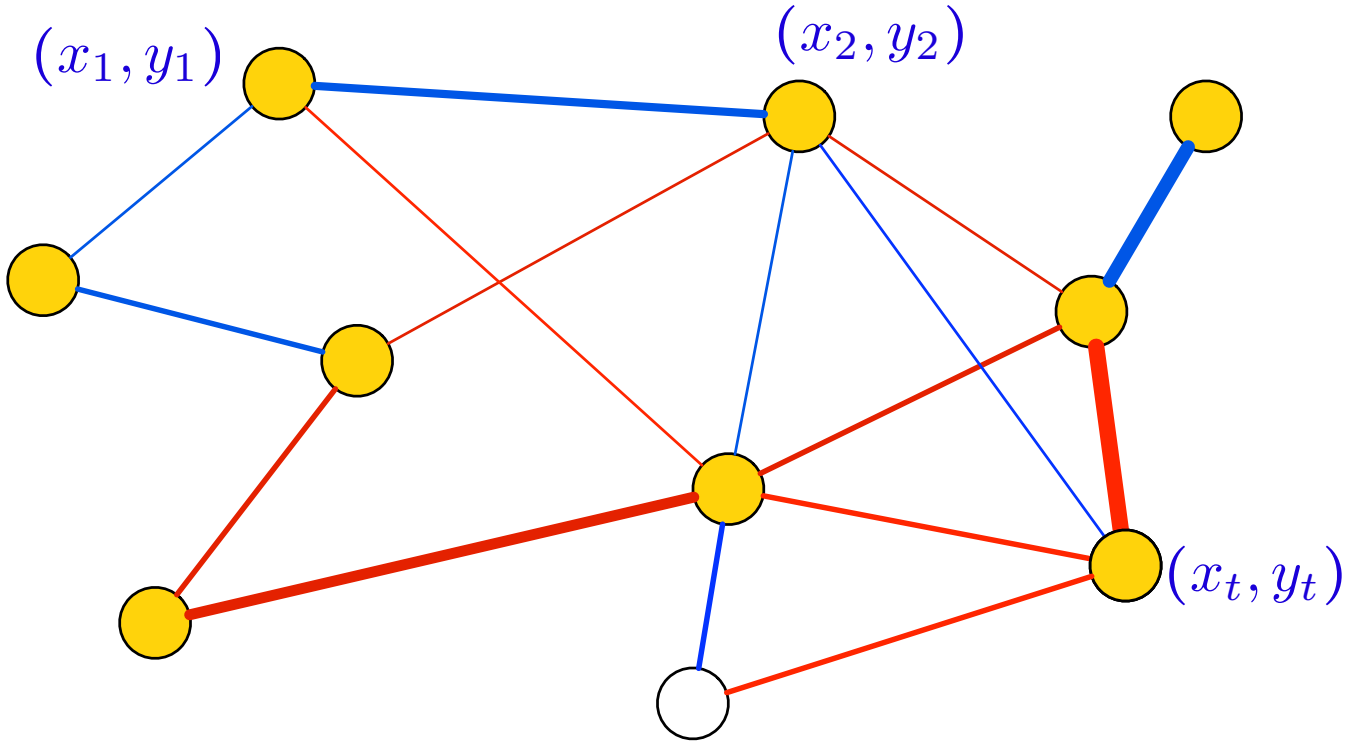

(x_1, y_1)
(x_2, y_2)

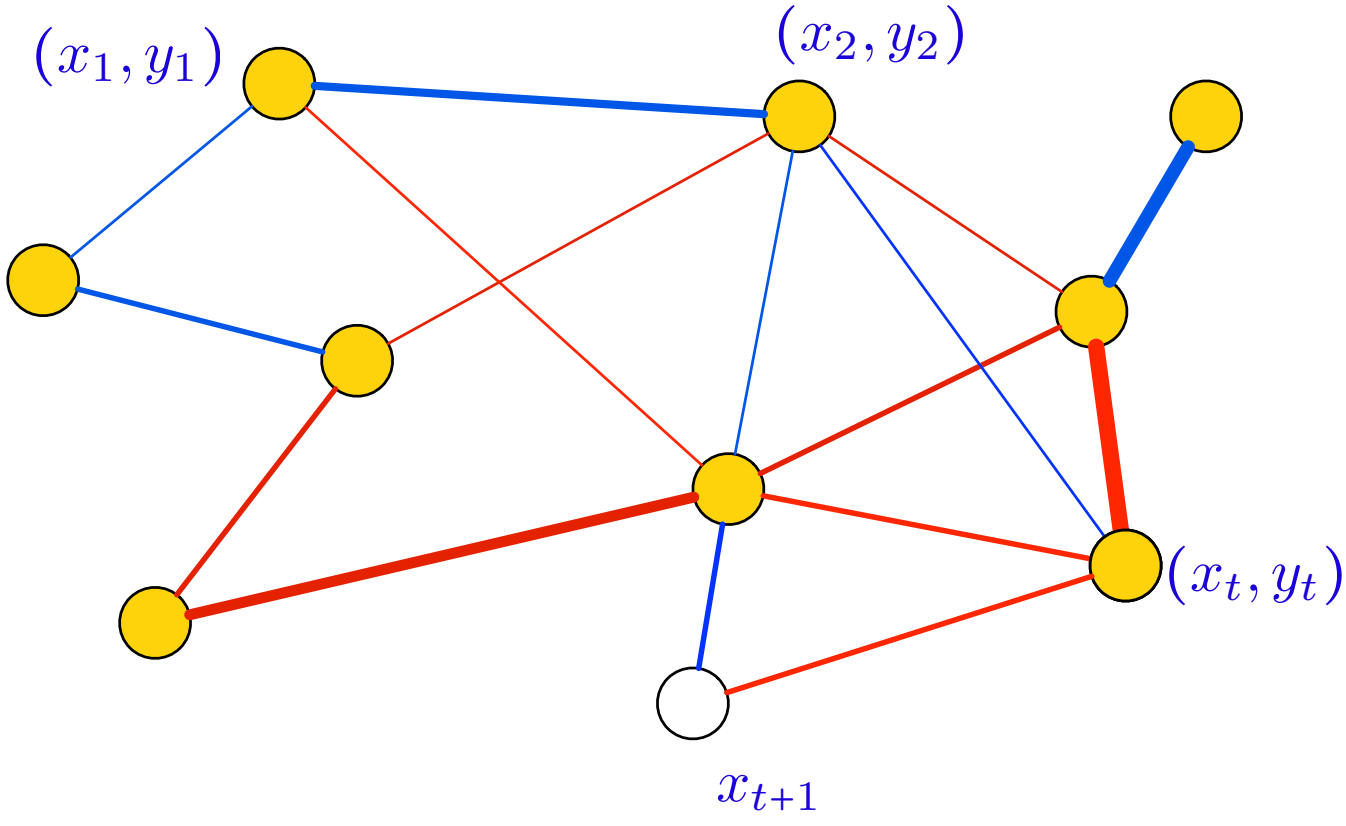

(x_1, y_1)
(x_2, y_2)

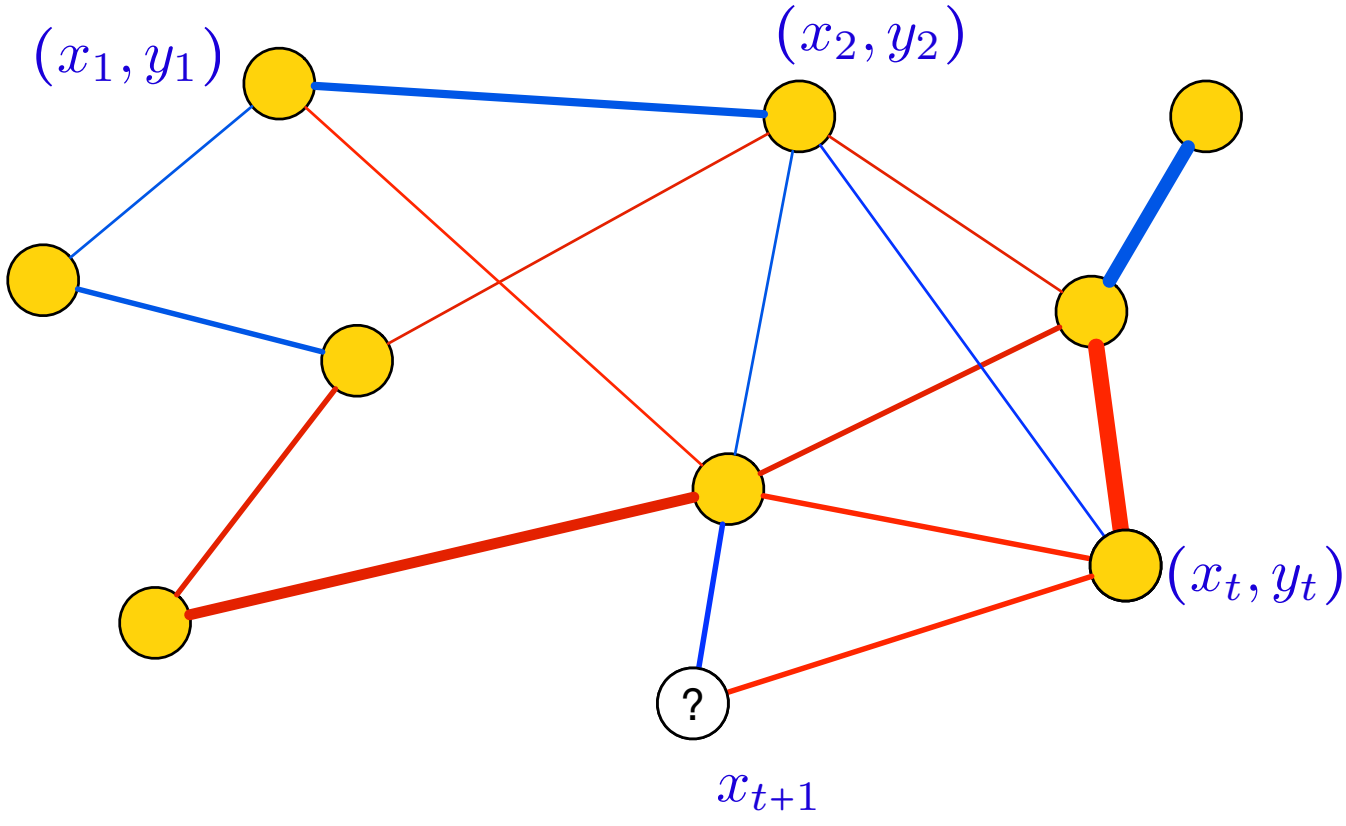

(x_1, y_1)
(x_2, y_2)
x_t

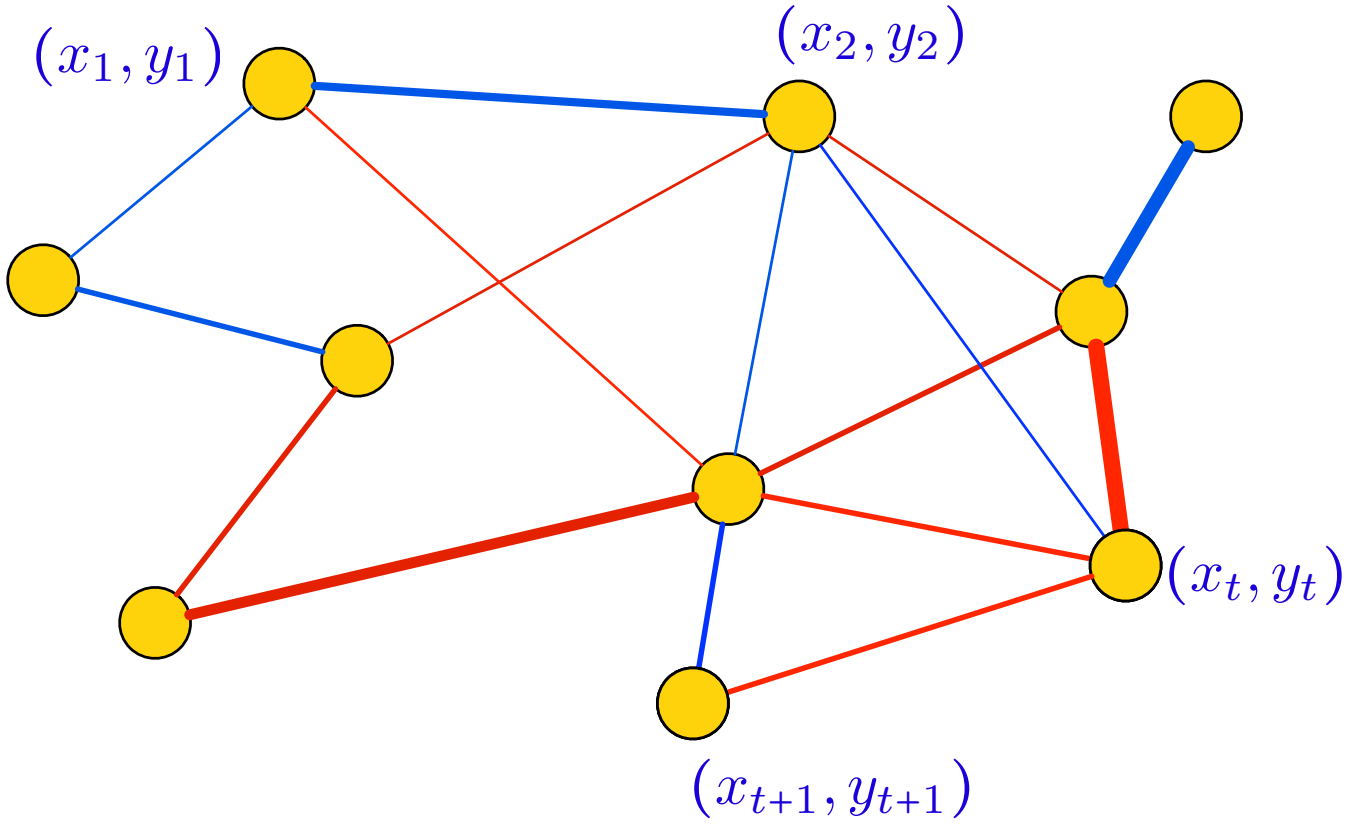

(x_1, y_1)
(x_2, y_2)
?
x_t


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)
x_{t+1}


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)
?
x_{t+1}


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)
(x_{t+1}, y_{t+1})

Prediction with sequentially-revealed data. This problem should be in the realm of statistics!

Questions:

- How to formulate the prediction problem? What does it mean to predict a sequence? What is the objective to be minimized?
- Can we employ Statistician's toolkit, or do we need new notions?
- What can we model in probabilistic way, and what can we not model? What can we treat as i.i.d.? How to incorporate assumptions? What if the model is misspecified?
- Is there a general algorithmic approach to such sequential prediction problems? (e.g. a substitute for the canonical 'maximum likelihood' principle in Statistics)
- How to develop computationally feasible methods?

Plan for 3 lectures

Before diving into a new area of online prediction, we will review some old and new results in Statistical Learning (first lecture).

We then turn to sequential prediction and develop some of the analogues in a surprising parallel to statistical learning (second lecture).

We discuss algorithmic techniques and examples in the third lecture.

Outline

Outline

Prediction

Unknown distribution $P = P_X \times P_{Y|X}$ on pair (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$.

Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. from P, find a function f that "explains the relationship."

Formally, construct estimator $\widehat{f}_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$

Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (e.g. quadratic $\ell(a, b) = (a - b)^2$, absolute $|a - b|$)

Benchmark class of functions $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$

Expected loss $\mathbf{L}(f) = \mathbb{E}\ell(f(X), Y)$. Excess loss $\mathbf{L}(f) - \inf_{f'} \mathbf{L}(f')$.

Empirical loss $\hat{\mathbf{L}}(f) = \frac{1}{n} \sum_{t=1}^{n} \ell(f(X_t), Y_t)$

Statistical learning theory

Let $\mathcal{P}$ be the set of all distributions on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{P}_0 \subseteq \mathcal{P}$.

Goal: find $\widehat{f}_n$ that approximately minimizes

$$\sup_{P \in \mathcal{P}_0} \left\{ \mathbb{E}\mathbf{L}(\widehat{f}_n) - \inf_{f \in \mathcal{F}} \mathbf{L}(f) \right\}$$

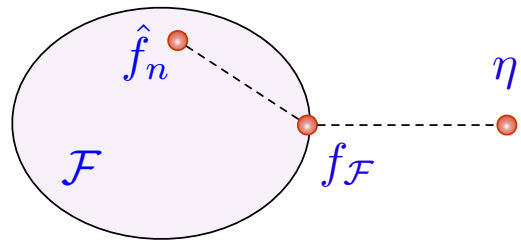
If $\mathcal{P}_0 = \mathcal{P}$, we say the setting is "distribution-free".

Let

$$\eta = \operatorname*{argmin}_f \mathbf{L}(f)$$

Bias-variance tradeoff

$$\mathbf{L}(\widehat{f}_n) - \mathbf{L}(\eta) = \underbrace{\mathbf{L}(\widehat{f}_n) - \inf_{f \in \mathcal{F}} \mathbf{L}(f)}_{\text{Estimation Error}} + \underbrace{\inf_{f \in \mathcal{F}} \mathbf{L}(f) - \mathbf{L}(\eta)}_{\text{Approximation Error}}$$



- Larger $\mathcal{F} \Rightarrow$ smaller approximation error but larger estimation error
- Larger $n \Rightarrow$ smaller estimation error and no effect on approx. error.
- Trade off size of $\mathcal{F}$ and n: *Structural Risk Minimization*, or *Method of Sieves*, or *Model Selection*.

Model selection can be done via penalization as soon as we have good bounds for *fixed* $\mathcal{F}$. We focus on the latter goal.

Square loss: prediction vs estimation

Regression function $\eta(x) = \mathbb{E}[Y|X = x]$ achieves $\operatorname{argmin}_f \mathbf{L}(f)$.

$$\mathbb{E}(\widehat{f}(X) - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2 = \mathbb{E}\|\widehat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2$$

where $\|\cdot\| = \|\cdot\|_{L_2(P_X)}$ (prove it)

Square loss: prediction vs estimation

Regression function $\eta(x) = \mathbb{E}[Y|X = x]$ achieves $\operatorname{argmin}_f \mathbf{L}(f)$.

$$\mathbb{E}(\widehat{f}(X) - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2 = \mathbb{E}\|\widehat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2$$

where $\|\cdot\| = \|\cdot\|_{L_2(P_X)}$ (prove it)

Model is *well-specified* if $\eta \in \mathcal{F}$ (this is a strong assumption!). In this case, expected excess loss of $\widehat{f}$ is same as

$$\mathbb{E}\|\widehat{f} - \eta\|^2$$

which is the problem of *estimation* in $L_2(P_X)$ norm.

Square loss: prediction vs estimation

Regression function $\eta(x) = \mathbb{E}[Y|X = x]$ achieves $\operatorname{argmin}_f \mathbf{L}(f)$.

$$\mathbb{E}(\widehat{f}(X) - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2 = \mathbb{E}\|\widehat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2$$

where $\|\cdot\| = \|\cdot\|_{L_2(P_X)}$ (prove it)

Model is *well-specified* if $\eta \in \mathcal{F}$ (this is a strong assumption!). In this case, expected excess loss of $\widehat{f}$ is same as

$$\mathbb{E}\|\widehat{f} - \eta\|^2$$

which is the problem of *estimation* in $L_2(P_X)$ norm.

If $\eta \notin \mathcal{F}$, the model is *misspecified* and we are asking for oracle inequalities.

For other loss functions, we don't have this nice connection between estimation and prediction.

Empirical risk minimization

When is it a good idea to take

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(X_t), Y_t) \quad ?$$

Sufficient condition is uniform "closeness" of empirical and expected loss:

$$\begin{aligned} \mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}}) &= \left\{\mathbf{L}(\widehat{f}_n) - \hat{\mathbf{L}}(\widehat{f}_n)\right\} + \left\{\hat{\mathbf{L}}(\widehat{f}_n) - \hat{\mathbf{L}}(f_{\mathcal{F}})\right\} + \left\{\hat{\mathbf{L}}(f_{\mathcal{F}}) - \mathbf{L}(f_{\mathcal{F}})\right\} \\ &\leq \sup_{f \in \mathcal{F}} \left\{\mathbf{L}(f) - \hat{\mathbf{L}}(f)\right\} + \left\{\hat{\mathbf{L}}(f_{\mathcal{F}}) - \mathbf{L}(f_{\mathcal{F}})\right\} \end{aligned}$$

and so in expectation

$$\mathbb{E}\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}}) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \left\{\mathbf{L}(f) - \hat{\mathbf{L}}(f)\right\}$$

Empirical risk minimization

When is it a good idea to take

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(X_t), Y_t) \quad ?$$

Sufficient condition is uniform "closeness" of empirical and expected loss:

$$\begin{aligned} \mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}}) &= \left\{\mathbf{L}(\widehat{f}_n) - \hat{\mathbf{L}}(\widehat{f}_n)\right\} + \left\{\hat{\mathbf{L}}(\widehat{f}_n) - \hat{\mathbf{L}}(f_{\mathcal{F}})\right\} + \left\{\hat{\mathbf{L}}(f_{\mathcal{F}}) - \mathbf{L}(f_{\mathcal{F}})\right\} \\ &\leq \sup_{f \in \mathcal{F}} \left\{\mathbf{L}(f) - \hat{\mathbf{L}}(f)\right\} + \left\{\hat{\mathbf{L}}(f_{\mathcal{F}}) - \mathbf{L}(f_{\mathcal{F}})\right\} \end{aligned}$$

and so in expectation

$$\mathbb{E}\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}}) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \left\{\mathbf{L}(f) - \hat{\mathbf{L}}(f)\right\}$$

Next: detour into empirical process theory.

A bit of notation to simplify things...

To ease the notation,

- $z_i = (x_i, y_i)$ so that data are $\{z_1, \ldots, z_n\}$
- $g(z) = \ell(f(x), y)$ for $z = (x, y)$
- Loss class $\mathcal{G} = \{g : g(z) = \ell(f(x), y)\} = \ell \circ \mathcal{F}$
- $\hat{g}_n = \ell(\widehat{f}_n(\cdot), \cdot)$, $g_{\mathcal{G}} = \ell(f_{\mathcal{F}}(\cdot), \cdot)$
- $g^* = \operatorname{argmin}_g \mathbb{E}g(z) = \ell(f^*(\cdot), \cdot)$

We can now work with the set $\mathcal{G}$, but keep in mind that each $g \in \mathcal{G}$ corresponds to an $f \in \mathcal{F}$:

$$g \in \mathcal{G} \quad \longleftrightarrow \quad f \in \mathcal{F}$$

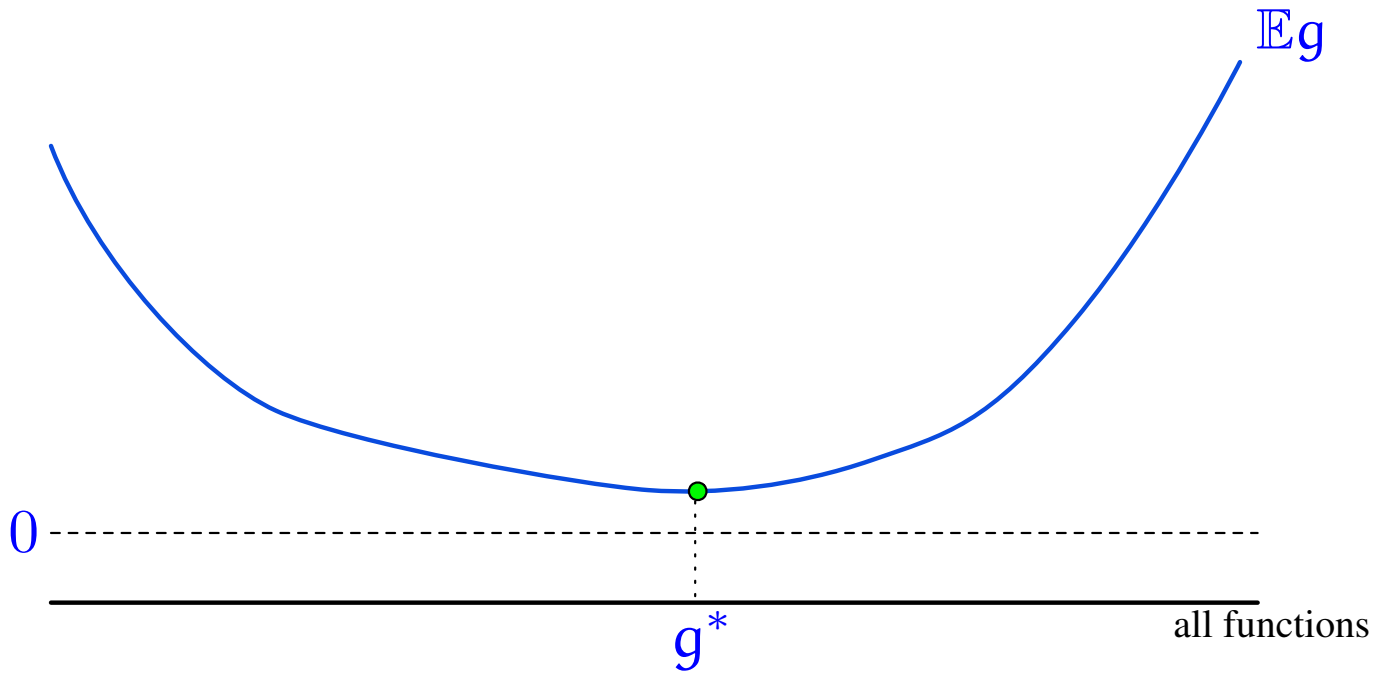
Once again, the quantity of interest is

$$\sup_{g \in \mathcal{G}} \left\{ \mathbb{E}g(z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right\}$$

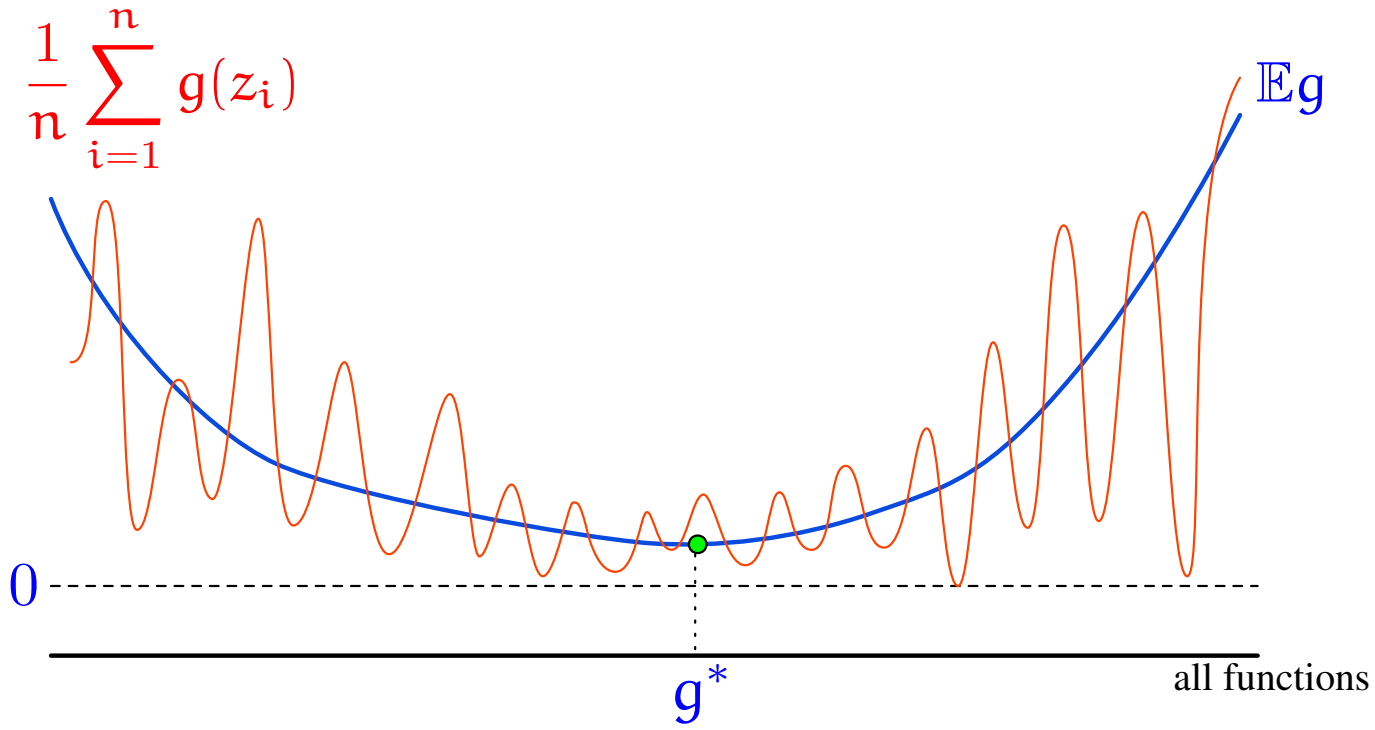
Next: visualize deviations $\mathbb{E}g(z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i)$ for all possible functions g and discuss all the concepts introduces so far.

