Sequential complexities and uniform martingale laws of large numbers

Ambuj Tewari

(based on joint work with Alexander Rakhlin and Karthik Sridharan)

Department of Statistics, and
Department of EECS,
University of Michigan, Ann Arbor

November 15, 2014

Some Prediction Problems

- Will a friendship relation form between two Facebook users?
- Which ads should Google show me when I search for `flights to Mexico`?
- 507,000 webpages match `game-theoretic probability`: in which order should Google show them to me?
- Should Gmail put the email with subject `FREE ONLINE COURSES!!!` in the spam folder?

Mathematical Formulation of Prediction Problems

- Input space $\mathcal{X}$ (vectors, matrices, text, graphs)
- Label space $\