Outline

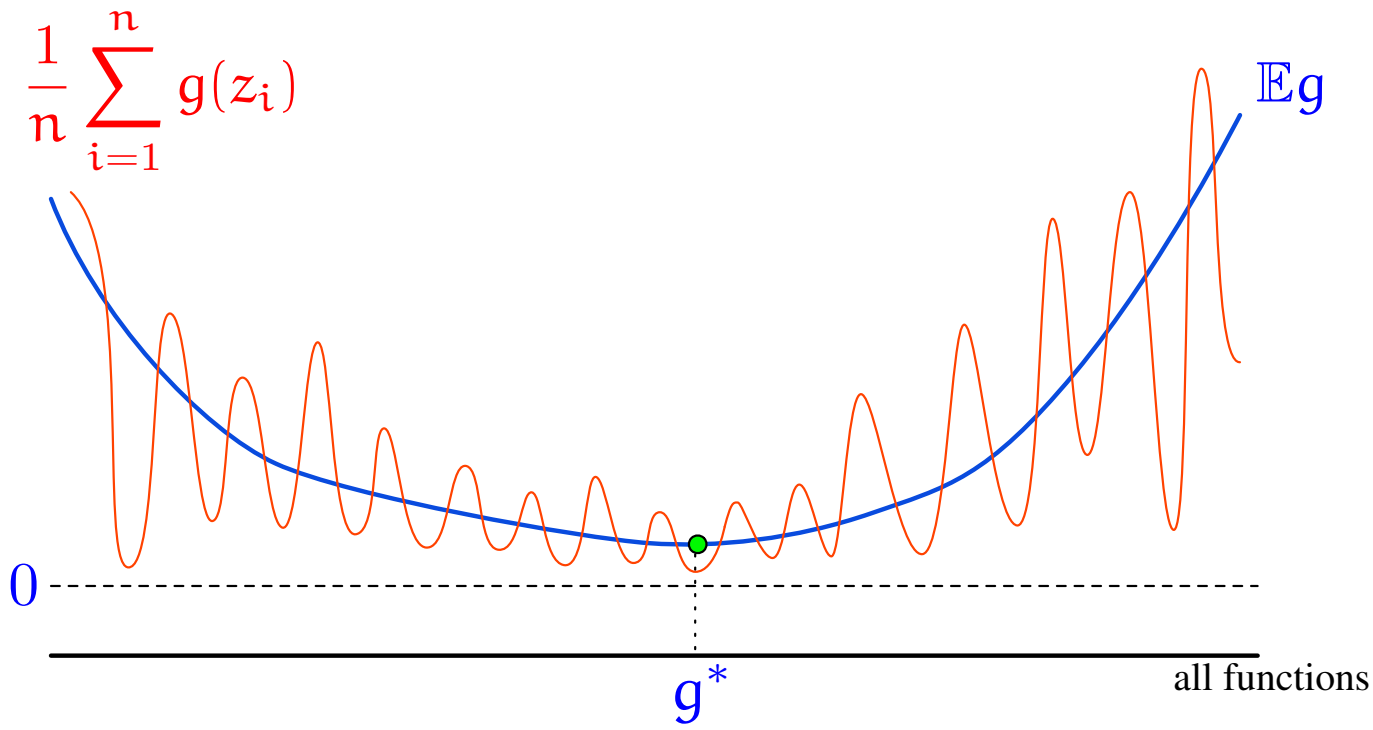
Empirical process viewpoint



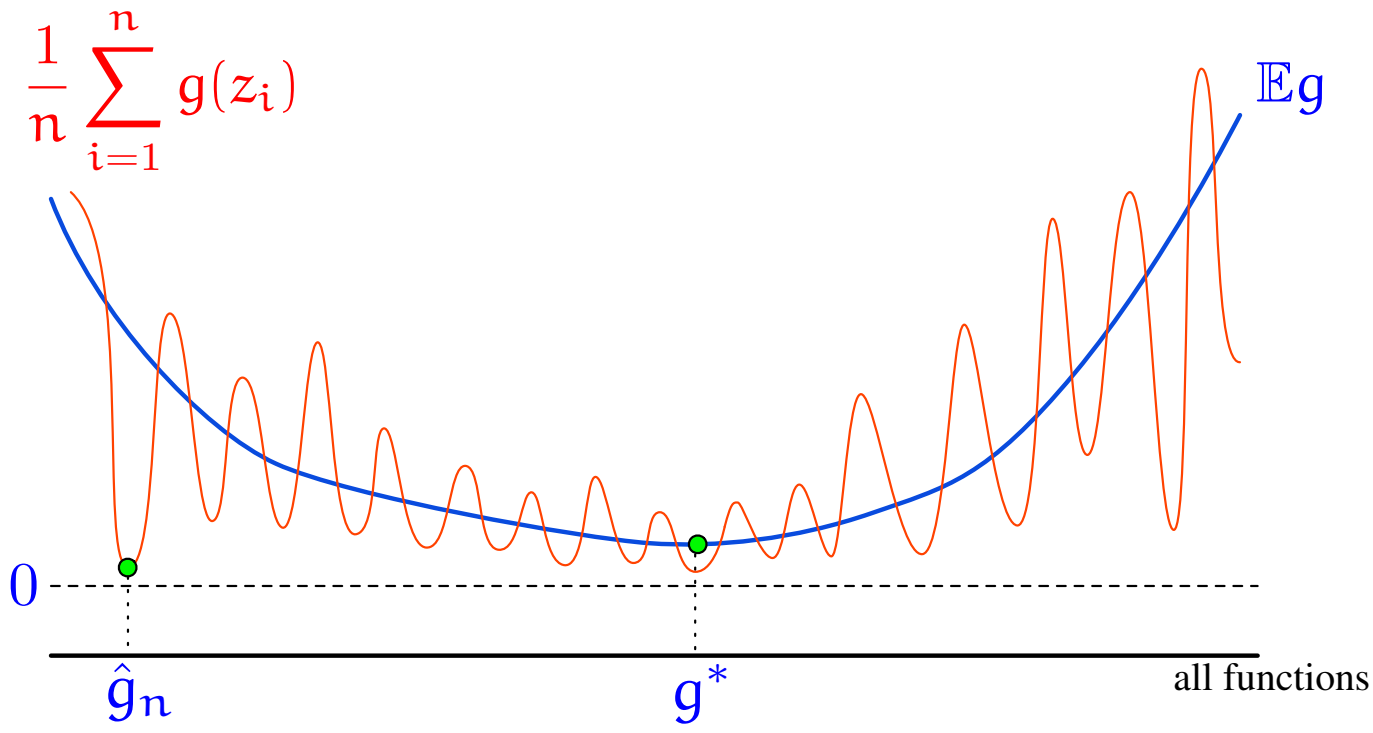
Empirical process viewpoint



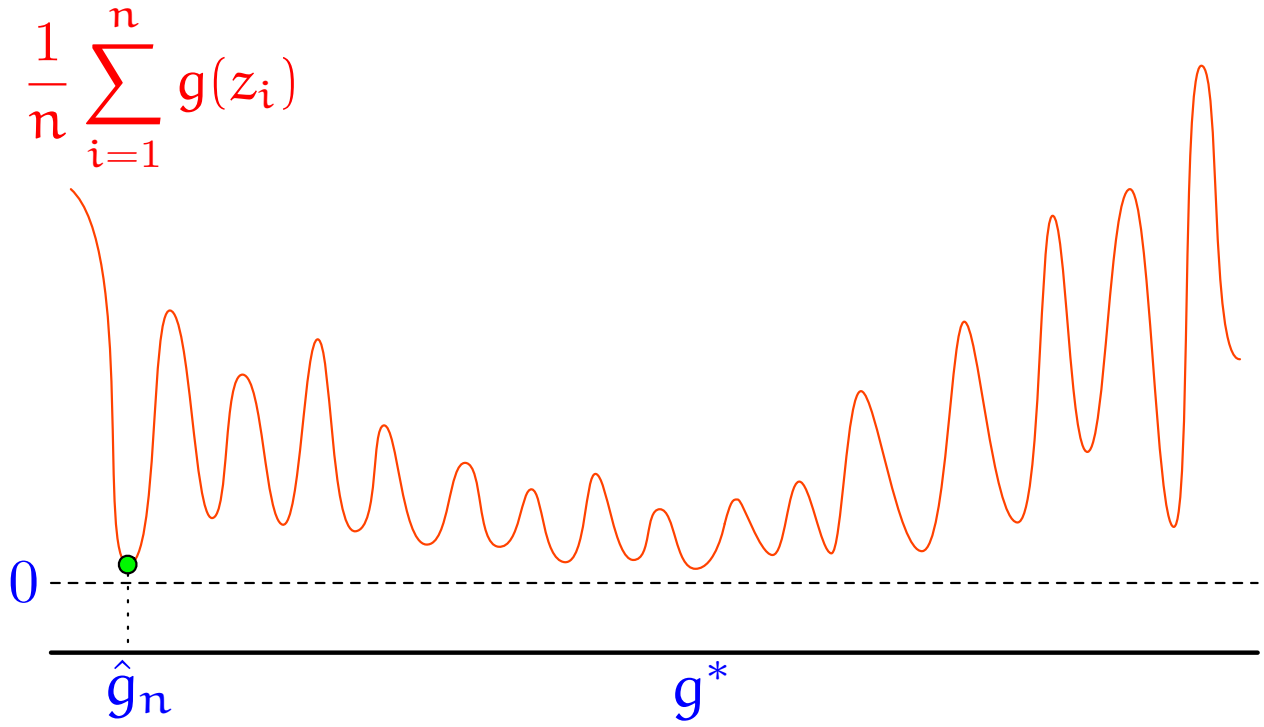
Empirical process viewpoint



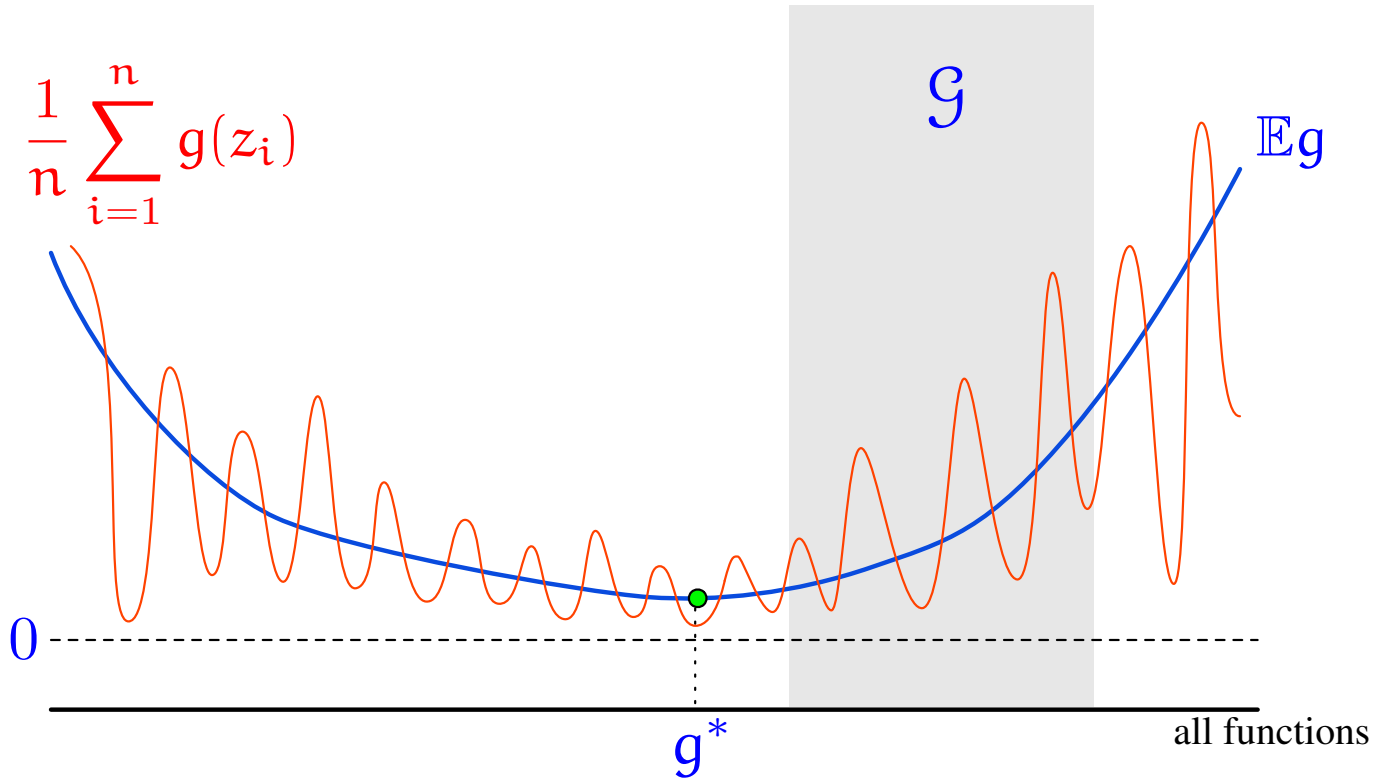
Empirical process viewpoint



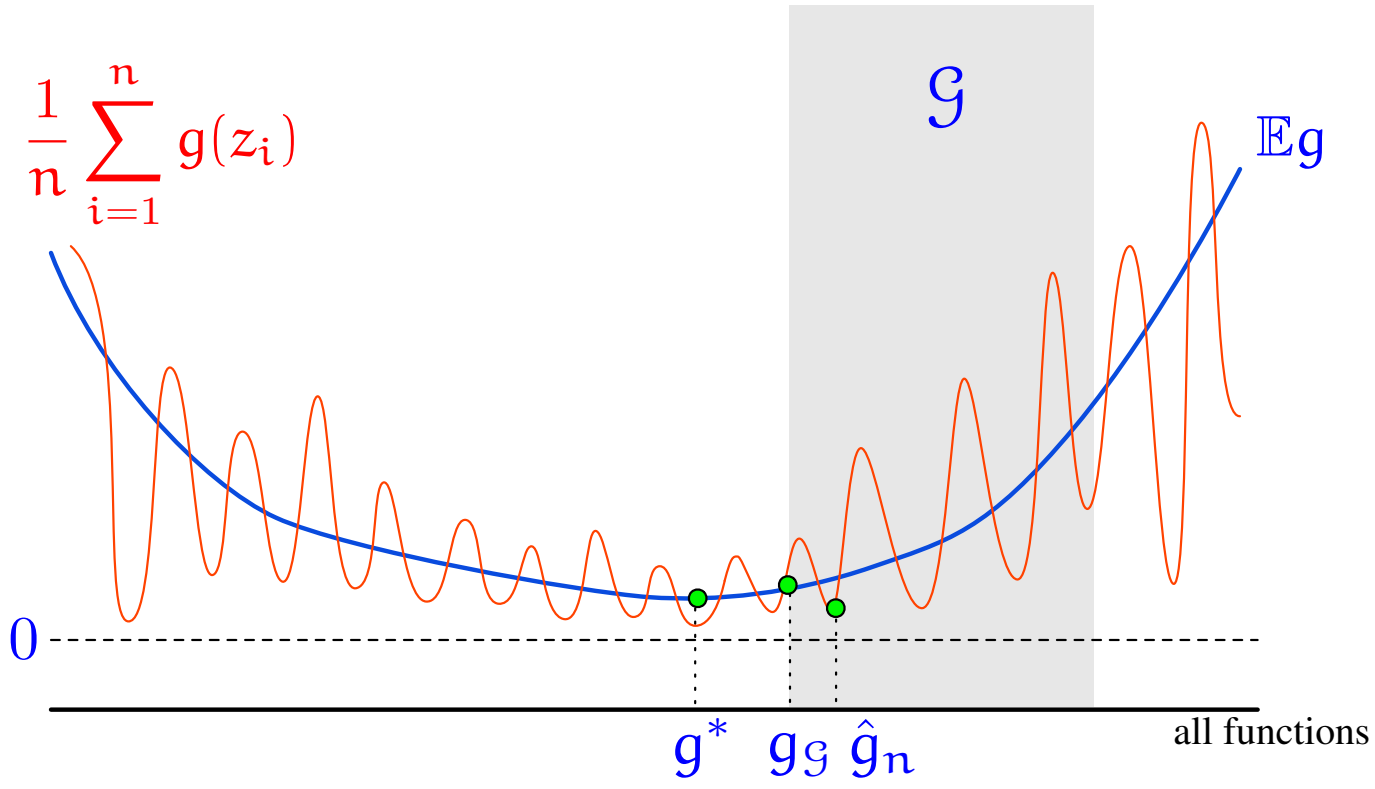
Empirical process viewpoint



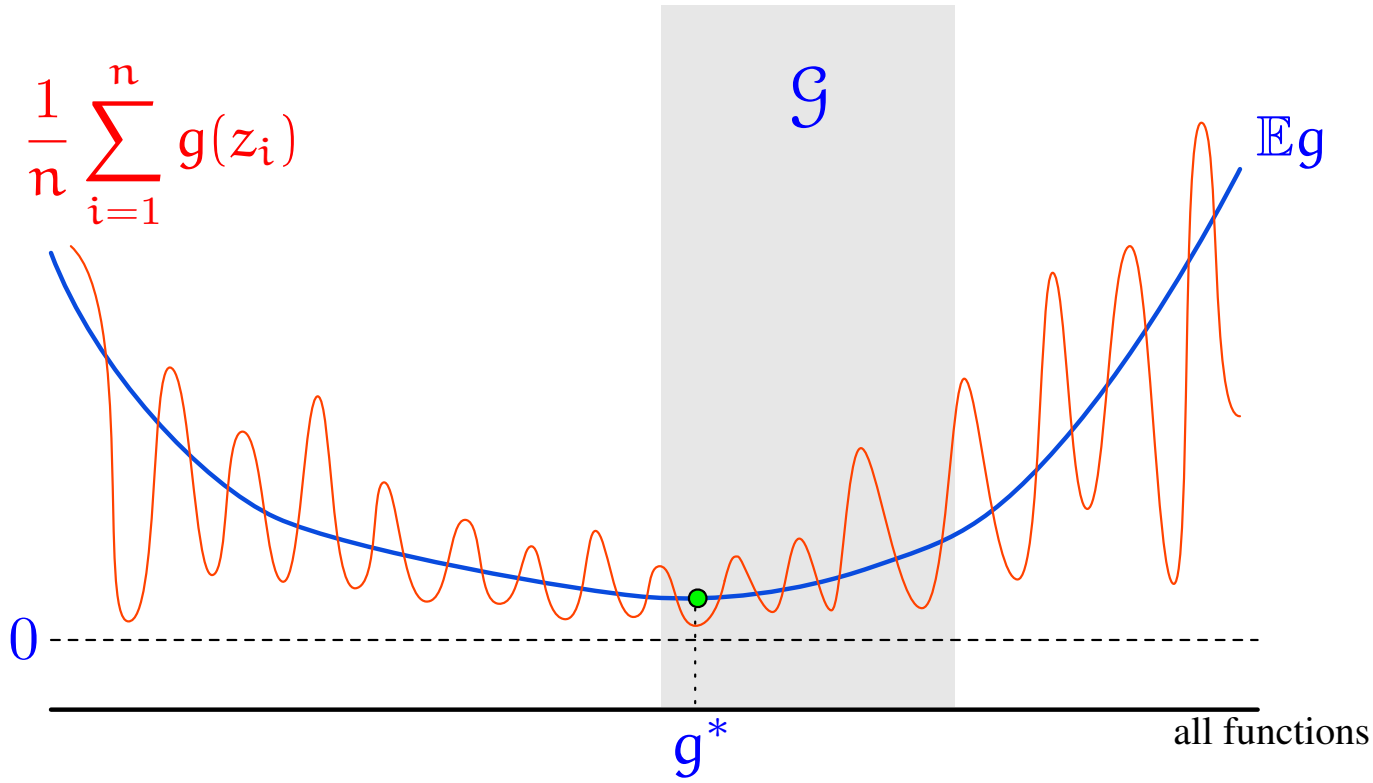
Empirical process viewpoint



Empirical process viewpoint



Empirical process viewpoint



Empirical process viewpoint

A *stochastic process* is a collection of random variables indexed by some set.

An *empirical process* is a stochastic process

$$\left\{ \mathbb{E}g(z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right\}_{g \in \mathcal{G}}$$

indexed by a function class $\mathcal{G}$.

(one-sided) *Uniform Law of Large Numbers:*

$$\sup_{g \in \mathcal{G}} \left\{ \mathbb{E}g - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right\} \to 0$$

in probability.

How does one quantify this rate of convergence when P is not known?

Rademacher process

Conditionally on $z_1, \ldots, z_n$, consider the Rademacher process

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(z_i) \right\}_{g \in \mathcal{G}}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. symmetric $\{\pm 1\}$-valued random vars.

Symmetrization (similar result holds for tails):

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\{ \mathbb{E} g - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right\} \leq 2 \mathbb{E} \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(z_i) \right\}$$

Rademacher process

The *empirical Rademacher averages* of $\mathcal{G}$ are defined as

$$\widehat{\mathscr{R}}_n(\mathcal{G}) = \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(z_i) \right\}$$

and let $\mathscr{R}(\mathcal{G}) = \mathbb{E}_{z_{1:n}} \widehat{\mathscr{R}}_n(\mathcal{G})$.

Why? Empirical Rademacher averages are a fully *data-dependent* quantity.

Maximal inequalities

A random variable X is ν-subgaussian if for any $\lambda \geq 0$,

$$\log \mathbb{E} \exp\{\lambda X\} \leq \nu\lambda^2/2.$$

If $X_1, \ldots, X_N$ are ν-subgaussian,

$$\mathbb{E} \max_i X_i \leq \sqrt{2\nu \log N}$$

Hoeffding: $a \leq X \leq b$ a.s. then $X - \mathbb{E}X$ is $(b-a)^2/4$-subgaussian.

First step: finite class

If $\mathcal{G} \subseteq [-1,1]^{\mathcal{Z}}$, $|\mathcal{G}| = N$, each $\sum_{i=1}^n \epsilon_i g(z_i)$ is n-subgaussian, $g \in \mathcal{G}$.

Thus,

$$\mathbb{E} \max_{g \in \mathcal{G}} \left\{ \sum_{i=1}^n \epsilon_i g(z_i) \right\} \leq \sqrt{2n \log N}.$$

In fact, a better bound is (prove!)

$$\mathbb{E} \max_{g \in \mathcal{G}} \left\{ \sum_{i=1}^n \epsilon_i g(z_i) \right\} \leq r\sqrt{2 \log N}, \quad r = \max_{g \in \mathcal{G}} \sqrt{\sum_{i=1}^n g(z_i)^2}$$

Empirical covering numbers

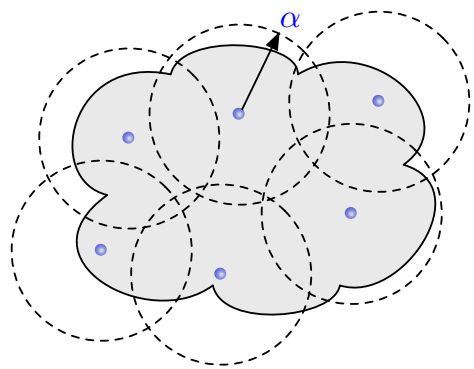
Projection

$$\mathcal{G}|_{z_{1:n}} = \mathcal{G}|_{z_1,\ldots,z_n} = \{(g(z_1),\ldots,g(z_n)) : g \in \mathcal{G}\} \subseteq \mathbb{R}^n$$

Can write empirical Rademacher averages as

$$\mathbb{E}_\epsilon \sup_{a \in \mathcal{G}|_{z_{1:n}}} \langle \epsilon, a \rangle$$

Very similar quantity: Gaussian widths (e.g. compressed sensing literature)



Given $\alpha > 0$, suppose we can find $V \subset \mathbb{R}^n$ of finite cardinality such that

$$\forall g, \exists v \in V, \quad \text{s.t.} \quad \frac{1}{n}\sum_{i=1}^{n} |g(z_i) - v_i|^p \leq \alpha^p$$

Empirical covering numbers

Such a set V is called an α-cover (or α-net) with respect to ℓ_p ($p \geq 1$). The size of the smallest α-cover is denoted by $\mathcal{N}_p(\mathcal{G}|_{z_{1:n}}, \alpha)$.

Using $p = 1$,

$$\begin{aligned}
\widehat{\mathscr{R}}_n(\mathcal{G}) &= \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(z_i) \\
&= \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (g(z_i) - v_i^g) + \mathbb{E}_{\epsilon_{1:n}} \sup_{g \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i v_i^g \\
&\leq \alpha + \mathbb{E}_\epsilon \max_{v \in V} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i v_i
\end{aligned}$$

Thus,

$$\widehat{\mathscr{R}}_n(\mathcal{G}) \leq \alpha + B\sqrt{\frac{2 \log \mathcal{N}_1(\mathcal{G}|_{z_{1:n}}, \alpha)}{n}}$$

where $B = \sup_f |f|_\infty$.

Chaining

Suppose $\mathcal{G} \subseteq [-1,1]^{\mathcal{Z}}$.

We have proved that conditionally on $z_1, \ldots, z_n$,

$$\widehat{\mathscr{R}}_n(\mathcal{G}) \leq \inf_{\alpha \geq 0} \left\{ \alpha + \frac{1}{\sqrt{n}} \sqrt{2 \log \mathcal{N}_1(\mathcal{G}|_{z_{1:n}}, \alpha)} \right\}$$

A better bound (called Dudley entropy integral):

$$\widehat{\mathscr{R}}_n(\mathcal{G}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\mathcal{G}|_{z_{1:n}}, \delta)} \, d\delta \right\}$$

Example: nondecreasing functions.

Consider the set $\mathcal{F}$ of nondecreasing functions $\mathbb{R} \to [-1, 1]$.

While $\mathcal{F}$ is a very large set, $\mathcal{F}|_{x_{1:n}}$ is not that large:

$$\mathcal{N}_1(\mathcal{F}|_{x_{1:n}}, \alpha) \le \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha) \le n^{2/\alpha}.$$

The first bound on the previous slide yields

$$\inf_{\alpha \ge 0} \left\{ \alpha + \frac{1}{\sqrt{\alpha n}} \sqrt{4 \log(n)} \right\} = \tilde{O}(n^{-1/3})$$

while the second bound (the Dudley entropy integral)

$$\inf_{\alpha \ge 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{4/\delta \log(n)} d\delta \right\} = \tilde{O}(n^{-1/2})$$

where the $\tilde{O}$ notation hides logarithmic factors.

Note: pointwise cover (w.r.t $d(f, g) = \sup_x |f(x) - g(x)|$) does not exist!

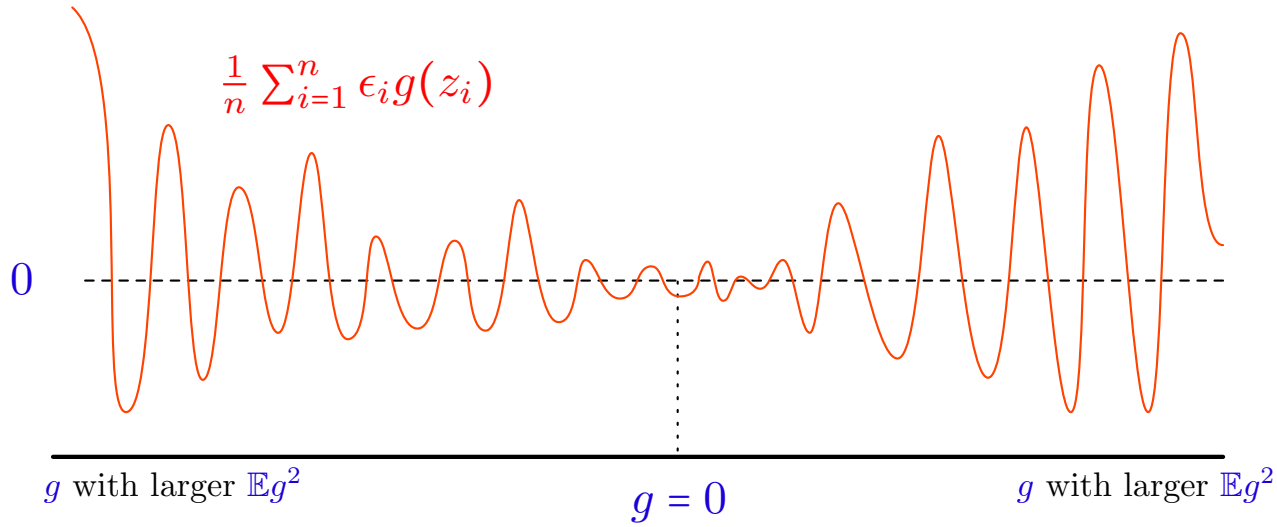
Offset Rademacher

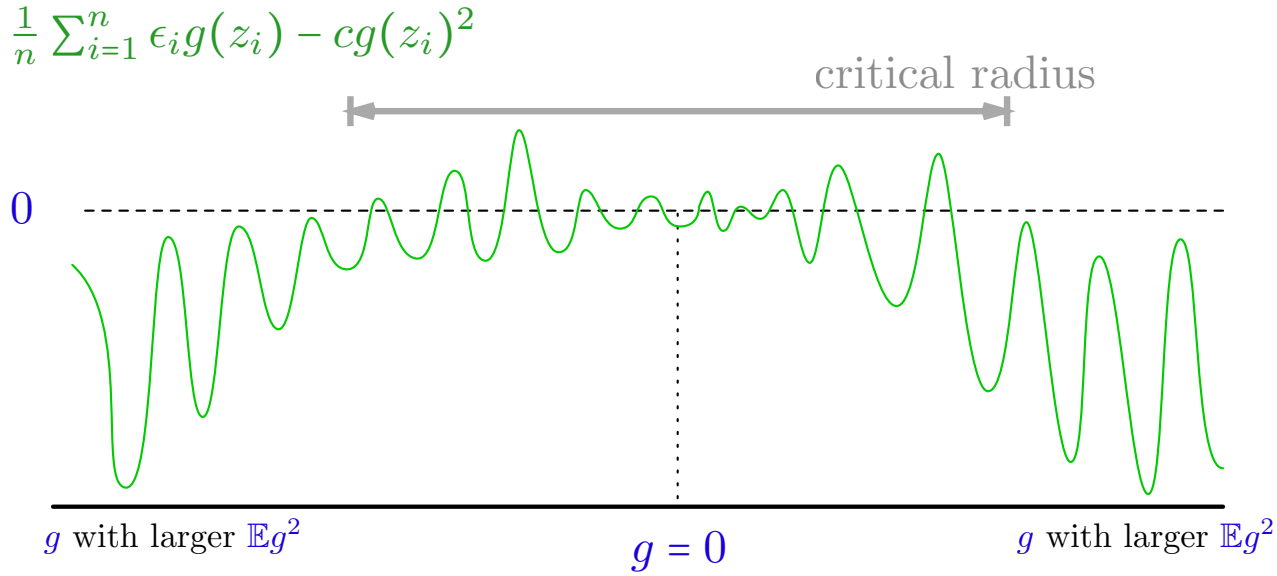
Offset Rademacher averages of $\mathcal{G}$ and constant $c \geq 0$ are defined as

$$\widehat{\mathscr{R}}_n^{\text{off}}(\mathcal{G}; c) = \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(z_i) - c g^2(z_i) \right\}$$

Empirical Rademacher averages correspond to $c = 0$.

Intuition


$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(z_i)$
0
g with larger $\mathbb{E}g^2$
$g = 0$
g with larger $\mathbb{E}g^2$


$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(z_i) - cg(z_i)^2$
critical radius
0
g with larger $\mathbb{E}g^2$
$g = 0$
g with larger $\mathbb{E}g^2$

Lemma.

Let $\mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$ be a finite class of cardinality N. Then for any $C > 0$,

$$\mathbb{E}_\epsilon \max_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(z_i) - C g(z_i)^2 \right] \leq \frac{1}{2C} \frac{\log N}{n}.$$

When the noise ξ is unbounded,

$$\mathbb{E}_\epsilon \max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \xi_i g(z_i) - C g(z_i)^2 \right] \leq M \cdot \frac{\log N}{n}$$

where

$$M := \max_{g \in \mathcal{G}} \frac{\sum_{i=1}^{n} g(z_i)^2 \xi_i^2}{2C \sum_{i=1}^{n} g(z_i)^2}.$$

(Liang, R., Sridharan '15)

Lemma (Localized Chaining).

Let $\mathcal{G}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}$. Then for any $z_1, \ldots, z_n \in \mathcal{Z}$

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_i g(z_i) - C g(z_i)^2 \right]$$
$$\leq \inf_{\gamma \geq 0, \alpha \in [0,\gamma]} \left\{ \frac{(2/C) \log \mathcal{N}_2(\mathcal{G}, \gamma)}{n} + 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)} d\delta \right\}$$

where $\mathcal{N}_2(\mathcal{G}, \gamma)$ is an ℓ_2-cover of $\mathcal{G}$ on $(z_1, \ldots, z_n)$ at scale γ (assumed to contain $\mathbf{0}$).

γ is an upper bound on critical radius.

(Liang, R., Sridharan '15)

Back-of-the-envelope calculation: if $\mathcal{N}_2(\mathcal{G}, \delta) \leq (1/\delta)^d$ (parametric class), then choosing $\gamma = 1/\sqrt{n}$ and $\alpha = 1/n$,

$$\inf_{\gamma \geq 0, \alpha \in [0,\gamma]} \left\{ \frac{(2/C)\log \mathcal{N}_2(\mathcal{G}, \gamma)}{n} + 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)}\, d\delta \right\}$$
$$\leq O\left(\frac{d \log n}{n}\right) + O\left(\frac{1}{\sqrt{n}} \int_{1/n}^{1/\sqrt{n}} \sqrt{d \log(1/\delta)}\, d\delta\right) = O\left(\frac{d \log n}{n}\right)$$

In contrast, the usual (non-offset) complexity will only give $n^{-1/2}$ rates.

Outline

The bottom line of the following slides in this lecture: the minimax rate for excess loss

- for loss functions without strong convexity (indicator loss, absolute loss) is given by Rademacher averages
- for square loss – by offset Rademacher averages.

Classification: loss function disappears

Consider binary classification with indicator loss, $\mathcal{F}$ a class of $\{0,1\}$-valued functions, and

$$\ell(f(x), y) = \mathbf{I}\{f(x) \neq y\} = (1-2y)f(x) + y.$$

Then

$$\begin{aligned}\widehat{\mathscr{R}}_n(\ell \circ \mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(x_i)(1-2y_i)+y_i)\right\} \,\middle|\, (x_1,y_1)\ldots,(x_n,y_n)\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right\} \,\middle|\, x_1,\ldots,x_n\right] = \widehat{\mathscr{R}}_n(\mathcal{F})\end{aligned}$$

because, given $y_1,\ldots,y_n$, the distribution of $\epsilon_i(1-2y_i)$ is the same as ϵ_i.

Absolute loss: disappears again

$\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and $\ell(f(x), y) = |f(x) - y|$.

Prove that

$$\widehat{\mathscr{R}}_n(\ell \circ \mathcal{F}) \leq \widehat{\mathscr{R}}_n(\mathcal{F})$$

This contraction (or, comparison inequality) holds for any Lipschitz loss. In fact, because there are no absolute values inside the supremum, the proof is easier than that in (Talagrand & Ledoux '91).

Lower bound

For indicator loss and for absolute loss, minimax rates in the distribution-free setting are given by Rademacher averages:

$$\bar{\mathcal{R}}^{\mathrm{iid}}(\mathcal{F}; n) \triangleq \sup_{x_1,\ldots,x_n} \widehat{\mathscr{R}}_n(\mathcal{F}; x_1, \ldots, x_n)$$

$$\inf_{\widehat{f}} \sup_{P_{X\times Y}} \left\{ \mathbb{E}|\widehat{f}(X) - Y| - \inf_{f\in\mathcal{F}} \mathbb{E}|f(X) - Y| \right\} \geq \bar{\mathcal{R}}^{\mathrm{iid}}(\mathcal{F}, 2n) - \frac{1}{2}\bar{\mathcal{R}}^{\mathrm{iid}}(\mathcal{F}, n)$$

Square Loss: no contraction, please!

If we attempt to pass to the Rademacher averages of $\widehat{\mathscr{R}}_n(\mathcal{F})$, the upper bound will be too loose (why? recall $O(\sigma^2 d/n)$ rate for classical regression)

Thankfully, offset Rademacher retains the curvature information. But we should not get to $\widehat{\mathscr{R}}_n(\ell \circ \mathcal{F})$ in the first place. We should come up with a method that directly gets us the offset Rademacher complexity.

ERM is the right algorithm for absolute/indictor loss. How about square loss?

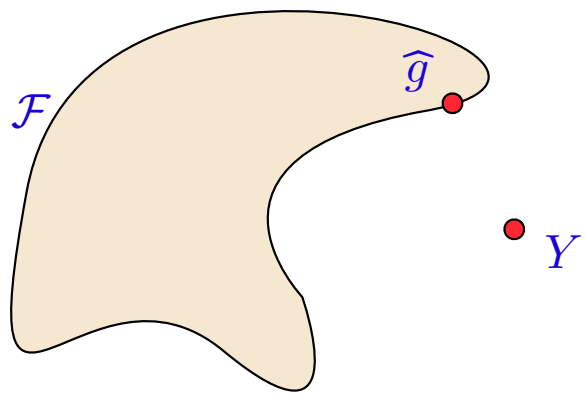
Turns out that for square (or other "curved") loss, ERM is suboptimal if $\mathcal{F}$ is non-convex.

Outline

The Star algorithm

$\widehat{\mathbb{E}}$ empirical expectation, $\text{star}(\mathcal{F}, g) = \{\lambda g + (1-\lambda) f : f \in \mathcal{F}, \lambda \in [0,1]\}$

$$\widehat{g} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \widehat{\mathbb{E}}(f(X) - Y)^2, \qquad \widehat{f} = \underset{f \in \text{star}(\mathcal{F}, \widehat{g})}{\text{argmin}}\, \widehat{\mathbb{E}}(f(X) - Y)^2$$

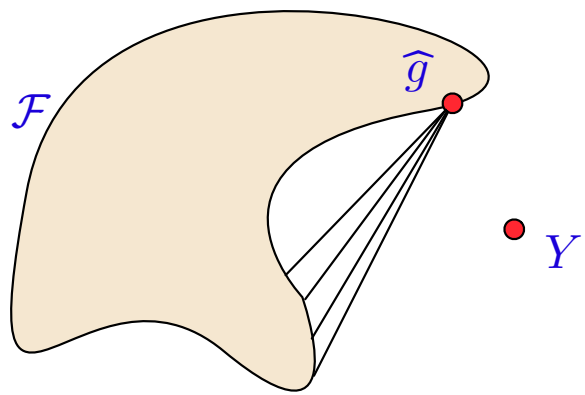


If $\mathcal{F}$ is convex, the Star algorithm coincides with ERM.

The Star algorithm

$\widehat{\mathbb{E}}$ empirical expectation, $\text{star}(\mathcal{F}, g) = \{\lambda g + (1-\lambda) f : f \in \mathcal{F}, \lambda \in [0,1]\}$

$$\widehat{g} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \widehat{\mathbb{E}}(f(X) - Y)^2, \qquad \widehat{f} = \underset{f \in \text{star}(\mathcal{F}, \widehat{g})}{\text{argmin}}\, \widehat{\mathbb{E}}(f(X) - Y)^2$$

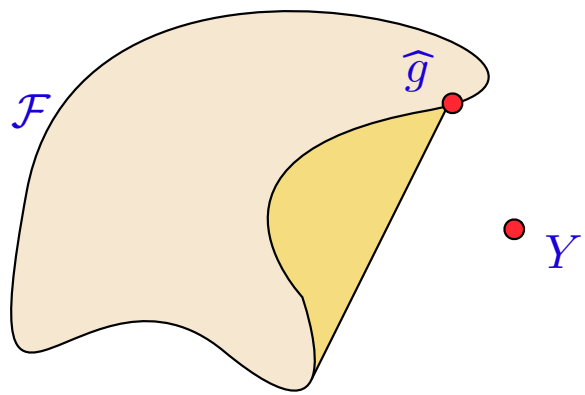


If $\mathcal{F}$ is convex, the Star algorithm coincides with ERM.

The Star algorithm

$\widehat{\mathbb{E}}$ empirical expectation, $\mathrm{star}(\mathcal{F}, g) = \{\lambda g + (1-\lambda) f : f \in \mathcal{F}, \lambda \in [0,1]\}$

$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mathbb{E}}(f(X) - Y)^2, \qquad \widehat{f} = \operatorname*{argmin}_{f \in \mathrm{star}(\mathcal{F}, \widehat{g})} \widehat{\mathbb{E}}(f(X) - Y)^2$$

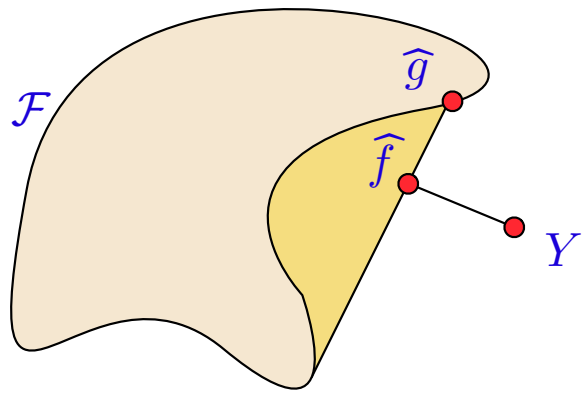


If $\mathcal{F}$ is convex, the Star algorithm coincides with ERM.

The Star algorithm

$\widehat{\mathbb{E}}$ empirical expectation, $\mathrm{star}(\mathcal{F}, g) = \{\lambda g + (1-\lambda) f : f \in \mathcal{F}, \lambda \in [0,1]\}$

$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mathbb{E}}(f(X) - Y)^2, \qquad \widehat{f} = \operatorname*{argmin}_{f \in \mathrm{star}(\mathcal{F}, \widehat{g})} \widehat{\mathbb{E}}(f(X) - Y)^2$$



If $\mathcal{F}$ is convex, the Star algorithm coincides with ERM.

Key geometric inequality

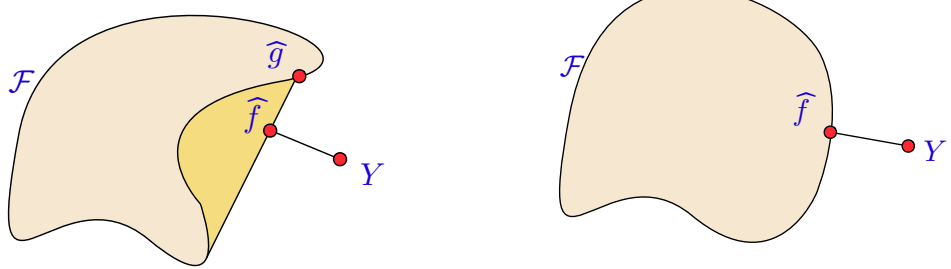
$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mathbb{E}}(f(X) - Y)^2, \qquad \widehat{f} = \operatorname*{argmin}_{f \in \mathrm{star}(\mathcal{F}, \widehat{g})} \widehat{\mathbb{E}}(f(X) - Y)^2$$

Lemma (Liang, R., Sridharan '15)**.**

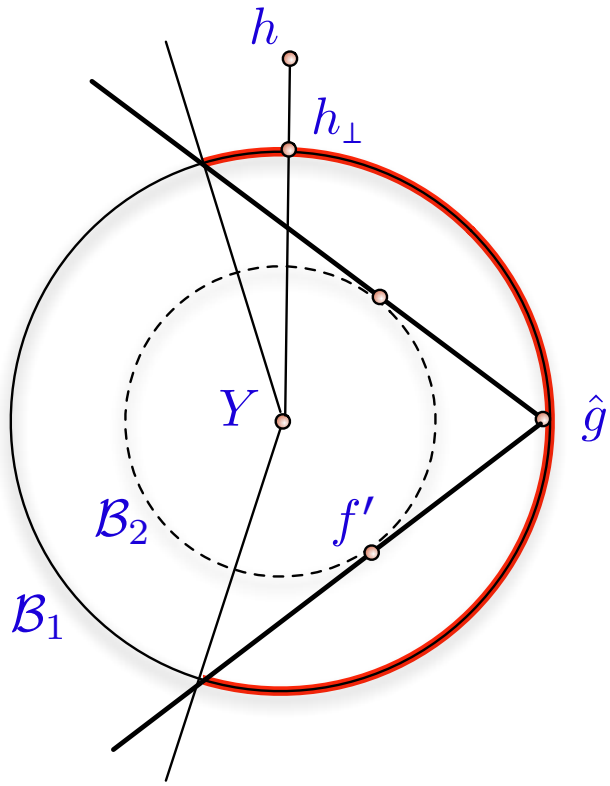
The Star algorithm $\widehat{f}$ satisfies

$$\widehat{\mathbb{E}}(h - Y)^2 - \widehat{\mathbb{E}}(\widehat{f} - Y)^2 \geq c \cdot \widehat{\mathbb{E}}(\widehat{f} - h)^2 \tag{1}$$

for any $h \in \mathcal{F}$ and $c = 1/18$. If $\mathcal{F}$ is convex, (1) holds with $c = 1$. If $\mathcal{F}$ is a linear subspace, (1) holds with equality and $c = 1$ by the Pythagorean theorem.



Proof of key geometric inequality



Corollary.

Excess loss of $\widehat{f}$ is upper bounded by

$$(\widehat{\mathbb{E}} - \mathbb{E})[2(f_{\mathcal{F}} - Y)(f_{\mathcal{F}} - \widehat{f})] + \mathbb{E}(f_{\mathcal{F}} - \widehat{f})^2 - (1 + c) \cdot \widehat{\mathbb{E}}(f_{\mathcal{F}} - \widehat{f})^2$$

Proof is immediate:

$$\begin{aligned}
&\mathbb{E}(\widehat{f} - Y)^2 - \mathbb{E}(f_{\mathcal{F}} - Y)^2 + \left[\widehat{\mathbb{E}}(f_{\mathcal{F}} - Y)^2 - \widehat{\mathbb{E}}(\widehat{f} - Y)^2 - c \cdot \widehat{\mathbb{E}}(\widehat{f} - f_{\mathcal{F}})^2\right] \\
&= (\widehat{\mathbb{E}} - \mathbb{E})[2(f_{\mathcal{F}} - Y)(f_{\mathcal{F}} - \widehat{f})] + \mathbb{E}(f_{\mathcal{F}} - \widehat{f})^2 - (1 + c) \cdot \widehat{\mathbb{E}}(f_{\mathcal{F}} - \widehat{f})^2.
\end{aligned}$$

The mismatch between coefficients 1 *and* $(1 + c)$ *allows to perform symmetrization and get offset Rademacher complexity as an upper bound on excess loss of* $\widehat{f}$.

Theorem (Liang, R., Sridharan, 15).

Define $\mathcal{H} := \mathcal{F} - f_{\mathcal{F}} + \text{star}(\mathcal{F} - \mathcal{F}, 0)$. The following expectation bound on excess loss of the Star estimator holds:

$$c'' \cdot \mathbb{E}\widehat{\mathscr{R}}_n^{\text{off}}(\mathcal{H}; c')$$

where $c' = \min\{\frac{c}{4M}, \frac{c}{4K(2+c)}\}$, $K = \sup_f |f|_\infty$, $M = \sup_f |Y - f|_\infty$, and $c'' = (2M + K(2+c)/2)$.

- A similar bound in terms of offset Rademacher holds in high probability and without the boundedness assumption (but under a weak lower isometry assumption).
- Easy to show that complexity of $\mathcal{H}$ is of same order as that of $\mathcal{F}$ (except for finite $\mathcal{F}$).

Example: ordinary least squares

$\mathcal{G} = \{x \mapsto w^T x : w \in \mathbb{R}^p\}$. Offset Rademacher becomes

$$\frac{1}{n} \sup_{w \in \mathbb{R}^p} \left\{ w^T \left(\sum_{t=1}^n \epsilon_t x_t \right) - c\|w\|_\Sigma^2 \right\} = \frac{c'}{n} \left\| \sum_{t=1}^n \epsilon_t x_t \right\|_{\Sigma^{-1}}^2$$

where $\Sigma = \sum_{t=1}^n x_t x_t^T$.

Example: OLS, a more precise statement

Lemma.

Consider parametric regression $Y_i = X_i^T \beta^* + \xi_i, 1 \le i \le n$, where ξ_i need not be centered. The offset Rademacher complexity is bounded as

$$\mathbb{E}_\epsilon \sup_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i \xi_i X_i^T \beta - C\beta^T X_i X_i^T \beta \right\} = \frac{\mathrm{tr}\left(G^{-1} H\right)}{Cn}$$

where $G := \sum_{i=1}^n X_i X_i^T$ and $H = \sum_{i=1}^n \xi_i^2 X_i X_i^T$. In well-specified case (ξ_i are zero-mean), assuming that conditional var is σ^2, then conditionally on the design, $\mathbb{E} G^{-1} H = \sigma^2 I_p$ and excess loss is order $\frac{\sigma^2 p}{n}$.

A high-probability statement holds as well.

High probability statement for unbounded functions

We say that a function class $\mathcal{F}$ satisfies the lower isometry bound with parameters $0 < \epsilon < 1$ and $0 < \delta < 1$ if

$$\mathbb{P}\left(\inf_{f \in \mathcal{F} \setminus \{0\}} \frac{1}{n} \sum_{i=1}^{n} \frac{f^2(X_i)}{\mathbb{E}f^2} \geq 1 - \epsilon \right) \geq 1 - \delta$$

for all $n \geq n_0(\mathcal{F}, \delta, \epsilon)$, where $n_0(\mathcal{F}, \delta, \epsilon)$ depends on the complexity of the class.

This holds under small ball assumption of Mendelson + norm comparison (e.g. $\|f\|_q \leq L\|f\|_2$ for all $f \in \mathcal{F}$). It also holds for subgaussian classes.

High probability statement for unbounded functions

Theorem (Liang, R., Sridharan, 15).

$\mathcal{H} := \mathrm{star}\left(\mathcal{F} - f^* + \mathrm{star}(\mathcal{F} - \mathcal{F})\right)$. Assume lower isometry holds with $\epsilon = 1/72$. Let $\xi_i = Y_i - f^*(X_i)$.

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}) > 4u\right) \le 4\delta + 4\mathbb{P}\left(\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \xi_i h(X_i) - \tilde{c} h(X_i)^2 > u\right)$$

for any $u > 0$, as long as

$$n > \sup_{h \in \mathcal{H}} \frac{2 \cdot \mathrm{Var}\left[2\xi h + (1 - c') \cdot h^2\right]}{\left[c' \cdot \mathbb{E}h^2\right]^2} \vee n_0(\mathcal{H}, \delta, c/4).$$

Example: nonparametric function classes

For many nonparametric classes, we can compute an estimate of empirical entropy. Suppose an upper bound is of the form

$$\log \mathcal{N}_2(\mathcal{F}|_{x_1,\ldots,x_n}, \alpha) \leq \alpha^{-p}$$

By plugging this into the bound for localized chaining, we obtain rate $n^{-\frac{2}{2+p}}$ for $p \in (0,2)$, $n^{-1/p}$ for $p > 2$, and $n^{-1/2}\log(n)$ at $p = 2$.

This gives an upper bound on excess square loss (and, hence, estimation in the sense of oracle inequalities).

One can show that for well-specified models, transition at $p = 2$ does not happen, and the rate remains $n^{-\frac{2}{2+p}}$. For instance, for estimation of bounded convex functions on $[0,1]^d$, it has been shown that $p = d/2$.

Outline

Exercise 1

For $A \subseteq [-1,1]^n$ define Rademacher averages of A as

$$\mathscr{R}(A) = \mathbb{E}_\epsilon \sup_{a \in A} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t a_t$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. ± 1 Rademacher random variables.

Prove that for any $r_1, \ldots, r_n \in [0,1]$,

$$\mathbb{E} \sup_{a \in A} \sum_{t=1}^{n} \epsilon_t r_t a_t \leq \mathbb{E} \sup_{a \in A} \sum_{t=1}^{n} \epsilon_t a_t$$

Exercise 2

Define Gaussian averages of A as

$$G(A) = \mathbb{E} \sup_{a \in A} \frac{1}{n} \sum_{t=1}^{n} \gamma_t a_t$$

where $\gamma_1, \ldots, \gamma_n$ are independent $N(0,1)$. Show that

$$c\mathscr{R}(A) \leq G(A) \leq C\sqrt{\log(n)}\mathscr{R}(A)$$

and find explicit constants c, C.

Exercise 3

Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz. Prove that

$$\mathbb{E} \sup_{a \in A} \sum_{t=1}^{n} \epsilon_t \phi(a_t) \leq L \mathbb{E} \sup_{a \in A} \sum_{t=1}^{n} \epsilon_t a_t$$

Hint: condition on all but one ϵ_t, write out the two possibilities for ϵ_t, and combine the suprema. Make sure the argument does not leave any absolute values.

Exercise 4

Prove that for a finite collection $A \subset \mathbb{R}^n$ and any $c > 0$,

$$\mathbb{E} \max_{a \in A} \left\{ \sum_{t=1}^{n} \epsilon_t a_t - c a_t^2 \right\} \leq C \log |A|$$

for some C that does not depend on the magnitude of vectors in A.

Hint: write out the moment-generating function and use $(e^{-x} + e^{x})/2 \leq e^{x^2/2}$.

Exercise 5

We argued in the lecture that for a finite collection $A \subset [-1,1]^n$,

$$\mathbb{E} \max_{a \in A} \sum_{t=1}^{n} \epsilon_t a_t \le r\sqrt{2 \log N}, \quad r = \max_{a \in A} \|a\|_2$$

Now suppose B is a set of predictable processes with respect to $\{\mathcal{F}_t = \sigma(\epsilon_1, \ldots, \epsilon_t)\}_{t=0}^{n}$. That is, each $\mathbf{b} \in B$ is a sequence $\mathbf{b}_1, \ldots, \mathbf{b}_n$ where each $\mathbf{b}_t$ is $\mathcal{F}_{t-1}$-measurable. Prove that

$$\mathbb{E} \max_{\mathbf{b} \in B} \sum_{t=1}^{n} \epsilon_t \mathbf{b}_t \le r\sqrt{2 \log N}, \quad r = \max_{\epsilon \in \{\pm 1\}^n} \max_{\mathbf{b} \in B} \sqrt{\sum_{t=1}^{n} \mathbf{b}_t^2}.$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$.

Hint: Consider the moment generating function and peel off one term at a time, from n backwards to $t = 1$.

Exercise 6

Let W be a random variable with values in $\mathcal{A}$. Prove that for a measurable function $\Psi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$,

$$\mathbb{E}_W \sup_{b \in \mathcal{B}} \Psi(W, b) = \sup_{\gamma} \mathbb{E}_W \Psi(W, \gamma(W))$$

where the supremum ranges over all functions $\gamma : \mathcal{A} \to \mathcal{B}$. (Assume compactness or boundedness if needed to make the argument rigorous).

Exercise 7

Let $\epsilon_{1:n} \triangleq (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ be n i.i.d. Rademacher random variables. Use the previous exercise to conclude that for $\Psi : \mathcal{X}^n \times \{\pm 1\}^n \to \mathbb{R}$,

$$\sup_{x_1 \in \mathcal{X}} \mathbb{E}_{\epsilon_1} \ldots \sup_{x_n \in \mathcal{X}} \mathbb{E}_{\epsilon_n} \Psi(x_{1:n}, \epsilon_{1:n}) = \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_n} \mathbb{E}_{\epsilon_{1:n}} \Psi(\mathbf{x}_1, \mathbf{x}_2(\epsilon_1) \ldots, \mathbf{x}_n(\epsilon_{1:n-1}), \epsilon_{1:n})$$

where the last supremum is taken over functions $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$.

Exercise 8

Let Q be the set of distributions on some set $\mathcal{A}$ and P the set of distributions on $\mathcal{B}$. Under very general conditions on $\ell, \mathcal{A}, \mathcal{B}$,

$$\min_{q \in Q} \max_{b \in \mathcal{B}} \mathbb{E}_{a \sim q} \ell(a, b) = \max_{p \in P} \min_{a \in \mathcal{A}} \mathbb{E}_{b \sim p} \ell(a, b).$$

This is known as the minimax theorem. Note that the inner max/min can be taken at a pure strategy (delta distribution) because a linear function achieves its max/min at a corner of the probability simplex.

Prove the following: if $\ell(a, b)$ is convex in a and $\mathcal{A}$ is a convex set, then the outer minimization

$$\min_{q \in Q} \max_{b \in \mathcal{B}} \mathbb{E}_{a \sim q} \ell(a, b) = \min_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \ell(a, b)$$

is achieved at a pure strategy. We will use this result to restrict our attention to deterministic strategies.

Exercise 9

Let W be a random variable, and suppose that for any realization of W,

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} \{\ell(a, b) + \Psi_t(b, W)\} \leq \Psi_{t-1}(W)$$

Prove that

$$\inf_{q \in \Delta(\mathcal{A})} \sup_{b \in \mathcal{B}} \{\mathbb{E}_{a \sim q} \ell(a, b) + \mathbb{E}_W \Psi_t(b, W)\} \leq \mathbb{E}_W \Psi_{t-1}(W)$$

by exhibiting a strategy for the infimum. This statement will be useful for defining computationally-efficient random playout methods in Lecture #3.

Exercise 10

Consider the following online prediction problem, taking place over rounds $t = 1, \dots, n$. On each round, we make a prediction $\widehat{y}_t \in [0,1]$, observe an outcome $y_t \in \{0,1\}$, and suffer the loss of $\ell(\widehat{y}_t, y_t) = y_t + \widehat{y}_t - 2\widehat{y}_t \cdot y_t$. Take a potential function $\Phi : \{\pm 1\}^n \to \mathbb{R}$ with two properties: first, it is stable with respect to flip of any coordinate:

$$|\Phi(\dots, -1, \dots) - \Phi(\dots, +1, \dots)| \le 1.$$

Second, $\mathbb{E}\Phi(b_1, \dots, b_n) \ge n/2$ where b_i's are i.i.d. Bernoulli with bias $1/2$. Show that

$$\min_{\widehat{y}_t} \max_{y_t} \left\{ \ell(\widehat{y}_t, y_t) + \mathbb{E}_{b_{t+1:n}} \Phi(y_1, \dots, y_t, b_{t+1}, \dots, b_n) \right\} \le \mathbb{E}_{b_{t:n}} \Phi(y_1, \dots, y_{t-1}, b_t, \dots, b_n) + \frac{1}{2}$$

Conclude that there is a prediction strategy that guarantees

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) \le \Phi(y_1, \dots, y_n) \tag{2}$$

for any sequence $y_1, \dots, y_n$ of binary outcomes. Conversely, argue that if there is a function Φ that satisfies (2) for all sequences, then it must hold that $\mathbb{E}\Phi \ge n/2$.

Exercise 11

Write the loss function in the previous exercise as expected indicator loss under the randomized strategy with bias $\widehat{y}_t$. Use the previous exercise to argue that there must exist a randomized algorithm that predicts an arbitrary sequence of bits with the following strong guarantee:

the expected average number of mistakes (per n rounds) is at most the ***minimum*** *of proportion of 1's and proportion of 0's in the sequence, up to a $O(1/\sqrt{n})$ additive factor.*

That is, if the sequence, say, has 40% of 0's, then the method will only err roughly 40% of the time, even though the locations of 0's. The method is adaptive: it does not need to know any prior information about the sequence. This result might seem surprising, given that the sequence is not governed by any stochastic process that we can describe.

Outline

Outline

Last lecture: $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_X \times P_{Y|X}$ i.i.d.

Objective: excess loss

$$\mathbb{E}\ell(\widehat{f}(X), Y) - \inf_{f \in \mathcal{F}} \mathbb{E}\ell(f(X), Y)$$

We showed that this quantity is controlled by Rademacher averages or by offset Rademacher averages.

Consider a time-averaged variant

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\ell(\widehat{f}_t(X),Y)-\inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\ell(f(X),Y)$$

where $\widehat{f}_t:(\mathcal{X}\times\mathcal{Y})^{t-1}\to\mathcal{Y}^{\mathcal{X}}$ is calculated based on $(X_1,Y_1),\ldots,(X_{t-1},Y_{t-1})$.

Since data are i.i.d. we can write this equivalently as

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{f}_t(X_t),Y_t)\right]-\inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\ell(f(X_t),Y_t)$$

Via Jensen's, a harder objective is

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{f}_t(X_t),Y_t)-\inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(X_t),Y_t)\right]$$

It's time to discuss the online protocol:

1. At time t, compute $\widehat{f}_t$ based on $(X_1,Y_1),\ldots,(X_{t-1},Y_{t-1})$
2. Observe (X_t,Y_t)
3. Pay loss $\ell(\widehat{f}_t(X_t),Y_t)$

Or, equivalently (from the point of view of the objective),

1. At time t, observe X_t
2. Compute $\widehat{y}_t$ based on $(X_1,Y_1),\ldots,(X_{t-1},Y_{t-1})$ and X_t
3. Observe Y_t
4. Pay loss $\ell(\widehat{y}_t,Y_t)$

A leap

Our objective (which we shall call *regret*) is

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{y}_t, Y_t) - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(X_t), Y_t)\right]$$

We now make a step that might be hard to digest.

> We assume that Y's are not i.i.d. Moreover, *they do not come from a stochastic process that is easy to describe.*

We analyze this case by making two simplifying assumptions:

- each $Y_t \in [-1, 1]$ (this boundedness assumption can be removed)
- we know P_X (we will remove this assumption in the next lecture)

An optimization problem?

Wald's decision theory: min over decision rules, max over problems. Not possible to solve this in general.

Here, we only need to compute one number per time step. Let us try to solve for it!

On round t we observe X_t and need to choose $\widehat{y}_t$. Here is the optimal choice:

$$\operatorname*{argmin}_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ \ell(\widehat{y}_t, y_t) + \mathrm{OPT}(\mathrm{past}, y_t) \right\}$$

where $\mathrm{OPT}(\mathrm{past}, y_t)$ is the "future" minimax value of regret with $\sum_{s=1}^{t} \ell(\widehat{y}_t, y_t)$ removed. This future depends on the choice of y_t.

Recursive definition, and not clear how to solve it. It definitely does not look like ERM or Star algorithm!

Dynamic programming

Suppose we can find a function $\mathbf{Rel} : \cup_{t=0}^{n} (\mathcal{X} \times \mathcal{Y})^t \to \mathbb{R}$ satisfying these two conditions:

1. For any $x_1, y_1, \ldots, x_n, y_n$,

$$\mathbf{Rel}(x_1, y_1, \ldots, x_n, y_n) \geq -\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)$$

2. For any $x_1, y_1, \ldots, x_{t-1}, y_{t-1}$,

$$\mathbb{E}_{x_t} \inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ \ell(\widehat{y}_t, y_t) + \mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) \right\} \leq \mathbf{Rel}(x_1, y_1, \ldots, x_{t-1}, y_{t-1})$$

A relaxation satisfying these conditions will be called *admissible*.

If $\ell(\cdot, y)$ is not convex, we need an extra expectation for mixed strategies. But for now assume it is convex.

Lemma.

If **Rel** is admissible, the algorithm

$$\widehat{y}_t = \operatorname*{argmin}_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ \ell(\widehat{y}_t, y_t) + \mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) \right\}$$

has regret bound of

$$\mathbb{E}\left[\frac{1}{n} \sum_{t=1}^{n} \ell(\widehat{y}_t, Y_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(X_t), Y_t) \right] \leq \frac{1}{n} \mathbb{E}\mathbf{Rel}(\varnothing)$$

(R., Shamir, Sridharan, '12)

Lemma.

Let $\ell(\widehat{y}, y) = |\widehat{y} - y|$. Rademacher-based relaxation

$$\mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) = \mathop{\mathbb{E}}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t} |f(x_s) - y_s| \right\}$$

is admissible.

Hence, expected regret is upper bounded by

$$\frac{1}{n}\mathbb{E}\mathbf{Rel}(\varnothing) = \mathbb{E}_{x,\epsilon} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{s=1}^{n} 2\epsilon_s f(x_s) \right\} = 2\mathbb{E}\mathscr{R}_n(\mathcal{F}),$$

the i.i.d. Rademacher averages.

see e.g. (R. and Sridharan, '15)

We removed the assumption that Y's are i.i.d. and still obtained a bound of Rademacher complexity, as in the previous lecture!

The proposed algorithm needs to solve an optimization problem involving **Rel** and so it needs to approximate this value. We assumed that P_X is known, but one can sample or use unlabeled data (we will discuss this in Lecture #3)

Why is

$$\mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) = \mathop{\mathbb{E}}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t} |f(x_s) - y_s| \right\}$$

admissible? Recall that $-\inf[\ldots]$ is same as $\sup -[\ldots]$. Relaxation interpolates ($t = n$ to $t = 1$) between

$$\mathbf{Rel}(x_1, y_1, \ldots, x_n, y_n) = -\inf_{f \in \mathcal{F}} \sum_{s=1}^{n} |f(x_s) - y_s|$$

and

$$\mathbf{Rel}(\varnothing) = \mathbb{E}_{x,\epsilon} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=1}^{n} 2\epsilon_s f(x_s) \right\}$$

However, this observation is not enough for admissibility. Need to show that change of potential function from t to $t-1$ is related to loss on that step (condition #2).

Notation: $\Delta = \Delta([-1, 1])$, $L_t(f) = \sum_{s=1}^{t} |f(x_s) - y_s|$, $A_{t+1}(f) = \sum_{s=t+1}^{n} 2\epsilon_s f(x_s)$

Proof of admissibility (it's not as scary as it looks, really!)

$$\inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$\leq \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ A_t(f) - L_{t-1}(f) \right\}$$

Proof of admissibility (it's not as scary as it looks, really!)

$$\inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$\leq \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ A_t(f) - L_{t-1}(f) \right\}$$

Proof of admissibility (it's not as scary as it looks, really!)

$$\inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \left\{ \inf_{\widehat{y}_t} \mathbb{E}_{y'_t} |\widehat{y}_t - y'_t| + \mathbb{E}_{x_{t+1:n}, y_t, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$\leq \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ A_t(f) - L_{t-1}(f) \right\}$$

Proof of admissibility (it's not as scary as it looks, really!)

$$\inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \left\{ \inf_{\widehat{y}_t} \mathbb{E}_{y'_t} |\widehat{y}_t - y'_t| + \mathbb{E}_{x_{t+1:n}, y_t, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}, y_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + \mathbb{E}_{y'_t} |f(x_t) - y'_t| - |f(x_t) - y_t| \right\}$$

$$\leq \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ A_t(f) - L_{t-1}(f) \right\}$$

Proof of admissibility (it's not as scary as it looks, really!)

$$\inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \left\{ \inf_{\widehat{y}_t} \mathbb{E}_{y'_t} |\widehat{y}_t - y'_t| + \mathbb{E}_{x_{t+1:n}, y_t, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}, y_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + \mathbb{E}_{y'_t} |f(x_t) - y'_t| - |f(x_t) - y_t| \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}, y_t, y'_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + \epsilon_t (|f(x_t) - y'_t| - |f(x_t) - y_t|) \right\}$$

$$\leq \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ A_t(f) - L_{t-1}(f) \right\}$$

Proof of admissibility (it's not as scary as it looks, really!)

$$\inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \left\{ \inf_{\widehat{y}_t} \mathbb{E}_{y'_t} |\widehat{y}_t - y'_t| + \mathbb{E}_{x_{t+1:n}, y_t, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}, y_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + \mathbb{E}_{y'_t} |f(x_t) - y'_t| - |f(x_t) - y_t| \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}, y_t, y'_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + \epsilon_t (|f(x_t) - y'_t| - |f(x_t) - y_t|) \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}, y_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + 2\epsilon_t |f(x_t) - y_t| \right\}$$

$$\leq \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ A_t(f) - L_{t-1}(f) \right\}$$

Proof of admissibility (it's not as scary as it looks, really!)

$$\inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$= \sup_{p_t \in \Delta} \left\{ \inf_{\widehat{y}_t} \mathbb{E}_{y'_t} |\widehat{y}_t - y'_t| + \mathbb{E}_{x_{t+1:n}, y_t, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_t(f) \right\} \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}, y_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + \mathbb{E}_{y'_t} |f(x_t) - y'_t| - |f(x_t) - y_t| \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}, y_t, y'_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + \epsilon_t (|f(x_t) - y'_t| - |f(x_t) - y_t|) \right\}$$

$$\leq \sup_{p_t \in \Delta} \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}, y_t} \sup_{f \in \mathcal{F}} \left\{ A_{t+1}(f) - L_{t-1}(f) + 2\epsilon_t |f(x_t) - y_t| \right\}$$

$$\leq \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ A_t(f) - L_{t-1}(f) \right\}$$

Taking expectation w.r.t x_t on both sides completes the proof.

In above, first step is minimax theorem. Third replaces inf with $f(x_t)$. Fourth – symmetrization.

Fifth – splitting sup into two equal terms. Last – contraction as per exercise.

Same proof without shorthand (skip it)

$$\inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t} |f(x_s) - y_s| \right\} \right\}$$

$$= \sup_{p_t \in \Delta([-1,1])} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \left\{ |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t} |f(x_s) - y_s| \right\} \right\}$$

$$= \sup_{p_t \in \Delta([-1,1])} \left\{ \inf_{\widehat{y}_t} \mathbb{E}_{y_t} |\widehat{y}_t - y_t| + \mathbb{E}_{x_{t+1:n}, y_t, \epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t} |f(x_s) - y_s| \right\} \right\}$$

$$\leq \sup_{p_t \in \Delta([-1,1])} \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}, y_t} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t-1} |f(x_s) - y_s| \right.$$
$$\left. + \mathbb{E}_{y'_t} |f(x_t) - y'_t| - |f(x_t) - y_t| \right\}$$

$$\leq \sup_{p_t \in \Delta([-1,1])} \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}, y_t, y'_t} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t-1} |f(x_s) - y_s| \right.$$
$$\left. + \epsilon_t (|f(x_t) - y'_t| - |f(x_t) - y_t|) \right\}$$

$$\leq \sup_{p_t \in \Delta([-1,1])} \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}, y_t} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t-1} |f(x_s) - y_s| + 2\epsilon_t |f(x_t) - y_t| \right\}$$

$$\leq \mathbb{E}_{x_{t+1:n}, \epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t-1} |f(x_s) - y_s| \right\}$$

Taking expectation w.r.t x_t on both sides completes the proof.

Outline

Suppose neither X's nor Y's come from a stochastic process that we can model.

We can still consider the objective

$$\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t), y_t)$$

1. At time t, observe x_t
2. Compute $\widehat{y}_t$ based on $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and x_t
3. Observe y_t
4. Pay loss $\ell(\widehat{y}_t, y_t)$

At every step, prediction is being evaluated (as in cross-validation). Then new datum added to dataset.

Objective is still coupled through $\mathcal{F}$ which we believe to perform well.

Duality: Modeling data sources vs modeling solutions.

Admissibility

Slight change for admissibility in 2nd condition– replace $\mathbb{E}_{x_t}$ with $\sup_{x_t}$:

$$\sup_{x_t} \inf_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ \ell(\widehat{y}_t, y_t) + \mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) \right\} \leq \mathbf{Rel}(x_1, y_1, \ldots, x_{t-1}, y_{t-1})$$

Observe that the main part of earlier proof of admissibility was done conditionally on x_t followed by expectation on both sides. Instead, we take supremum. It is easy to check that this leads to interleaved $\sup_{x_t}$ and $\mathbb{E}_{\epsilon_t}$:

$$\mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) = \sup_{x_{t+1}} \mathbb{E}_{\epsilon_{t+1}} \ldots \sup_{x_n} \mathbb{E}_{\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t} |f(x_s) - y_s| \right\}$$

Sequential Rademacher Complexity

Lemma.

Relaxation on previous slide is admissible for prediction with absolute loss (similar relaxations can be derived for other losses too).

The algorithm based on this relaxation is as before: observe x_t and predict

$$\widehat{y}_t = \operatorname*{argmin}_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ \ell(\widehat{y}_t, y_t) + \mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) \right\}$$

Upper bound on regret of this method is $\frac{1}{n}$ times

$$\sup_{x_1} \mathbb{E}_{\epsilon_1} \ldots \sup_{x_n} \mathbb{E}_{\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} 2\epsilon_t f(x_t) \right\}$$

Sequential Rademacher Complexity

The expression

$$\sup_{x_1} \mathbb{E}_{\epsilon_1} \dots \sup_{x_n} \mathbb{E}_{\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(x_t) \right\}$$

can be written as

$$\mathscr{R}_n^{\text{seq}}(\mathcal{F}) \triangleq \sup_{\mathbf{x}} \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right\}$$

where $\mathbf{x}$ is a *tree* (predictable process).

An $\mathcal{X}$-valued tree $\mathbf{x}$ is a sequence of functions $\mathbf{x}_1, \dots, \mathbf{x}_n$ with $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$.

When $\epsilon_1, \dots, \epsilon_n$ are taken i.i.d. Rademacher, $\{\mathbf{x}_t\}$ is a predictable process with $\mathbf{x}_t$ being $\sigma(\epsilon_1, \dots, \epsilon_{t-1})$-measurable. We write $\mathbf{x}_t(\epsilon)$ for $\mathbf{x}_t(\epsilon_{1:t-1})$.

Sequential Rademacher complexity of $\mathcal{F}$ on $\mathbf{x}$:

$$\mathscr{R}_n^{\text{seq}}(\mathcal{F}; \mathbf{x}) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right\}$$

Constant $\{\mathbf{x}_t(\epsilon) = x_t\}$ predictable process gives us the classical definition.

Minimax regret for online learning with absolute loss or indicator loss is upper-bounded by $2\mathscr{R}_n^{\text{seq}}(\mathcal{F})$ and lower-bounded by $\mathscr{R}_n^{\text{seq}}(\mathcal{F})$.

For square loss, the behavior of minimax regret is given by sequential offset Rademacher (defined later).

This story is in parallel to i.i.d. statistical learning. But there are even more parallels! Much of empirical process theory extends to the case of trees.

Outline

Finite class

If $\mathcal{G} \subseteq [-1,1]^{\mathcal{X}}$, $|\mathcal{G}| = N$. Let $\mathbf{z}$ be a $\mathcal{Z}$-valued tree of depth n. Then

$$\mathbb{E} \max_{g \in \mathcal{G}} \left\{ \sum_{i=1}^{n} \epsilon_i g(\mathbf{z}_i(\epsilon)) \right\} \leq \sqrt{2n \log N}.$$

Again, a better bound (note $\max_\epsilon$) is

$$\mathbb{E} \max_{g \in \mathcal{G}} \left\{ \sum_{i=1}^{n} \epsilon_i g(\mathbf{z}_i(\epsilon)) \right\} \leq r\sqrt{2 \log N}, \quad r = \max_{\epsilon} \max_{g \in \mathcal{G}} \sqrt{\sum_{i=1}^{n} g(\mathbf{z}_i(\epsilon))^2}$$

Covering numbers

In the i.i.d. case, we considered $\mathcal{G}|_{z_1,\ldots,z_n}$ as the effective n-dimensional projection. In the sequential case, $\mathcal{G}|_{\mathbf{z}} = \{g \circ \mathbf{z} : g \in \mathcal{G}\}$ might be too large. Instead, consider the notion of a 0-cover (for binary-valued classes first).

Fix A, a set of $\{0,1\}$-valued trees.

Definition.

A 0-cover is the smallest set V of $\{0,1\}$-valued trees with the property

$$\forall \mathbf{a} \in A, \epsilon \in \{\pm 1\}^n, \exists \mathbf{v} \in V \text{ s.t. } \forall t, \quad \mathbf{a}_t(\epsilon) = \mathbf{v}_t(\epsilon)$$

Clearly,

$$\mathbb{E} \max_{\mathbf{a} \in A} \sum_{t=1}^{n} \epsilon_t \mathbf{a}_t(\epsilon) \leq \mathbb{E} \max_{\mathbf{v} \in V} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon)$$

Covering numbers

Let $\mathcal{N}(\mathcal{G}, \mathbf{z}, 0)$ denote the size of the smallest 0-cover. We have shown that

$$\mathscr{R}_n^{\text{seq}}(\mathcal{G}; \mathbf{z}) \leq \sqrt{\frac{2 \log \mathcal{N}(\mathcal{G}, \mathbf{z}, 0)}{n}}$$

for a binary-valued class $\mathcal{G}$.

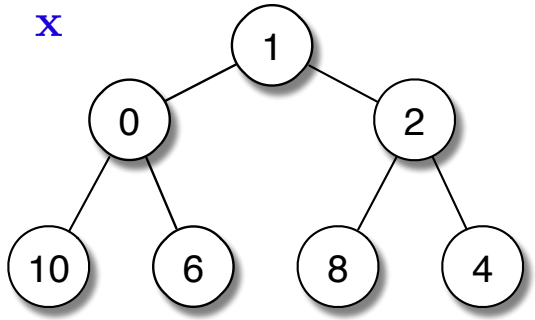
Gap between the size of $\mathcal{G}|_{\mathbf{z}}$ and the size of its 0-cover can be exponential.

Example: Fix $\mathbf{z}$ and take

$$\mathcal{G} = \left\{ g^{\hat{\epsilon}} : \ \forall \epsilon, \quad \forall t < n, \quad g^{\hat{\epsilon}}(\mathbf{z}_t(\epsilon)) = 0; \quad g^{\hat{\epsilon}}(\mathbf{z}_n(\epsilon)) = \mathbf{I}\{\hat{\epsilon}_{1:n-1} = \epsilon_{1:n-1}\} \right\}$$

Then cardinality of $\mathcal{G}|_{\mathbf{z}}$ is 2^{n-1}, but size of 0-cover is 2.

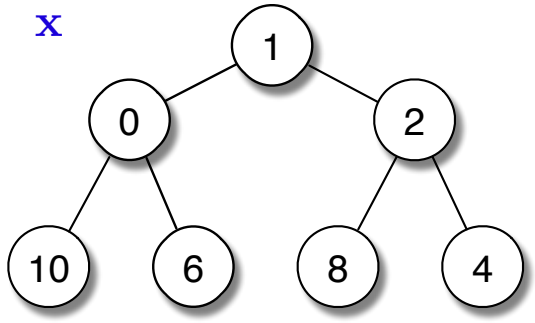
Example: Covering number



$$f_1(x) = \mathbf{I}\{x \in [9, 11]\}$$
$$f_2(x) = \mathbf{I}\{x \in [5, 7]\}$$
$$f_3(x) = \mathbf{I}\{x \in [7, 9]\}$$
$$f_4(x) = \mathbf{I}\{x \in [3, 5]\}$$

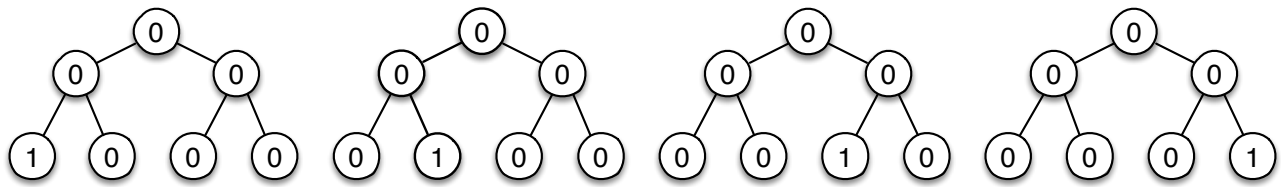
$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^n\ \exists \mathbf{v} \in V \qquad \text{s.t.} \qquad \mathbf{v}_t(\epsilon) = f(\mathbf{x}_t(\epsilon))$$

Example: Covering number



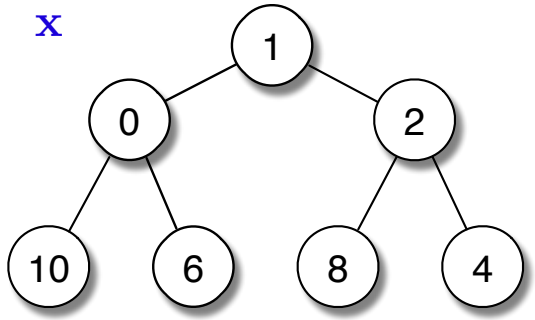
$$f_1(x) = \mathbf{I}\{x \in [9, 11]\}$$
$$f_2(x) = \mathbf{I}\{x \in [5, 7]\}$$
$$f_3(x) = \mathbf{I}\{x \in [7, 9]\}$$
$$f_4(x) = \mathbf{I}\{x \in [3, 5]\}$$

$f_1 \circ \mathbf{x}$ $f_2 \circ \mathbf{x}$ $f_3 \circ \mathbf{x}$ $f_4 \circ \mathbf{x}$



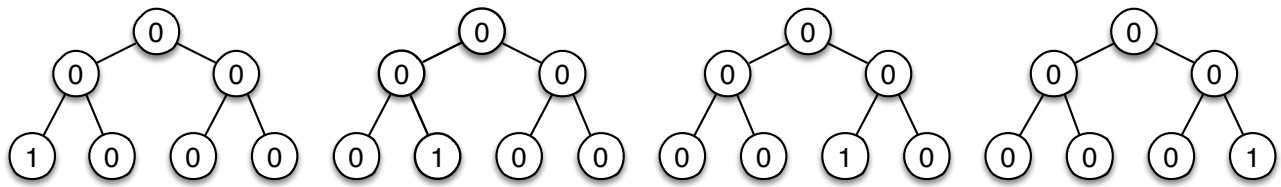
$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^n\ \exists \mathbf{v} \in V \qquad \text{s.t.} \qquad \mathbf{v}_t(\epsilon) = f(\mathbf{x}_t(\epsilon))$$

Example: Covering number

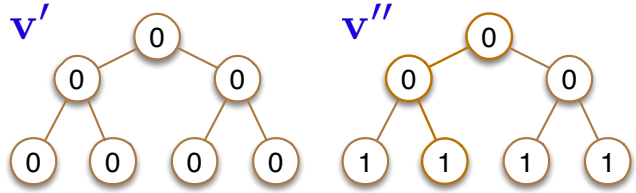


$$f_1(x) = \mathbf{I}\{x \in [9,11]\}$$
$$f_2(x) = \mathbf{I}\{x \in [5,7]\}$$
$$f_3(x) = \mathbf{I}\{x \in [7,9]\}$$
$$f_4(x) = \mathbf{I}\{x \in [3,5]\}$$

$f_1 \circ \mathbf{x}$ $f_2 \circ \mathbf{x}$ $f_3 \circ \mathbf{x}$ $f_4 \circ \mathbf{x}$

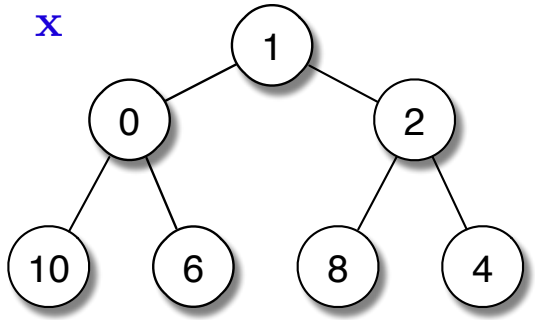


$\mathbf{v}'$ $\mathbf{v}''$



$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^n\ \exists \mathbf{v} \in V \qquad \text{s.t.} \qquad \mathbf{v}_t(\epsilon) = f(\mathbf{x}_t(\epsilon))$$

Example: Covering number

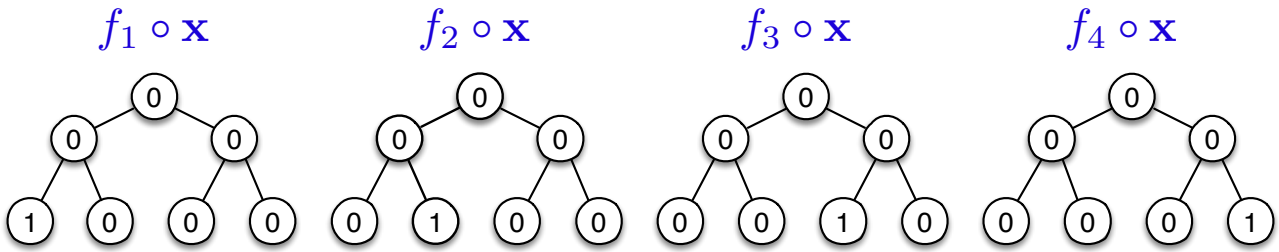


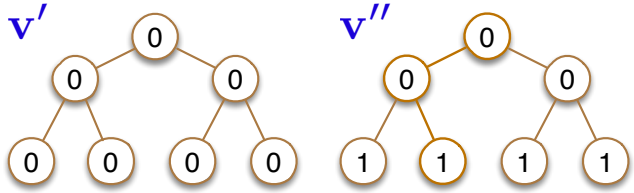
$$f_1(x) = \mathbf{I}\{x \in [9,11]\}$$

$$f_2(x) = \mathbf{I}\{x \in [5,7]\}$$

$$f_3(x) = \mathbf{I}\{x \in [7,9]\}$$

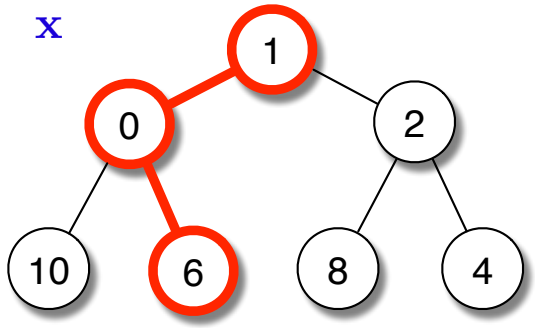
$$f_4(x) = \mathbf{I}\{x \in [3,5]\}$$





$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^n\ \exists \mathbf{v} \in V \qquad \text{s.t.} \qquad \mathbf{v}_t(\epsilon) = f(\mathbf{x}_t(\epsilon))$$

Example: Covering number

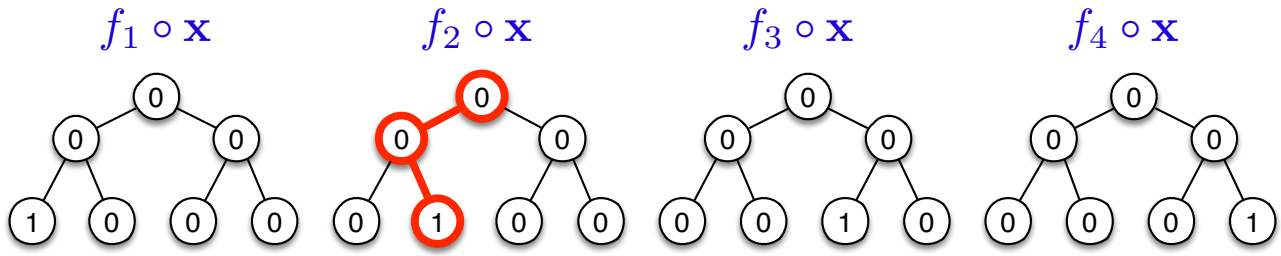


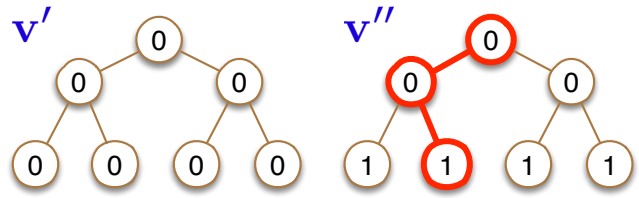
$$f_1(x) = \mathbf{I}\{x \in [9,11]\}$$

$$f_2(x) = \mathbf{I}\{x \in [5,7]\}$$

$$f_3(x) = \mathbf{I}\{x \in [7,9]\}$$

$$f_4(x) = \mathbf{I}\{x \in [3,5]\}$$





$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^n\ \exists \mathbf{v} \in V \qquad \text{s.t.} \qquad \mathbf{v}_t(\epsilon) = f(\mathbf{x}_t(\epsilon))$$

Covering numbers

For real-valued $\mathcal{G}$, a scale-sensitive cover is needed.

Definition.

A set V of $\mathbb{R}$-valued trees of depth n is an ℓ_p-cover at scale $\alpha > 0$ of $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ on a $\mathbf{z}$-valued tree $\mathbf{z}$ if

$$\forall g \in \mathcal{G}, \epsilon \in \{\pm 1\}^n, \quad \exists \mathbf{v} \in V, \quad \text{s.t.} \quad \frac{1}{n} \sum_{t=1}^{n} |g(\mathbf{z}_t(\epsilon)) - \mathbf{v}_t(\epsilon)|^p \leq \alpha^p.$$

The size of the smallest cover is denoted $\mathcal{N}_p(\mathcal{G}, \mathbf{z}, \alpha)$.

(R., Sridharan, Tewari '10)

Sequential Chaining

Suppose $\mathcal{G} \subseteq [-1,1]^{\mathcal{Z}}$.

For any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n,

$$\mathscr{R}_n^{\text{seq}}(\mathcal{G};\mathbf{z}) \le \inf_{\alpha \ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\sqrt{2\log \mathcal{N}_1(\mathcal{G},\mathbf{z},\alpha)}\right\}$$

A better bound (called Dudley entropy integral):

$$\mathscr{R}_n^{\text{seq}}(\mathcal{G};\mathbf{z}) \le \inf_{\alpha \ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{\log \mathcal{N}_2(\mathcal{G},\mathbf{z},\delta)}\,\mathrm{d}\delta\right\}$$

Combinatorial parameters: real-valued case

Definition.

A $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d is α-*shattered* by $\mathcal{G}$ if there exists an $\mathbb{R}$-valued tree $\mathbf{s}$ such that

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists g \in \mathcal{G} \quad \text{s.t. } \forall t \in [d],\ \epsilon_t (g(\mathbf{z}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \geq \alpha/2$$

The fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{G})$ at scale α is the largest d such that $\mathcal{G}$ α-shatters an $\mathcal{Z}$-valued tree of depth d.

This is a generalization of the scale-sensitive dimension introduced in (Kearns and Schapire, '94) and (Bartlett, Long, and Williamson, '94) for i.i.d. learning.

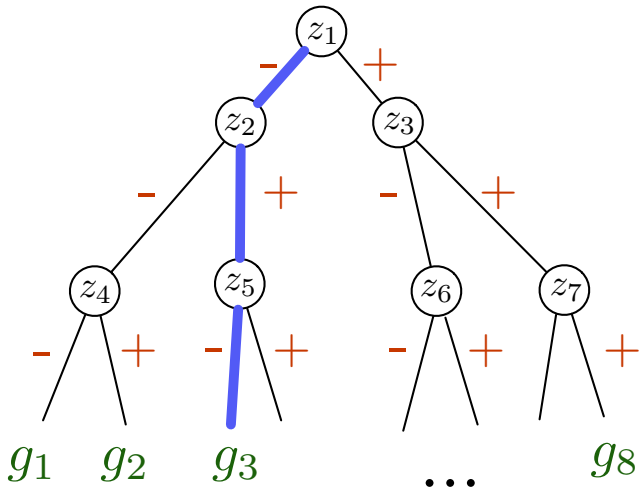
Combinatorial parameters: binary case

Definition (Littlestone 88; Ben-David, Pál, Shalev-Shwartz 09).

An $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d is *shattered* by a function class $\mathcal{G} \subseteq \{\pm 1\}^{\mathcal{Z}}$ if

$$\forall \epsilon \in \{\pm 1\}^d \quad \exists g \in \mathcal{G} \quad \text{s.t.} \quad \forall t \in [d] \quad g(\mathbf{z}_t(\epsilon)) = \epsilon_t .$$

The *Littlestone dimension* $\ell\dim(\mathcal{G})$ is the largest d such that $\mathcal{G}$ shatters an $\mathcal{Z}$-valued tree of depth d.



$\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3) = (-1, +1, -1)$

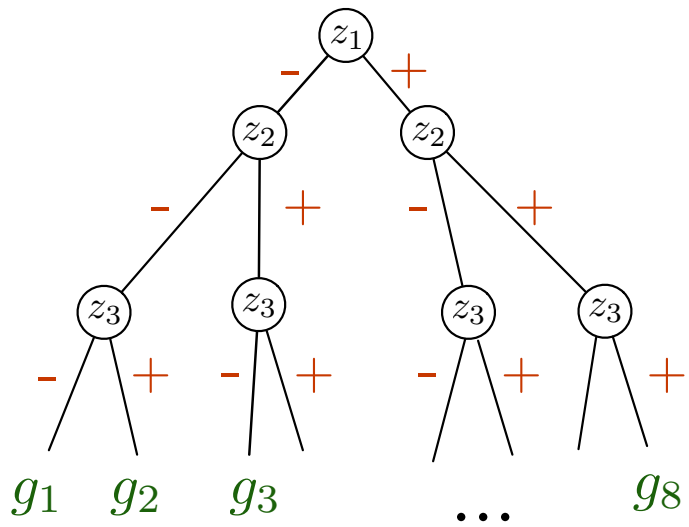
$g_3(x_1) = -1$

$g_3(x_2) = +1$

$g_3(x_5) = -1$

Combinatorial parameters: binary case

The notion of Vapnik-Chervonenkis dimension is recovered if the tree has constant z_t at each level: $\mathbf{z}_t(\epsilon_{1:t-1}) = z_t$.



In particular, this implies $\mathrm{vc}(\mathcal{G}) \leq \ell\dim(\mathcal{G})$.

Analogue of Vapnik-Chervonenkis-Sauer-Shelah Lemma for trees:

Theorem.

For binary-valued class $\mathcal{G}$ of $\ell\dim(\mathcal{G}) = d$ and any $\mathbf{z}$,

$$\mathcal{N}(0, \mathcal{G}, \mathbf{z}) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^{d}.$$

Theorem.

For $[-1, 1]$-valued class $\mathcal{G}$ and any $\mathbf{z}$,

$$\mathcal{N}_\infty(\alpha, \mathcal{G}, \mathbf{z}) \leq \left(\frac{2en}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{G})}$$

Open: dimension-free estimate on ℓ_2-cover *a la* Mendelson & Vershynin.

Martingale uGC

Definition.

Function class $\mathcal{F}$ satisfies *Sequential Uniform Convergence* if,

$$\forall \delta > 0, \lim_{n' \to \infty} \sup_{\mathbb{P}} \mathbb{P}\left(\sup_{n \geq n'} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[f(X_t)|X^{t-1}] - f(X_t) \right| > \delta \right) = 0$$

Definition of *uniform Glivenko-Cantelli* classes is recovered if supremum is taken over i.i.d. distributions.

Theorem (R., Sridharan, Tewari '10, '15)**.**

Let $\mathcal{F}$ be a class of $[-1,1]$-valued functions. The following are equivalent:

1. $\mathcal{F}$ satisfies uniform convergence of averages to conditional means (martingale extension of *uniform Glivenko-Cantelli*)
2. Sequential Rademacher $\mathscr{R}_n^{\text{seq}}(\mathcal{F}) \to 0$
3. Sequential version of Dudley entropy integral converges
4. Sequential $\text{fat}_\alpha(\mathcal{F})$ is finite for all $\alpha > 0$
5. $\mathcal{F}$ is "online learnable"

Outline

The minimax rate for excess loss

- for loss functions without strong convexity (indicator loss, absolute loss) is given by *sequential* Rademacher averages
- for square loss – by *sequential* offset Rademacher averages (we are not going to state this formally, but you can check out the paper "Online Non-Parametric Regression").

Outline

Problem I: Estimation (Random Design)

Model:

$$Y_i = \eta(X_i) + \xi_i, \qquad \eta \in \mathcal{F}$$

- $\mathcal{F}$ is a class of functions $\mathcal{X} \to \mathcal{Y}$
- $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. from P_η on $\mathcal{X} \times \mathcal{Y}$
- Regression function $\mathbb{E}[Y|X = x] = \eta(x)$, $\|\cdot\| = \|\cdot\|_{L_2(P_X)}$

Minimax risk:

$$W_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_\eta : \eta \in \mathcal{F}} \mathbb{E}\|\hat{f} - \eta\|^2.$$

From well-specified to misspecified

$\mathcal{P}$ is the set of all distributions on $\mathcal{X} \times \mathcal{Y}$ (or a superset of $\{P_\eta : \eta \in \mathcal{F}\}$)

$$W_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_\eta : \eta \in \mathcal{F}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 \right\}$$

From well-specified to misspecified

$\mathcal{P}$ is the set of all distributions on $\mathcal{X} \times \mathcal{Y}$ (or a superset of $\{P_\eta : \eta \in \mathcal{F}\}$)

$$W_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_\eta : \eta \in \mathcal{F}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 \right\}$$

$$= \inf_{\hat{f}} \sup_{P_\eta : \eta \in \mathcal{F}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\}$$

From well-specified to misspecified

$\mathcal{P}$ is the set of all distributions on $\mathcal{X} \times \mathcal{Y}$ (or a superset of $\{P_\eta : \eta \in \mathcal{F}\}$)

$$
\begin{aligned}
W_n(\mathcal{F}) &= \inf_{\hat{f}} \sup_{P_\eta : \eta \in \mathcal{F}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 \right\} \\
&= \inf_{\hat{f}} \sup_{P_\eta : \eta \in \mathcal{F}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\} \\
&\leq \inf_{\hat{f}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\}
\end{aligned}
$$

From well-specified to misspecified

$\mathcal{P}$ is the set of all distributions on $\mathcal{X} \times \mathcal{Y}$ (or a superset of $\{P_\eta : \eta \in \mathcal{F}\}$)

$$
\begin{aligned}
W_n(\mathcal{F}) &= \inf_{\hat{f}} \sup_{P_\eta : \eta \in \mathcal{F}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 \right\} \\
&= \inf_{\hat{f}} \sup_{P_\eta : \eta \in \mathcal{F}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\} \\
&\leq \inf_{\hat{f}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\} \\
&= \inf_{\hat{f}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E} (\hat{f}(X) - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E} (f(X) - Y)^2 \right\} \\
&= V_n(\mathcal{F})
\end{aligned}
$$

Problem II: Statistical Learning

Model: any distribution P on $\mathcal{X} \times \mathcal{Y}$

- $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d. from P
- Regression function $\mathbb{E}[Y|X = x] = \eta(x)$ not necessarily in $\mathcal{F}$.

Minimax regret:

$$V_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P} \left\{ \mathbb{E}(\hat{f}(X) - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2 \right\}$$

From statistical to online learning

Shorthand $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $Z_i = (X_i, Y_i)$, $\ell(f, Z) = (f(X) - Y)^2$. Sequence of estimators: $\hat{f}_t = \hat{f}_t(Z_1, \ldots, Z_{t-1})$, $t = 1, \ldots, n$.

$$V_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P} \left\{ \mathbb{E}\ell(\hat{f}, Z) - \inf_{f \in \mathcal{F}} \mathbb{E}\ell(f, Z) \right\}$$

From statistical to online learning

Shorthand $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $Z_i = (X_i, Y_i)$, $\ell(f, Z) = (f(X) - Y)^2$. Sequence of estimators: $\hat{f}_t = \hat{f}_t(Z_1, \ldots, Z_{t-1})$, $t = 1, \ldots, n$.

$$V_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P} \left\{ \mathbb{E}\ell(\hat{f}, Z) - \inf_{f \in \mathcal{F}} \mathbb{E}\ell(f, Z) \right\}$$

$$\approx \inf_{\{\hat{f}_t\}} \sup_{P^{\otimes n}} \left\{ \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\ell(\hat{f}_t, Z) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\ell(f, Z) \right\}$$

From statistical to online learning

Shorthand $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $Z_i = (X_i, Y_i)$, $\ell(f, Z) = (f(X) - Y)^2$. Sequence of estimators: $\hat{f}_t = \hat{f}_t(Z_1, \ldots, Z_{t-1})$, $t = 1, \ldots, n$.

$$
\begin{aligned}
V_n(\mathcal{F}) &= \inf_{\hat{f}} \sup_{P} \left\{ \mathbb{E}\ell(\hat{f}, Z) - \inf_{f \in \mathcal{F}} \mathbb{E}\ell(f, Z) \right\} \\
&\approx \inf_{\{\hat{f}_t\}} \sup_{P^{\otimes n}} \left\{ \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\ell(\hat{f}_t, Z) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\ell(f, Z) \right\} \\
&\leq \inf_{\{\hat{f}_t\}} \sup_{P^{\otimes n}} \mathbb{E} \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{f}_t, Z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, Z_t) \right\}
\end{aligned}
$$

From statistical to online learning

Shorthand $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $Z_i = (X_i, Y_i)$, $\ell(f, Z) = (f(X) - Y)^2$. Sequence of estimators: $\hat{f}_t = \hat{f}_t(Z_1, \ldots, Z_{t-1})$, $t = 1, \ldots, n$.

$$
\begin{aligned}
V_n(\mathcal{F}) &= \inf_{\hat{f}} \sup_{P} \left\{ \mathbb{E}\ell(\hat{f}, Z) - \inf_{f \in \mathcal{F}} \mathbb{E}\ell(f, Z) \right\} \\
&\approx \inf_{\{\hat{f}_t\}} \sup_{P^{\otimes n}} \left\{ \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\ell(\hat{f}_t, Z) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\ell(f, Z) \right\} \\
&\leq \inf_{\{\hat{f}_t\}} \sup_{P^{\otimes n}} \mathbb{E} \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{f}_t, Z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, Z_t) \right\} \\
&\leq \inf_{\{\hat{f}_t\}} \sup_{P} \mathbb{E} \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{f}_t, Z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, Z_t) \right\}
\end{aligned}
$$

From statistical to online learning

Shorthand $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $Z_i = (X_i, Y_i)$, $\ell(f, Z) = (f(X) - Y)^2$. Sequence of estimators: $\hat{f}_t = \hat{f}_t(Z_1, \ldots, Z_{t-1})$, $t = 1, \ldots, n$.

$$
\begin{aligned}
V_n(\mathcal{F}) &= \inf_{\hat{f}} \sup_{P} \left\{ \mathbb{E}\ell(\hat{f}, Z) - \inf_{f \in \mathcal{F}} \mathbb{E}\ell(f, Z) \right\} \\
&\approx \inf_{\{\hat{f}_t\}} \sup_{P^{\otimes n}} \left\{ \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\ell(\hat{f}_t, Z) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\ell(f, Z) \right\} \\
&\leq \inf_{\{\hat{f}_t\}} \sup_{P^{\otimes n}} \mathbb{E} \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{f}_t, Z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, Z_t) \right\} \\
&\leq \inf_{\{\hat{f}_t\}} \sup_{P} \mathbb{E} \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{f}_t, Z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, Z_t) \right\} \\
&= \inf_{\{\hat{f}_t\}} \sup_{(z_1, \ldots, z_n)} \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{f}_t, z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) \right\} \\
&= R_n(\mathcal{F})
\end{aligned}
$$

Problem III: Sequential Prediction (Online Regression)

Model: individual sequence $(x_1, y_1), \ldots, (x_n, y_n)$

At each time step $t = 1, \ldots, n$,

- x_t is revealed
- prediction $\widehat{y}_t \in \mathcal{Y}$ is made by the forecaster
- $y_t \in \mathcal{Y}$ is revealed

Minimax regret:

$$R_n(\mathcal{F}) = \inf_{\text{Algo}} \sup_{\{(x_t, y_t)\}_{t=1}^n} \left\{ \frac{1}{n} \sum_{t=1}^n (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n (f(x_t) - y_t)^2 \right\}$$

Whenever sequential offset complexity and i.i.d. offset Rademacher averages are of the same order of magnitude, one may claim no gaps in the above sequence of inequalities. In particular, one may claim that methods built for the online problem are near-optimal for statistical learning and the estimation problems with i.i.d. data.

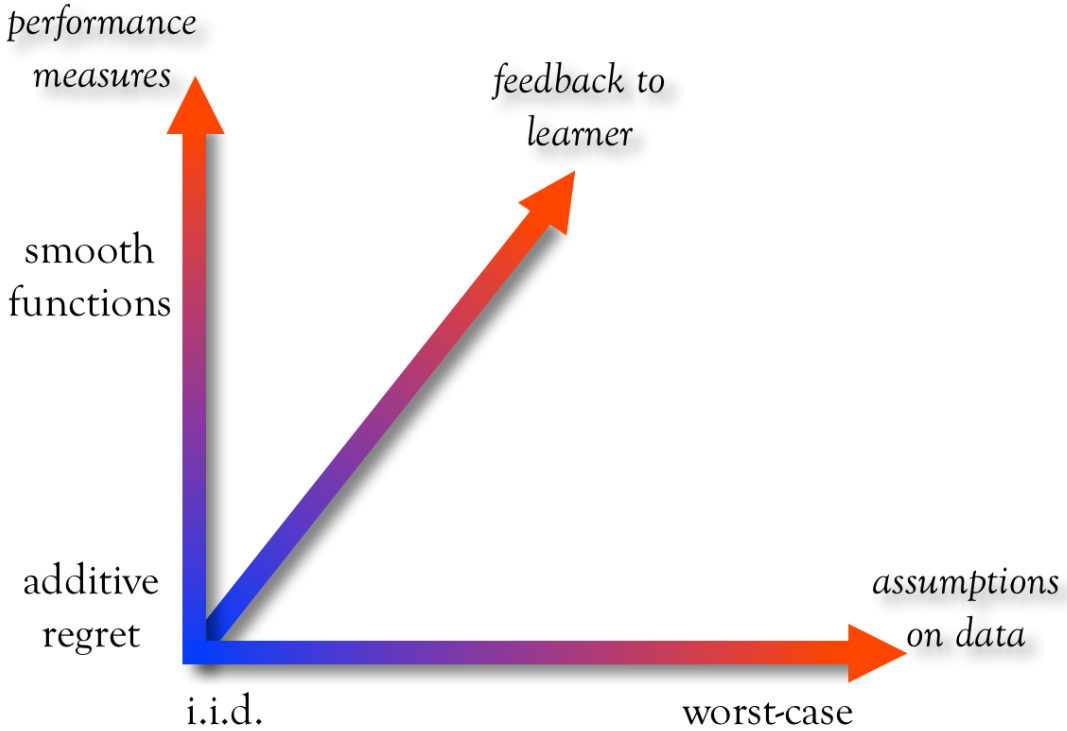
Outline

In this lecture, we focus on computation.

Pretty much all the online prediction methods can be derived through the relaxation framework, starting with sequential Rademacher complexity. We outline the main algorithmic techniques on several examples.

We have reduced the problem of finding a prediction method to the problem of finding a good admissible relaxation. Think about this statement: it is rare that we have an algorithmic parametrization like this (in general, the space of all algorithms is very large!).

First, we show that one can use these constructions for deriving estimators in the i.i.d. setting (1st lecture). This gives a new language for talking about improper estimators.


performance measures
feedback to learner
smooth functions
additive regret
assumptions on data
i.i.d.
worst-case

Outline

We will describe an array of tools for building low-regret algorithms. Given x_t we can solve for $\widehat{y}_t$ s.t. for any sequence $(x_1, y_1), \ldots, (x_n, y_n)$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t), y_t)\right] \leq \psi_n$$

for some decreasing ψ_n, where the expectation is with respect to a possible randomization of the prediction method.

We can compute the solution $\widehat{y}_t(x_t)$ as a function for all possible x_t: $\widehat{f}_t = \widehat{y}_t(\cdot)$. If data are i.i.d., we use Polyak averaging

$$\widehat{f} = \frac{1}{n}\sum_{t=1}^{n}\widehat{f}_t$$

If $\ell(\cdot, y)$ is convex and data are i.i.d., an easy argument shows that (prove!)

$$\mathbb{E}\mathbf{L}(\widehat{f}) - \inf_{f\in\mathcal{F}}\mathbf{L}(f) \leq \psi_n$$

Furthermore, if ψ_n is data-dependent, we may obtain data-dependent bounds.

Outline

The starting point is the lemma from previous lecture: the Rademacher-based relaxation

$$\mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) = \sup_{\mathbf{x}} \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(\mathbf{x}_s) - \sum_{s=1}^{t} |f(x_s) - y_s| \right\}$$

is admissible for prediction with absolute (in fact, any 1-Lipshitz or the indicator) loss for the problem when both x and y come from some unknown process.

The algorithm then is to find an (approximate) minimum of

$$\operatorname*{argmin}_{\widehat{y}_t} \sup_{y_t \in [-1,1]} \left\{ \ell(\widehat{y}_t, y_t) + \mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) \right\}$$

The main computational impediment to using this relaxation is the supremum over $\mathbf{x}$ which requires us to search over binary trees.

Getting rid of the trees

Computationally efficient methods are obtained by removing the tree in some fashion. This is typically done by further upper bounding **Rel** to obtain a new relaxation. *Note*: admissibility needs to be checked for any such upper bound.

The following are major approaches to getting rid of the tree:

1. Use a probabilistic upper bound to remove the tree (see, for instance, derivations of Exponential Weights and Dual Averaging below).
2. Show existence of a distribution on $\mathcal{X}$ that is "almost as bad" as a single worst-case choice of $x \in \mathcal{X}$. Sample from this distribution to imitate the supremum over trees (this can be made precise!)
3. Assume that x's come from some stochastic process from which we can sample. Example: use unlabeled data together with random playout (see below).
4. Suppose x's come from a finite set of size n without replacement.
5. Of course, the tree also disappears when there are effectively no x's and each $f = (f_1, \ldots, f_n)$ (this case is known as *static experts*).

In many (but not all) cases, sequential Rademacher complexity is of the same order as the classical Rademacher. In these cases, the second approach (showing existence of a distribution) appears to be extremely powerful.

Outline

Example: finite class $\mathcal{F}$

Assume $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$. Then

$$\begin{aligned}
\mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) &= \sup_{\mathbf{x}} \mathbb{E} \max_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(\mathbf{x}_s(\epsilon)) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \right\} \\
&\leq \sup_{\mathbf{x}} \frac{1}{\eta} \ln \mathbb{E} \sum_{f \in \mathcal{F}} \exp \left\{ \eta \sum_{s=t+1}^{n} 2\epsilon_s f(\mathbf{x}_s(\epsilon)) - \eta \sum_{s=1}^{t} \ell(f(x_s), y_s) \right\} \\
&\leq \frac{1}{\eta} \ln \sum_{f \in \mathcal{F}} \exp \left\{ -\eta \sum_{s=1}^{t} \ell(f(x_s), y_s) \right\} + 2\eta(n-t)
\end{aligned}$$

How did we get rid of the tree in the last inequality? By the standard subgaussian inequality, peeling off one term at a time from $s = n$ to $s = t+1$.

Example: finite class $\mathcal{F}$

It is an good exercise to check that

$$\inf_{\eta>0}\left\{\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{s=1}^{t}\ell(f(x_s),y_s)\right)\right)+2\eta(n-t)\right\}$$

is admissible and leads to a parameter-free version of the Exponential Weights Algorithm. Without any structural assumptions on the class, this is the tightest relaxation.

The bound on regret is simply

$$\mathbf{Rel}(\varnothing)=\inf_{\eta}\left\{\frac{1}{\eta}\log\sum_{f\in\mathcal{F}}\exp\{0\}+2\eta n\right\}=\sqrt{\frac{2\log N}{n}}$$

The Exponential Weights Algorithm was re-discovered several times: Vovk '90, Littlestone & Warmuth 94, etc. It can be seen as an instance of Mirror Descent with entropy function (Nemirovskii & Yudin '79). It remains the most basic prediction method because it does not assume any knowledge of how experts (functions f) make prediction, as long as there are N of them.

Example: linear loss and dual averaging

Slightly different setting: $\ell(f, z) = \langle f, z \rangle$, $\mathcal{F}$ and $\mathcal{Z}$ are unit balls in dual Banach spaces. Protocol: we forecast f_t and then observe z_t. Known as Online Convex Optimization (OCO).

Recall the definition of a dual norm: $\sup_{f \in \mathcal{F}} \langle f, z \rangle = \|z\|$. Write sequential Rademacher as

$$\mathbf{Rel}(z_1, \ldots, z_t) = \sup_{\mathbf{x}} \mathbb{E} \left\| \sum_{s=t+1}^{n} 2\epsilon_s \mathbf{z}_s(\epsilon) - \sum_{s=1}^{t} z_s \right\|$$

Example: linear loss and dual averaging

If norm is smooth (second derivative is bounded), we can expand

$$\|a + \epsilon_1 b\|^2 \leq \|a\|^2 + \epsilon_1 \left\langle \nabla \|a\|^2, b \right\rangle + C \|b\|^2$$

and the middle term is zero in expectation. We use this simple fact to get rid of the tree.

By Jensen's inequality and then by repeating the above manipulation $n - t$ times, an upper bound on sequential Rademacher is

$$\sqrt{\|\tilde{z}_{t-1}\|^2 + \left\langle \nabla \|\tilde{z}_{t-1}\|^2, z_t \right\rangle + C(n - t + 1)}$$

where $\tilde{z}_{t-1} = \sum_{i=1}^{t-1} z_t$.

This simple upper bound turns out to be an admissible relaxation and it leads to a projection-free "dual-averaging"-style method:

$$f_t = -\frac{\nabla \|\tilde{z}_{t-1}\|^2}{2\sqrt{\|\tilde{z}_{t-1}\|^2 + C(n - t + 1)}}$$

Outline

From Sequential to Classical Rademacher

Consider the proof of admissibility (the "scary proof") from the previous lecture. For the case of non-i.i.d. x_t, we took supremum on both sides. For the right-most term on that slide this supremum is

$$\sup_{x_t} \mathbb{E}_{\epsilon_t} \left[\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) + \epsilon_t f(x_t) - \sum_{s=1}^{t-1} |f(x_s) - y_s| \right\} \right]$$

Suppose we can find a distribution D on $\mathcal{X}$ such that the above quantity is upper bounded by

$$\mathbb{E}_{x_t \sim D} \; \mathbb{E}_{\epsilon_t} \left[\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) + C\epsilon_t f(x_t) - \sum_{s=1}^{t-1} |f(x_s) - y_s| \right\} \right]$$

for some constant $C \geq 2$.

If we can find such a distribution D, several nice things happen:

- sequential Rademacher is upper bounded by classical one (under the distribution D) up to constant C
- regret of the prediction method based on the classical Rademacher relaxation is bounded by classical Rademacher under D
- we may use the knowledge of D to gain efficiency by "sampling future"

Random Playout

On the last point, as soon as the relaxation involves $\mathbb{E}_{x_{t:n}, \epsilon_{t:n}}$, we can sample these random variables.

Recall the exercise:

Let W be a random variable, and suppose that for any realization of W,

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} \{\ell(a, b) + \Psi_t(b, W)\} \leq \Psi_{t-1}(W)$$

Prove that

$$\inf_{q \in \Delta(\mathcal{A})} \sup_{b \in \mathcal{B}} \{\mathbb{E}_{a \sim q} \ell(a, b) + \mathbb{E}_W \Psi_t(b, W)\} \leq \mathbb{E}_W \Psi_{t-1}(W)$$

by exhibiting a strategy for the infimum.

Random playout: draw $x_{t:n} \sim D$ and $\epsilon_{t:n}$ and solve for

$$\sup_{f \in \mathcal{F}} \left\{ \sum_{s=t}^{n} C \epsilon_s f(x_s) - \sum_{s=1}^{t-1} |f(x_s) - y_s| \right\}$$

Alternative: use unlabeled data $x_{t:n}$

Random Playout

Easiest example is linear loss $\ell(f, z) = \langle f, z \rangle$. Need to find a distribution $D \in \Delta(\mathcal{Z})$ and C such that for any w

$$\sup_{z \in \mathcal{Z}} \mathop{\mathbb{E}}_{\epsilon} \|w + \epsilon z\| \leq \mathop{\mathbb{E}}_{z \sim D} \|w + Cz\| \ .$$

At time t, draw $z_{t+1}, \ldots, z_n \sim D$ and and compute

$$f_t = \operatorname*{argmin}_{g \in \mathcal{F}} \sup_{z \in \mathcal{Z}} \left\{ \langle g, z \rangle + \left\| C \sum_{i=t+1}^{n} z_i - \sum_{i=1}^{t-1} z_i - z \right\| \right\}$$

This randomized strategy is an admissible algorithm w.r.t. conditional *classical* Rademacher complexity. Can find closed-form solutions.

The idea of replacing martingales with iid draws is quite general.

Smoothed Fictitious Play / Follow the Perturbed Leader

Add noise to the cumulative payoffs of each action and choose the best.

Turns out that algorithm on previous page can be of above form.

Idea goes back to (Hannan, 1957), re-discovered in (Kalai & Vempala, 2004)

"Smoothed" empirical risk minimization (or, *smooth fictitious play*).

General idea is related to random rollout in approximate dynamic programming.

Outline

Suppose x_t's come from a finite set of n items without replacement.

Then the tree disappears and the future $x_{t+1}, \ldots, x_n$ can be chosen in any prefixed order.

Algorithms for classification problems are especially simple. Relaxation is

$$\mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) = \mathbb{E}_\epsilon \max_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} 2\epsilon_s f(x_s) - \sum_{s=1}^{t} \mathbf{I}\{f(x_s) \neq y_s\} \right\}$$

and

$$\widehat{y}_t = \operatorname*{argmin}_{q_t \in \Delta(\{-1,1\})} \max_{y_t \in \{-1,1\}} \left\{ \mathbb{E}_{\widehat{y}_t \sim q_t} \mathbf{I}\{\widehat{y}_t \neq y_t\} + \mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) \right\}$$

Writing $\mathbf{I}\{\widehat{y}_t \neq y_t\} = \frac{1}{2}(1 - \widehat{y}_t y_t)$,

$$\widehat{y}_t = \operatorname*{argmin}_{\mu_t \in [-1,1]} \max_{y_t \in \{-1,1\}} \left\{ \frac{1}{2}(1 - \mu_t y_t) + \mathbf{Rel}(x_1, y_1, \ldots, x_t, y_t) \right\}$$

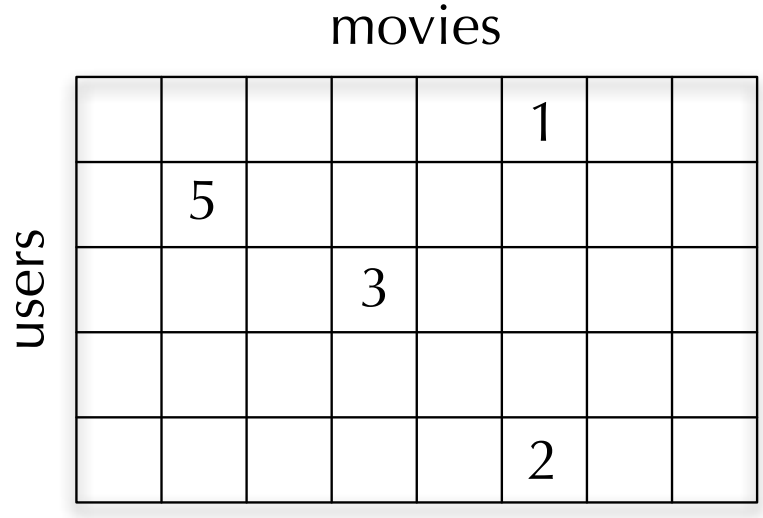
where μ_t is the mean of the distribution q_t. Equating the two possibilities, optimum is at

$$\mu_t = \mathbf{Rel}(x_1, y_1, \ldots, x_t, +1) - \mathbf{Rel}(x_1, y_1, \ldots, x_t, -1)$$

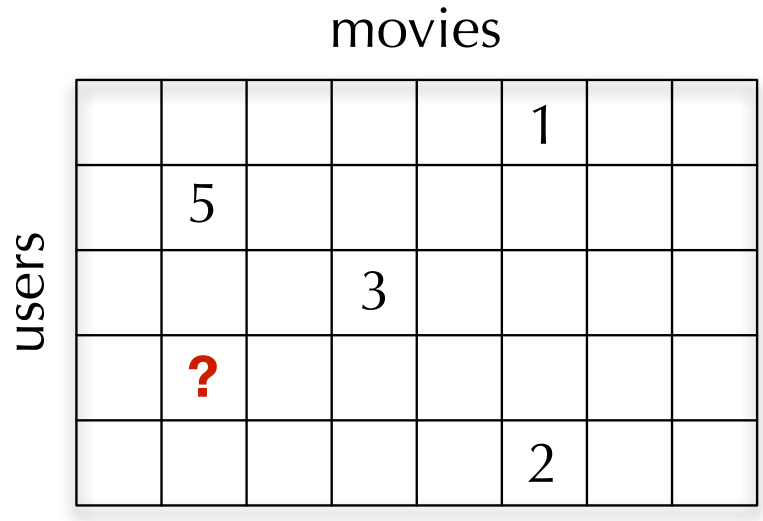
Interpretation: if potential does not change when changing -1 to 1, predict with probability $1/2$.

(to be precise, we require that solution be clipped to $[-1, 1]$)

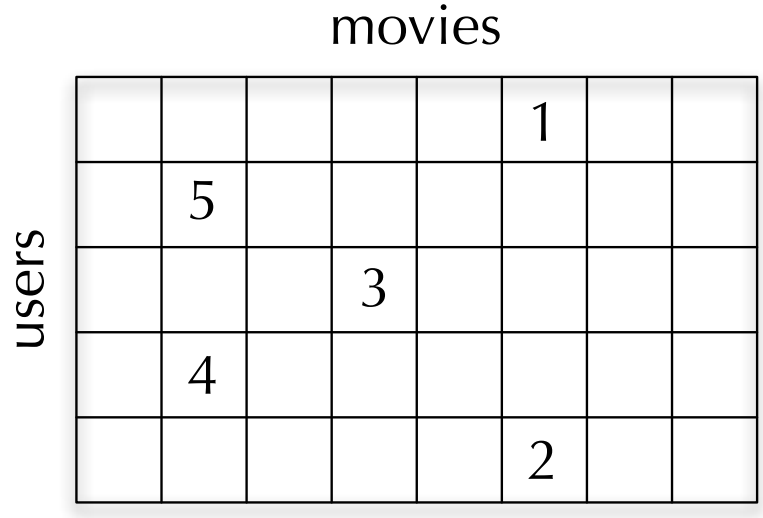
Matrix completion / collaborative filtering



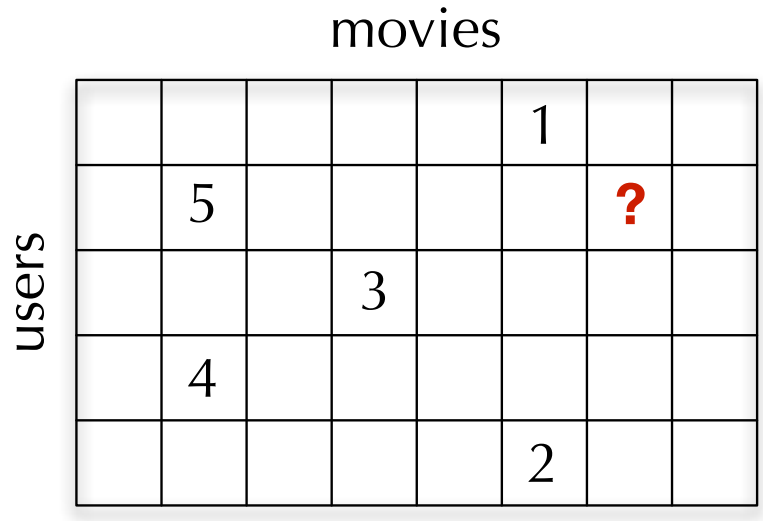
Matrix completion / collaborative filtering



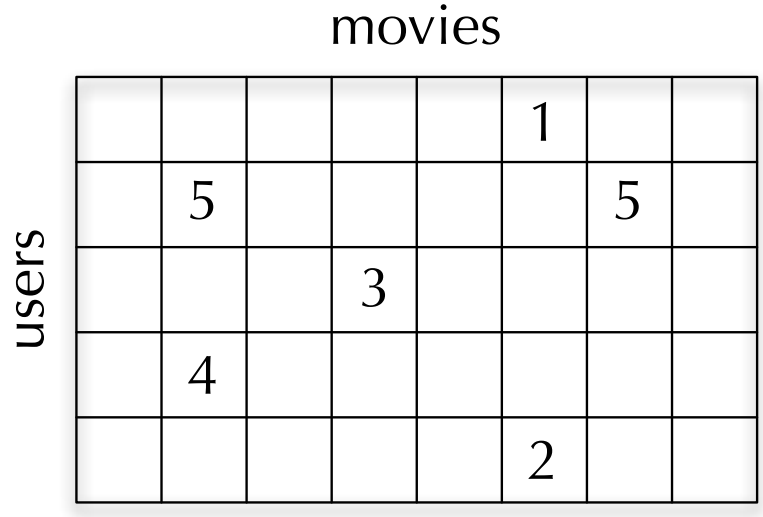
Matrix completion / collaborative filtering



Matrix completion / collaborative filtering



Matrix completion / collaborative filtering



Claim: can make number of mistakes not much worse than that made by a low-trace-norm matrix.

Set it up as a sequential optimization problem and solve the dynamic programming problem with some tricks.

Matrix completion / collaborative filtering

For $t = 1, \dots, T$
Observe person/movie identity $x_t = \mathbb{1}(i_t, j_t) \in \{0,1\}^{m \times n}$
Make randomized prediction $\widehat{y}_t \sim q_t \in \Delta(\{-1, 1\})$
Observe the outcome y_t.

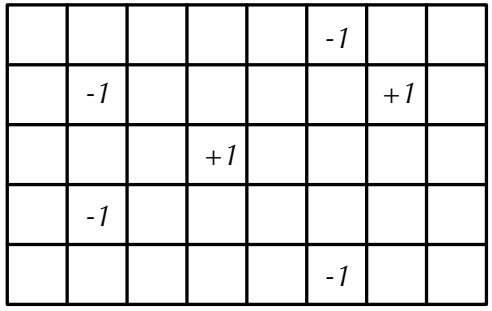
Optimization objective with respect to trace norm:

$$\frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{\widehat{y}_t \neq y_t\} - \inf_{M : \|M\|_\Sigma \leq B} \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{M(x_t) \neq y_t\}$$

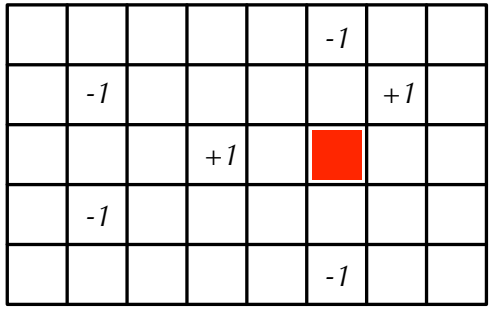
Relaxation:

$$\mathbf{Rel}(x_1, \dots, x_t) = B \; \mathbb{E} \left\| 2 \sum_{s=t+1}^{n} \epsilon_s x_s - \sum_{s=1}^{t} x_s \right\|_\sigma$$

Algorithm



Algorithm



Algorithm

-2	+2	-2	+2	-2	-1	-2	+2
+2	-1	+2	-2	-2	-2	+1	+2
-2	+2	+2	+1	+2		-2	+2
-2	-1	-2	-2	+2	-2	+2	-2
+2	+2	-2	-2	+2	-1	+2	+2

Algorithm

-2	+2	-2	+2	-2	-1	-2	+2
+2	-1	+2	-2	-2	-2	+1	+2
-2	+2	+2	+1	+2		-2	+2
-2	-1	-2	-2	+2	-2	+2	-2
+2	+2	-2	-2	+2	-1	+2	+2

-2	+2	-2	+2	-2	-1	-2	+2
+2	-1	+2	-2	-2	-2	+1	+2
-2	+2	+2	+1	+2		-2	+2
-2	-1	-2	-2	+2	-2	+2	-2
+2	+2	-2	-2	+2	-1	+2	+2

Algorithm

-2	+2	-2	+2	-2	-1	-2	+2
+2	-1	+2	-2	-2	-2	+1	+2
-2	+2	+2	+1	+2	**+1**	-2	+2
-2	-1	-2	-2	+2	-2	+2	-2
+2	+2	-2	-2	+2	-1	+2	+2

-2	+2	-2	+2	-2	-1	-2	+2
+2	-1	+2	-2	-2	-2	+1	+2
-2	+2	+2	+1	+2	**-1**	-2	+2
-2	-1	-2	-2	+2	-2	+2	-2
+2	+2	-2	-2	+2	-1	+2	+2

Algorithm

$$\left\| \begin{array}{|c|c|c|c|c|c|c|c|} \hline -2 & +2 & -2 & +2 & -2 & -1 & -2 & +2 \\ \hline +2 & -1 & +2 & -2 & -2 & -2 & +1 & +2 \\ \hline -2 & +2 & +2 & +1 & +2 & \mathbf{+1} & -2 & +2 \\ \hline -2 & -1 & -2 & -2 & +2 & -2 & +2 & -2 \\ \hline +2 & +2 & -2 & -2 & +2 & -1 & +2 & +2 \\ \hline \end{array} \right\|_{\sigma} - \left\| \begin{array}{|c|c|c|c|c|c|c|c|} \hline -2 & +2 & -2 & +2 & -2 & -1 & -2 & +2 \\ \hline +2 & -1 & +2 & -2 & -2 & -2 & +1 & +2 \\ \hline -2 & +2 & +2 & +1 & +2 & \mathbf{-1} & -2 & +2 \\ \hline -2 & -1 & -2 & -2 & +2 & -2 & +2 & -2 \\ \hline +2 & +2 & -2 & -2 & +2 & -1 & +2 & +2 \\ \hline \end{array} \right\|_{\sigma}$$

Algorithm

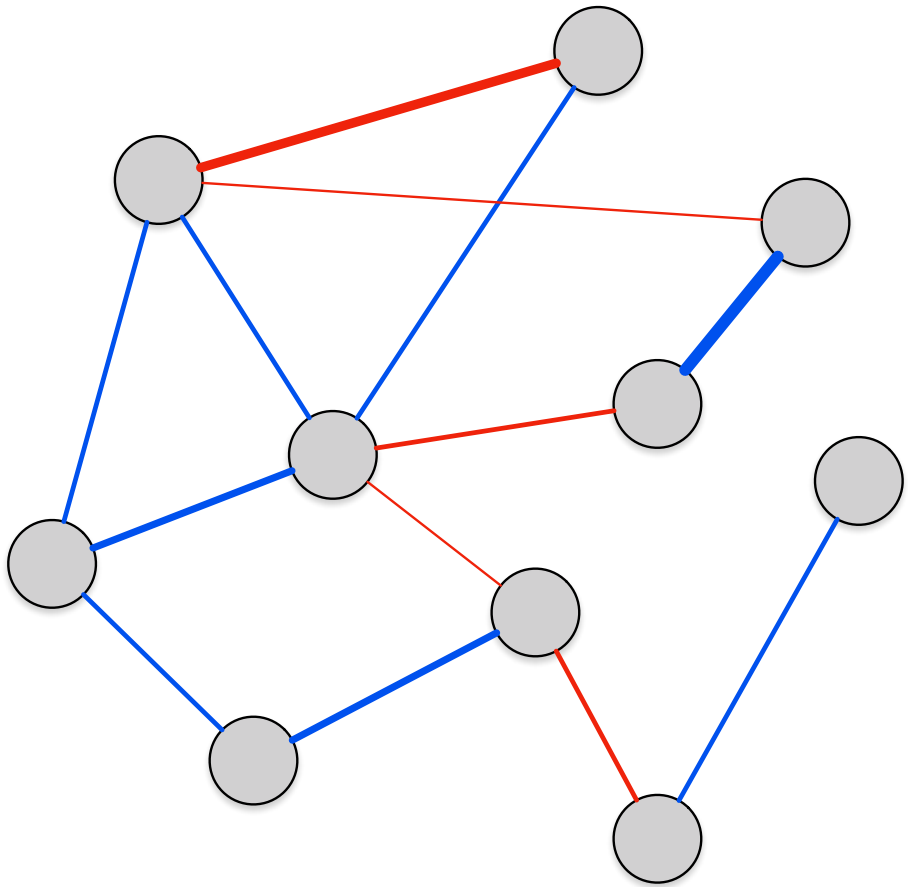
$$\left\| \begin{array}{|cccccccc|} \hline -2 & +2 & -2 & +2 & -2 & -1 & -2 & +2 \\ +2 & -1 & +2 & -2 & -2 & -2 & +1 & +2 \\ -2 & +2 & +2 & +1 & +2 & \mathbf{+1} & -2 & +2 \\ -2 & -1 & -2 & -2 & +2 & -2 & +2 & -2 \\ +2 & +2 & -2 & -2 & +2 & -1 & +2 & +2 \\ \hline \end{array} \right\|_\sigma - \left\| \begin{array}{|cccccccc|} \hline -2 & +2 & -2 & +2 & -2 & -1 & -2 & +2 \\ +2 & -1 & +2 & -2 & -2 & -2 & +1 & +2 \\ -2 & +2 & +2 & +1 & +2 & \mathbf{-1} & -2 & +2 \\ -2 & -1 & -2 & -2 & +2 & -2 & +2 & -2 \\ +2 & +2 & -2 & -2 & +2 & -1 & +2 & +2 \\ \hline \end{array} \right\|_\sigma$$

$$\text{Regret} \leq 2B\ \mathbb{E} \left\| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t x_t \right\|_\sigma \leq O\left(\frac{B\ (\sqrt{m} + \sqrt{n})}{n} \right)$$

Power method works well in practice.

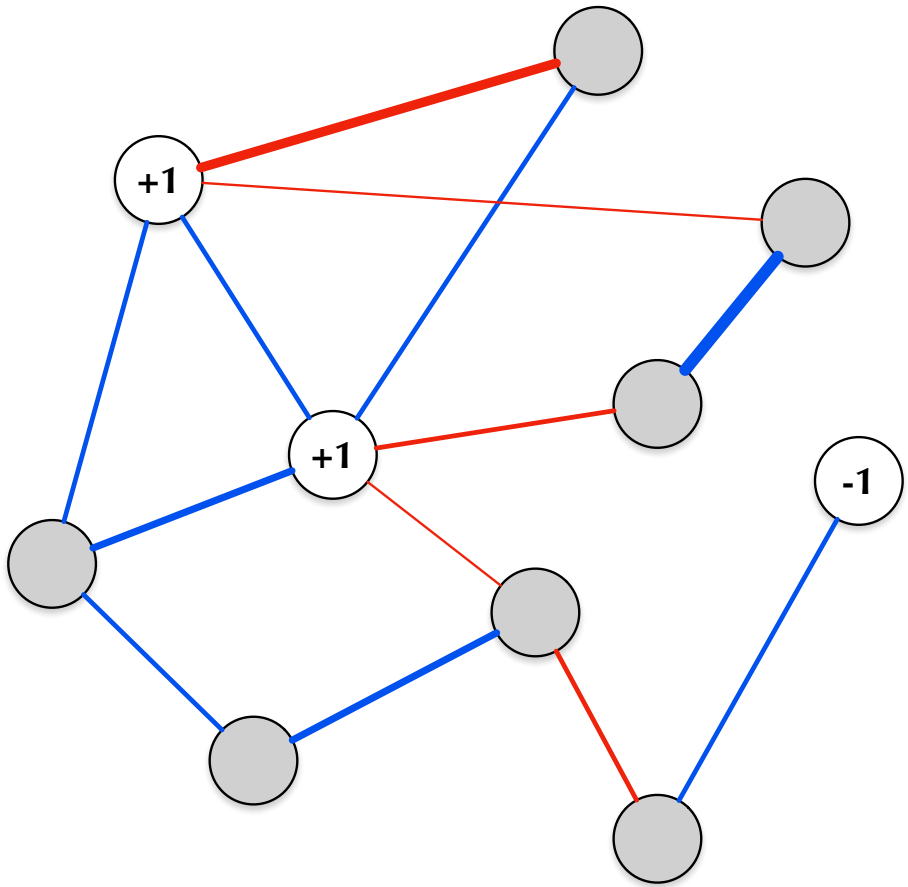
Node prediction

Weighted graph $G = (V, E, W)$, $W : E \to [-1, 1]$



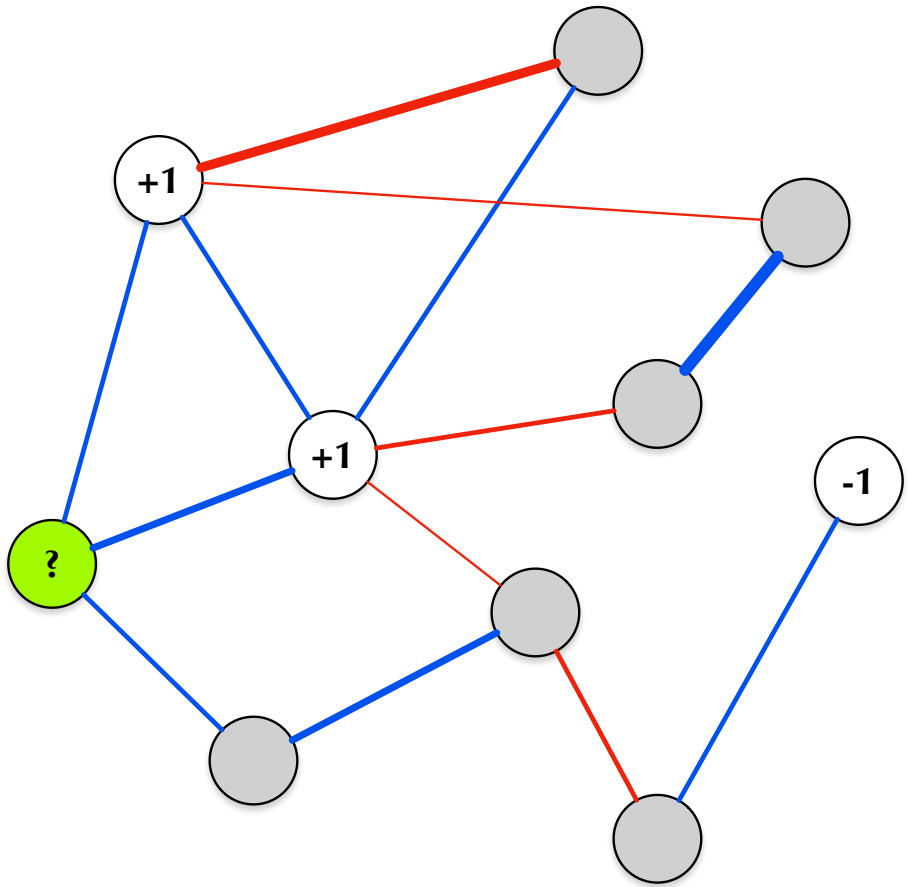
Node prediction

Weighted graph $G = (V, E, W)$, $W : E \to [-1, 1]$



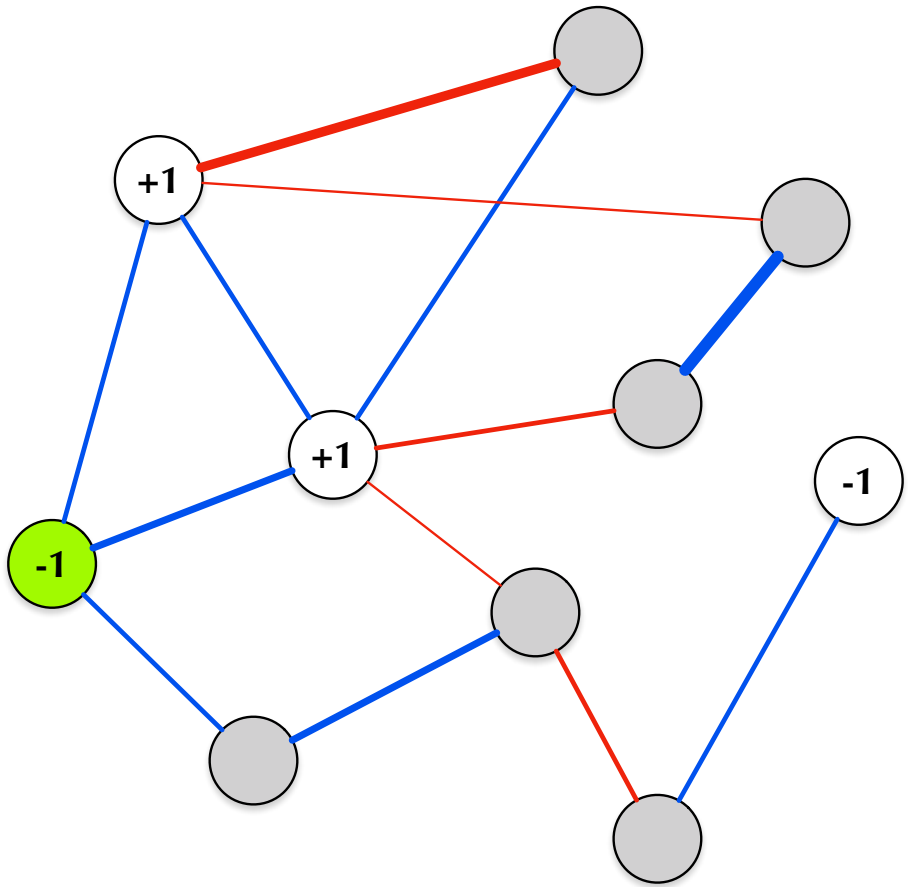
Node prediction

Weighted graph $G = (V, E, W)$, $W : E \to [-1, 1]$



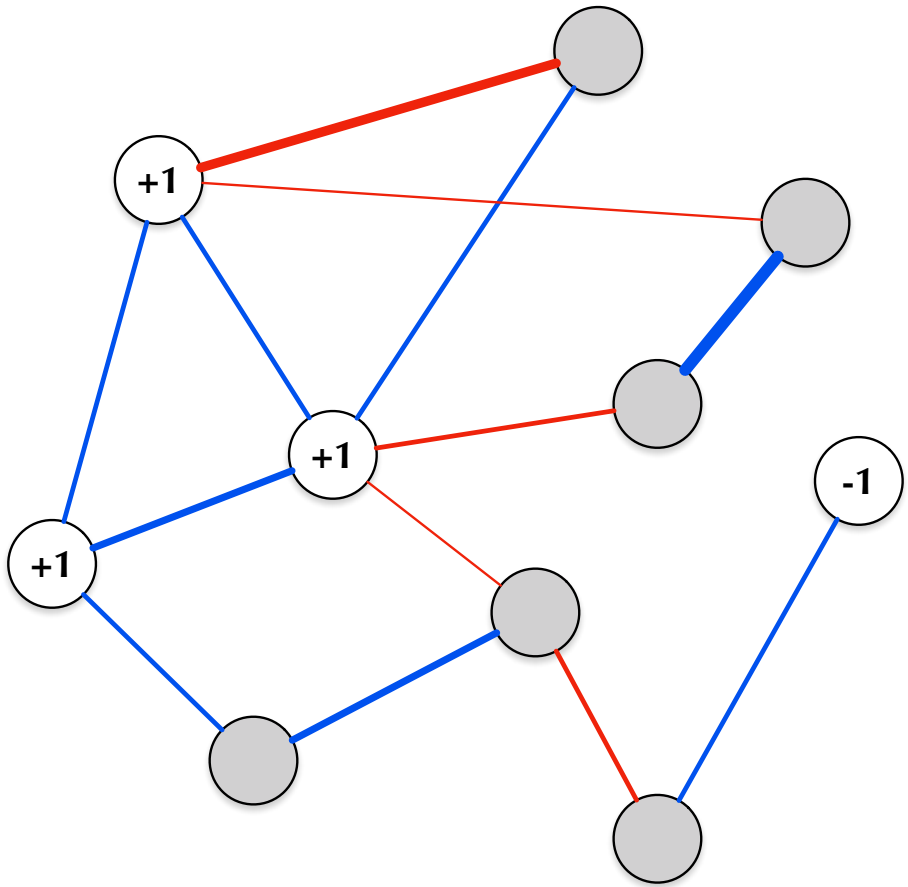
Node prediction

Weighted graph $G = (V, E, W)$, $W : E \to [-1, 1]$



Node prediction

Weighted graph $G = (V, E, W)$, $W : E \to [-1, 1]$



How do we formalize this prediction problem and find an algorithm?

Node prediction

For $t = 1, \ldots, T$
Observe node identity $v_t \in V$
Make randomized prediction $\widehat{y}_t \sim q_t \in \Delta(\{-1, 1\})$
Observe the outcome y_t.

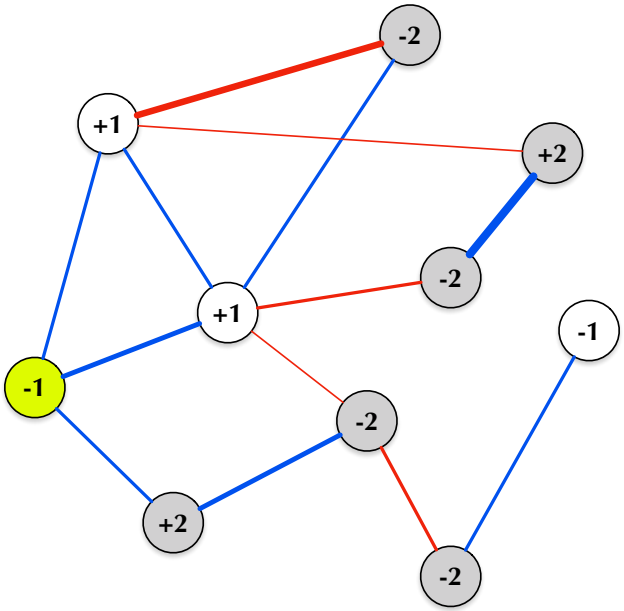
Optimization objective:

$$\frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{\widehat{y}_t \neq y_t\} - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{f(v_t) \neq y_t\}$$

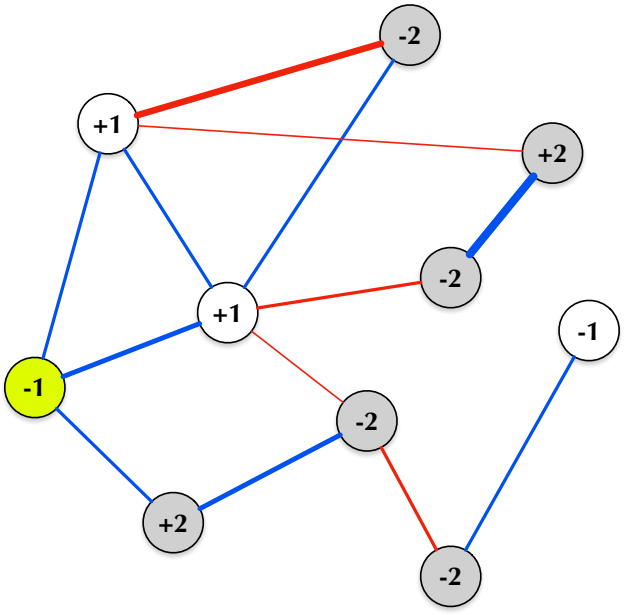
for a class $\mathcal{F} \subseteq \{\pm 1\}^V$ of labelings of vertices.

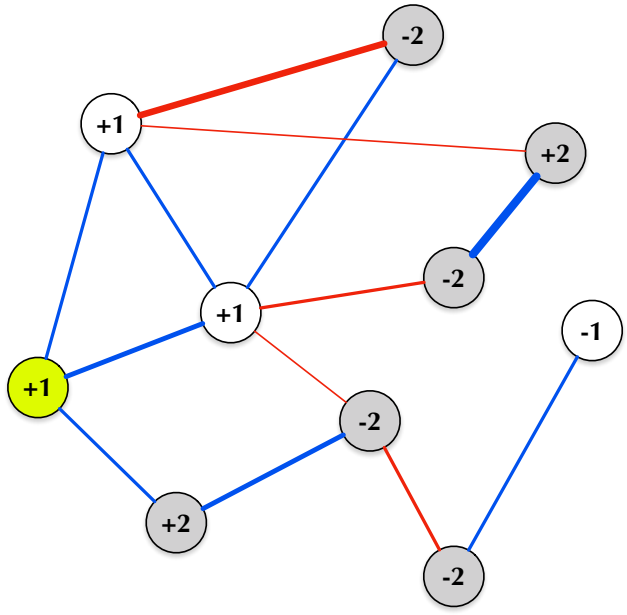
Similar nodes (as measured by W) should have similar labels and dissimilar nodes – different labels.

Algorithm

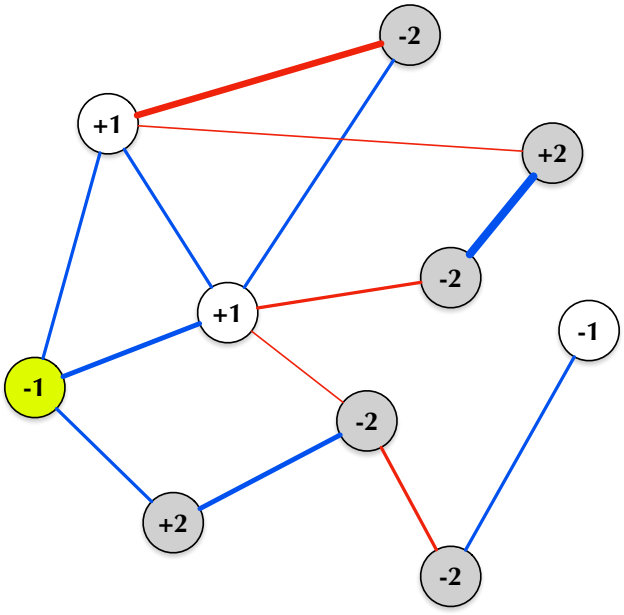


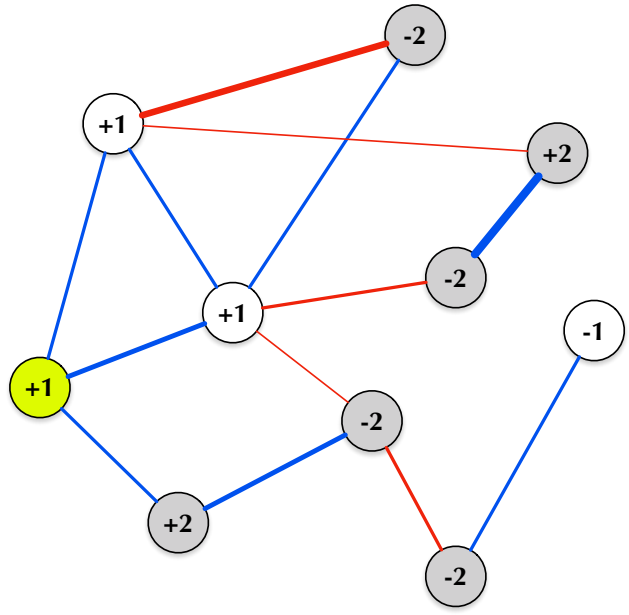
Algorithm





Algorithm





Solve two linear programs and subtract the objective values:

$$\begin{aligned} \mathrm{Val}_t^+ = \text{Maximize} \quad & f^\top X_t^+ \\ \text{subject to} \quad & f \in \mathcal{F} \end{aligned} \qquad \begin{aligned} \mathrm{Val}_t^- = \text{Maximize} \quad & f^\top X_t^- \\ \text{subject to} \quad & f \in \mathcal{F} \end{aligned} \tag{3}$$

Randomized predictor given distribution with mean

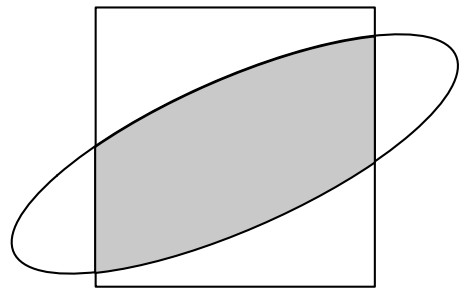
$$q_t = \frac{1}{2}\,\mathrm{Clip}(\mathrm{Val}_t^+ - \mathrm{Val}_t^-)$$

Algorithm (Using Graph Laplacian)

Low label disagreement:

$$\mathcal{F} = \left\{ f \in \{\pm 1\}^{|V|} : \sum_{(u,v)\in E} W_{(u,v)} \left(f(u) - f(v)\right)^2 \leq K \right\} = \left\{ f \in \{\pm 1\}^{|V|} : f^\top L f \leq K \right\}$$

Further relaxation for computational purposes:



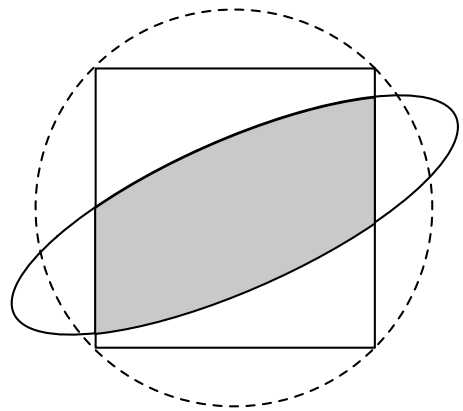
$$\mathcal{F} = \{\pm 1\}^{|V|} \cap \{f : f^\top L f \leq K\}$$

Algorithm (Using Graph Laplacian)

Low label disagreement:

$$\mathcal{F} = \left\{ f \in \{\pm 1\}^{|V|} : \sum_{(u,v)\in E} W_{(u,v)} \left(f(u) - f(v)\right)^2 \leq K \right\} = \left\{ f \in \{\pm 1\}^{|V|} : f^\top L f \leq K \right\}$$

Further relaxation for computational purposes:



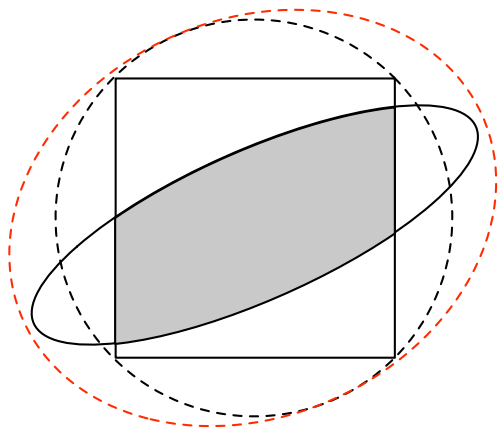
$$\mathcal{F} = \{\pm 1\}^{|V|} \cap \{f : f^\top L f \leq K\}$$

Algorithm (Using Graph Laplacian)

Low label disagreement:

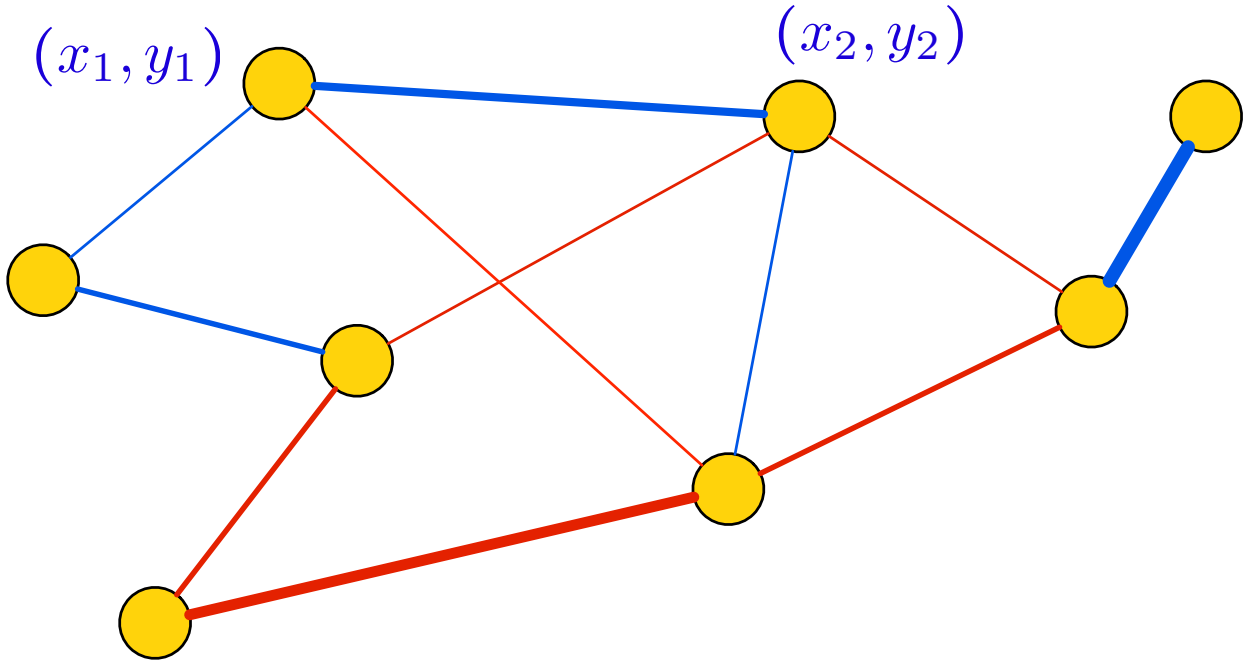
$$\mathcal{F} = \left\{ f \in \{\pm 1\}^{|V|} : \sum_{(u,v)\in E} W_{(u,v)} \left(f(u) - f(v)\right)^2 \leq K \right\} = \left\{ f \in \{\pm 1\}^{|V|} : f^\top L f \leq K \right\}$$

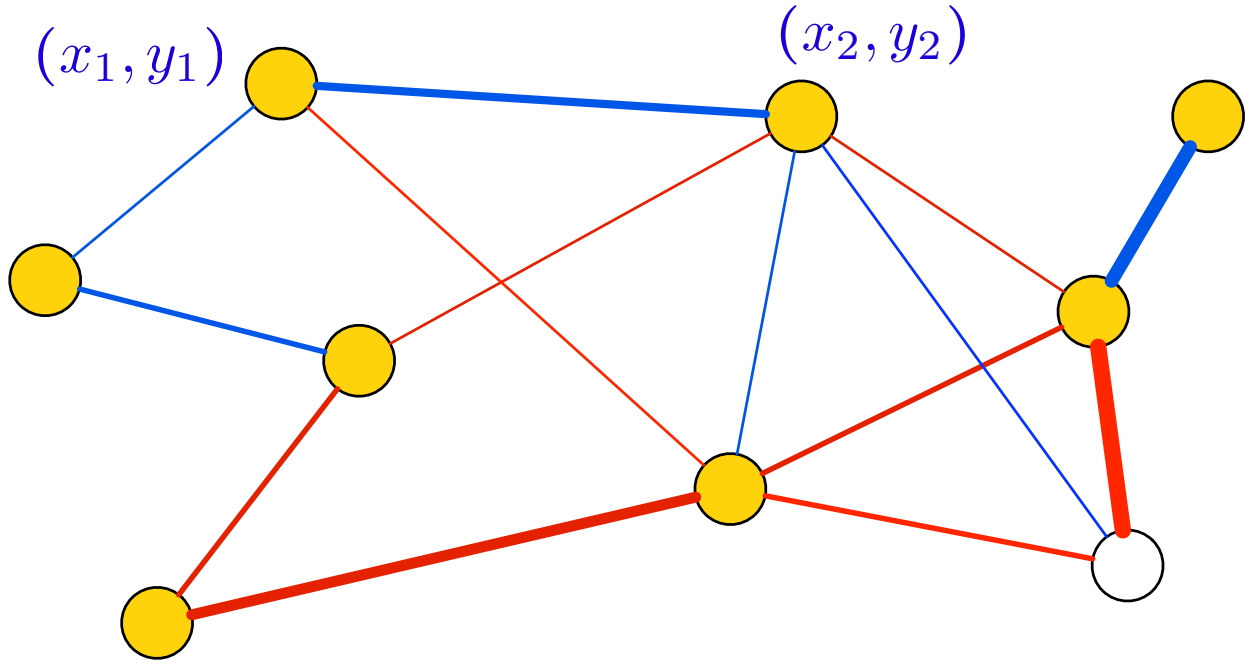
Further relaxation for computational purposes:

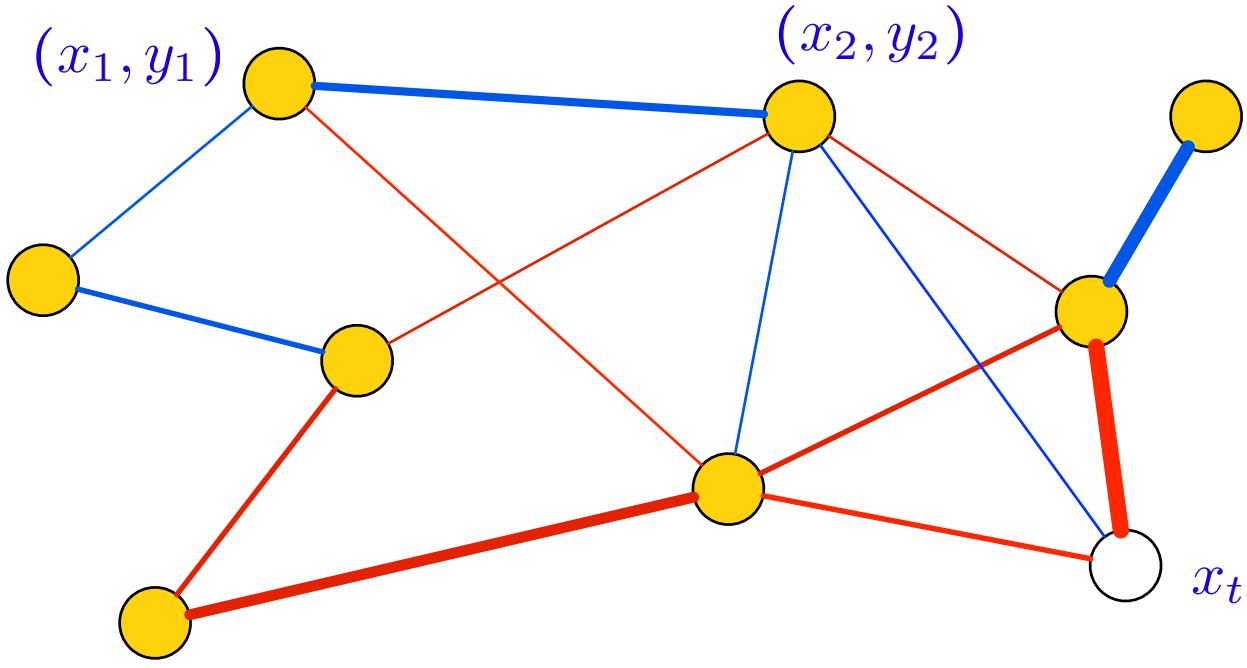


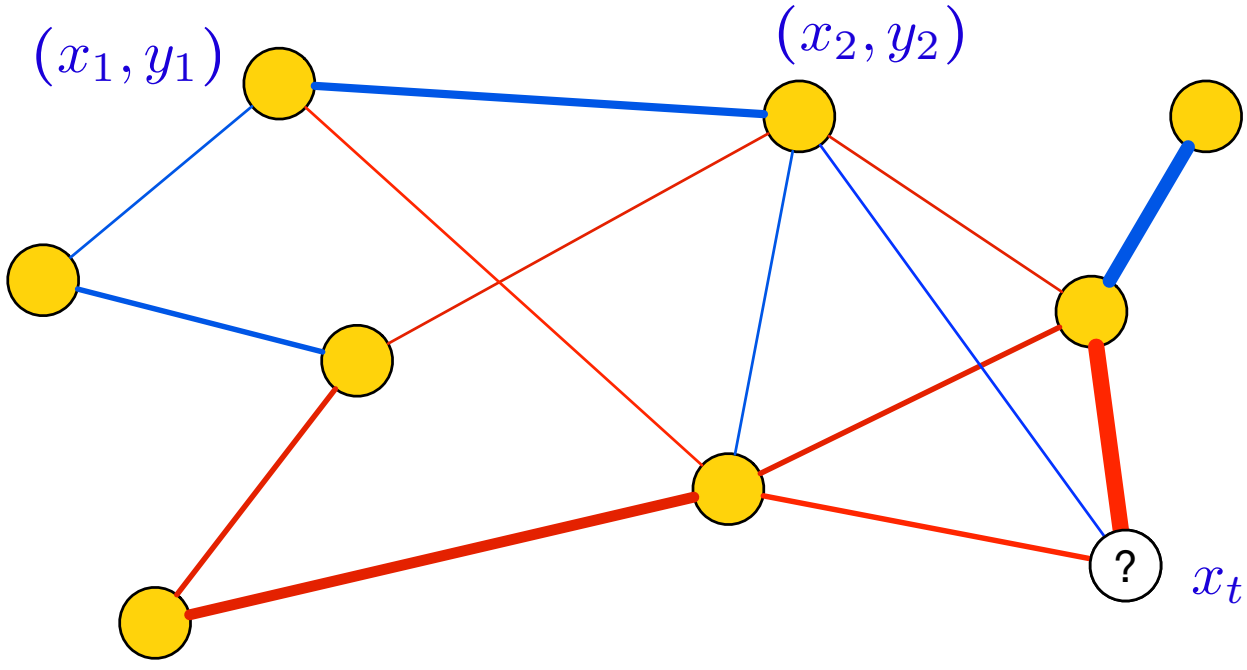
$$\mathcal{F} = \{\pm 1\}^{|V|} \cap \{f : f^\top L f \leq K\}$$

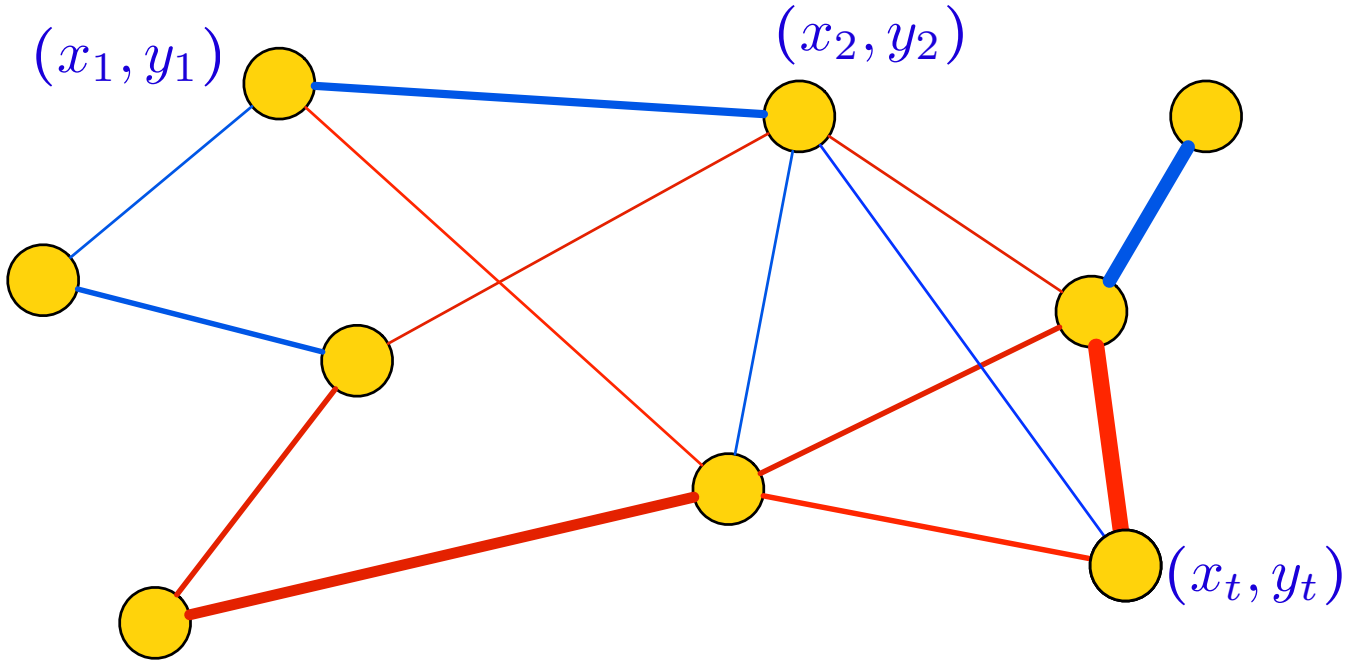
Outline

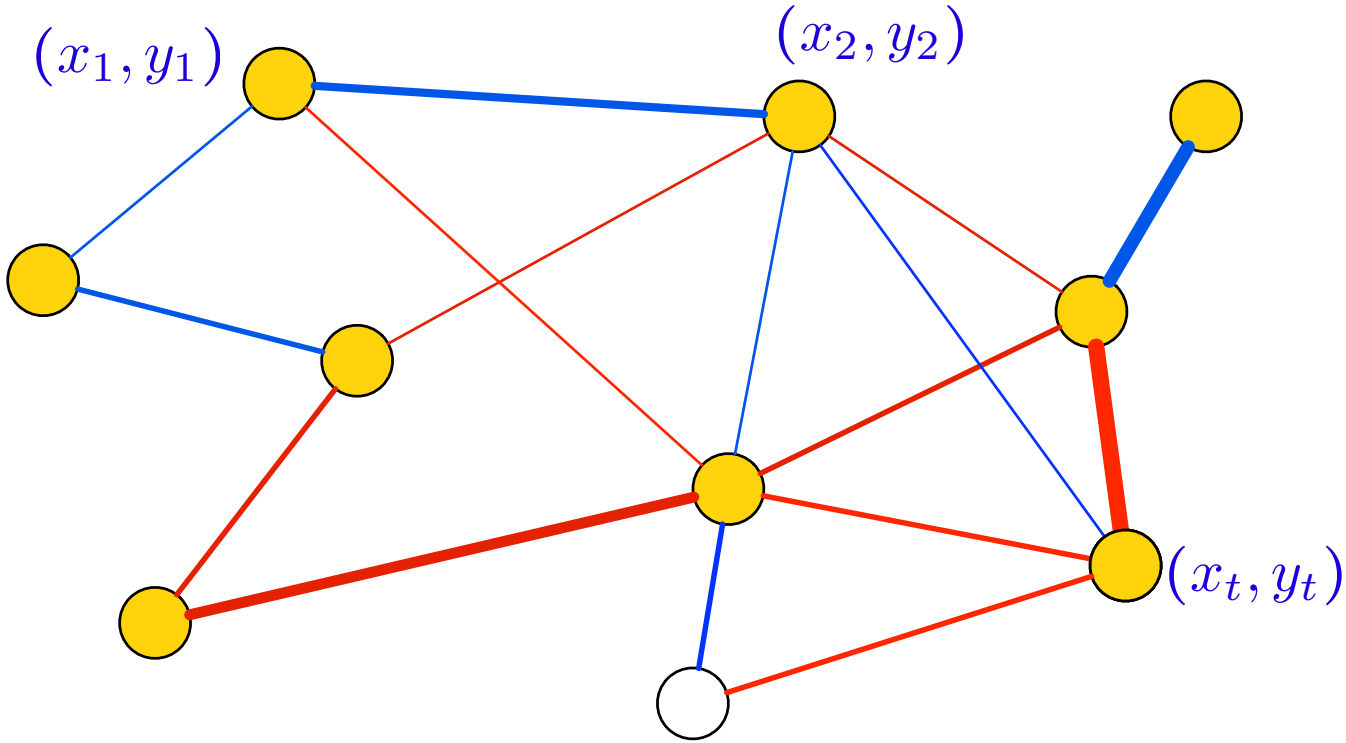

(x_1, y_1)
(x_2, y_2)

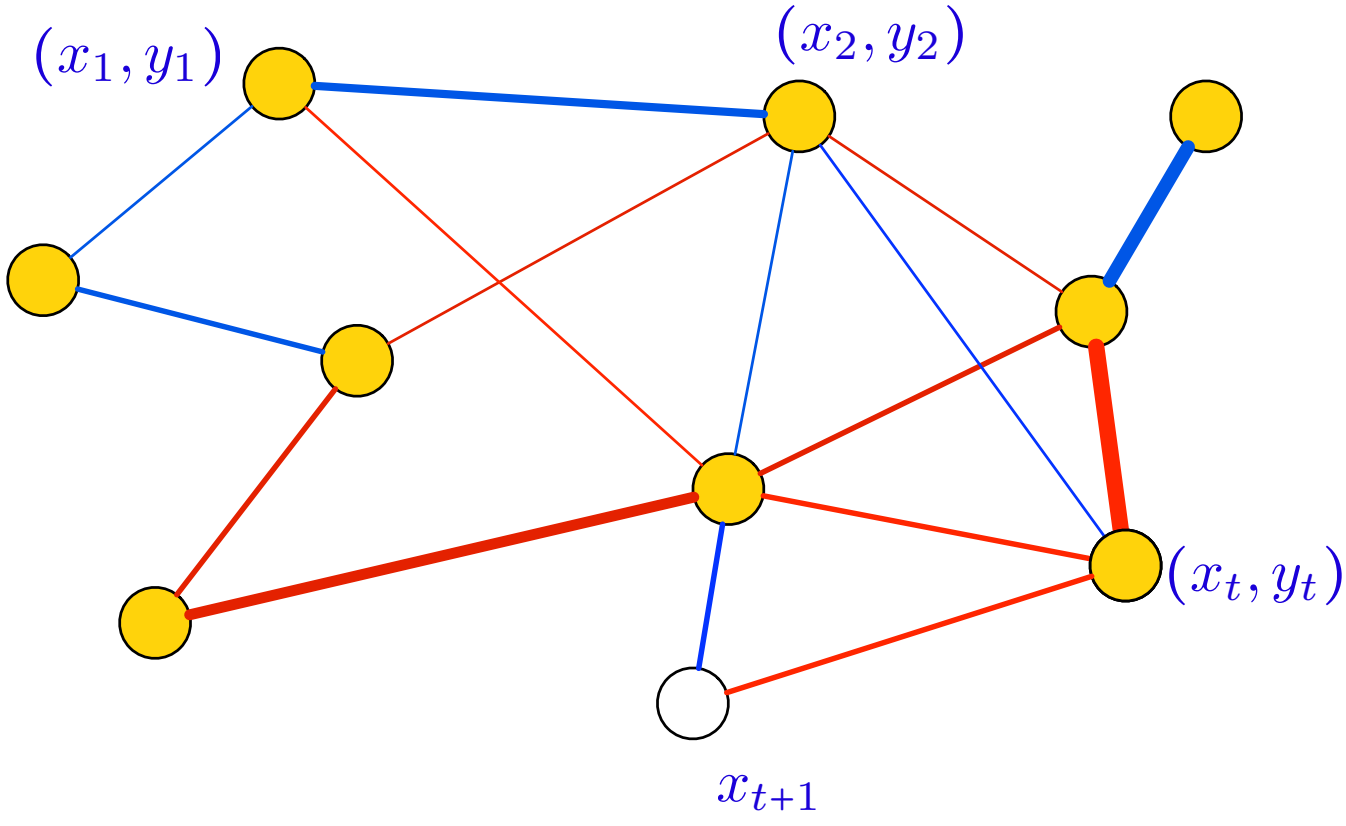

(x_1, y_1)
(x_2, y_2)

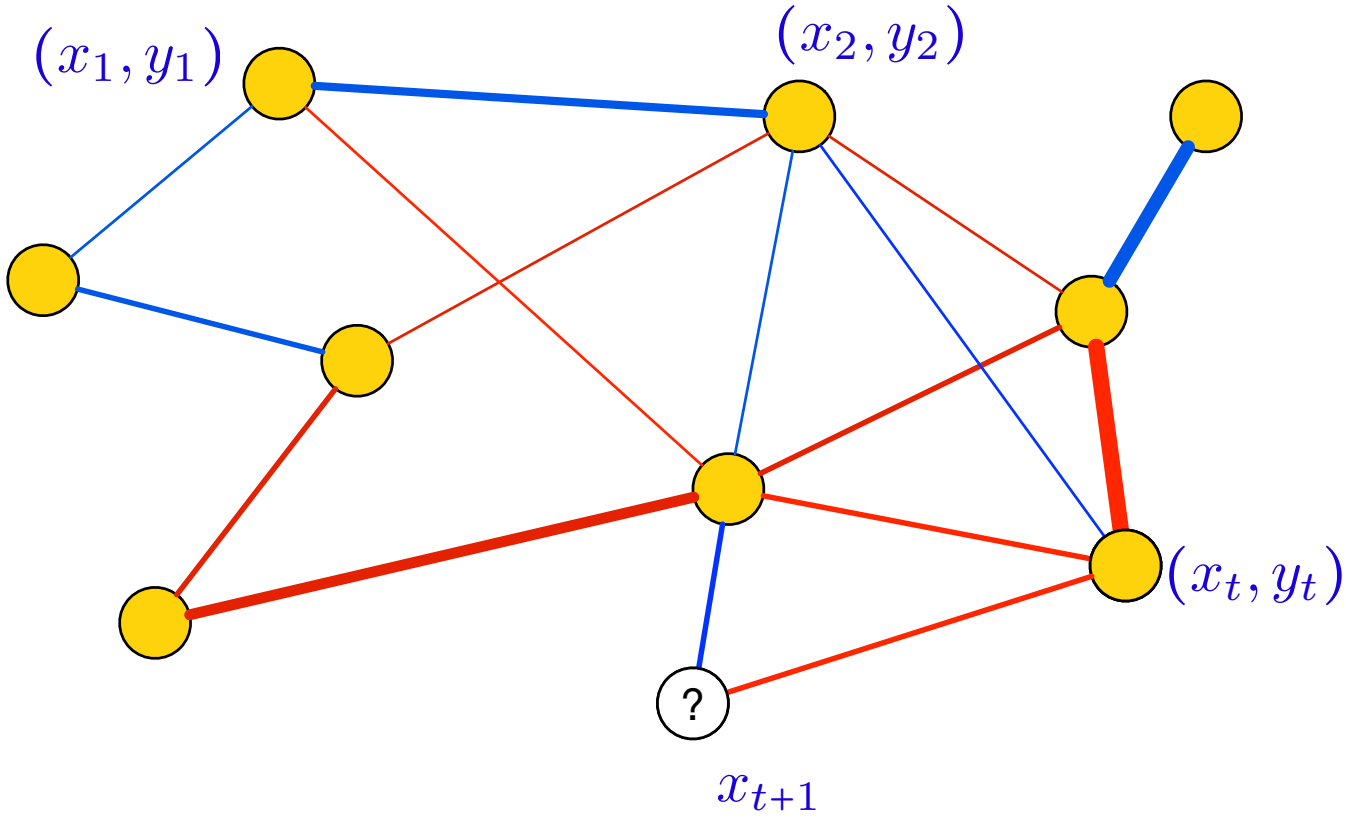

(x_1, y_1)
(x_2, y_2)
x_t

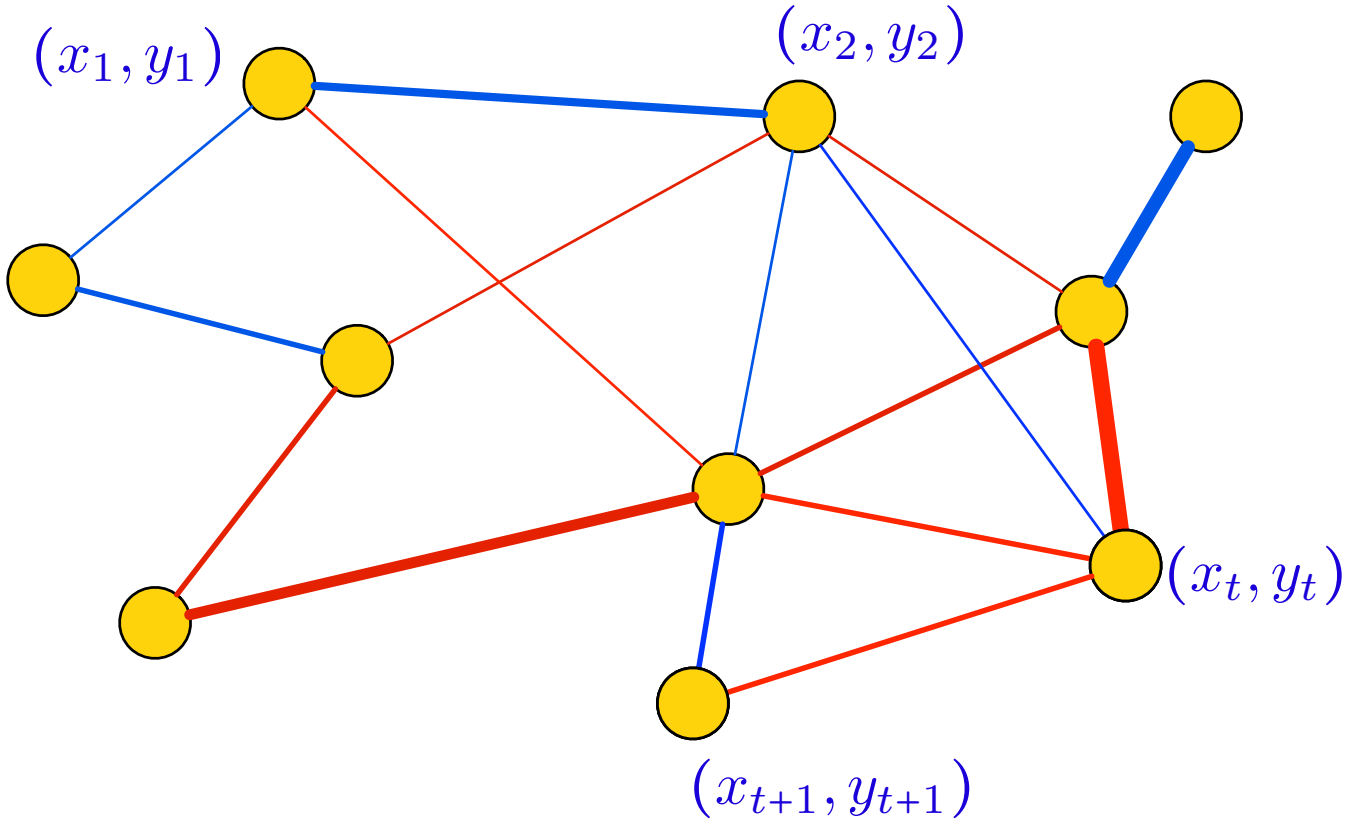

(x_1, y_1)
(x_2, y_2)
?
x_t


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)
x_{t+1}


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)
?
x_{t+1}


(x_1, y_1)
(x_2, y_2)
(x_t, y_t)
(x_{t+1}, y_{t+1})

At time t,

- x_t and a set of constraints $\mathcal{C}_t$ is revealed
- $\widehat{y}_t \in \{1, \ldots, k\}$ is made
- outcome $y_t \in \{1, \ldots, k\}$ is revealed

- $n = V$
- Assume we know generating process for $(x_t, \mathcal{C}_t)$
- Each constraint measures affinity of labeling for a group of nodes (e.g. whether labels of u, v match for edge (u, v))
- $(f(x_1), \ldots, f(x_n))$ are labels of all nodes at the end
- F is a set of labelings on x_t's that violate at most K constraints at the end of $n = V$ rounds

$$F(x_{1:n}, \mathcal{C}_{1:n}) = \left\{ f \in \mathcal{F}|_{x_{1:n}} : \sum_{c \in \cup \mathcal{C}_t} c(f) \leq K \right\}$$

Regret

$$\sum_{t=1}^{n} \mathbf{I}\{\widehat{y}_t \neq y_t\} - \inf_{f \in F(x_{1:n}, \mathcal{C}_{1:n})} \sum_{t=1}^{n} \mathbf{I}\{f_t \neq y_t\}$$

is against a time-changing target.

Randomized method:

- Observe $x_t, \mathcal{C}_t$
- Randomly generate $x_{t+1:n}, \mathcal{C}_{t+1:n}$ (or use unlabeled data)
- Approximately solve a version of CSP (constraint satisfaction problem) using semidefinite relaxations
- Use this approximate value in the relaxation framework to compute prediction

The power of improper learning

- Computing an offline solution (even approximately) is NP-hard in many interesting cases
- We do not need to round the solution but only need the approximate value of the relaxation per step
- Integrality gap multiplies the final regret bound, not the OPT
- *Lasserre hierarchy to trade off computation and prediction performance*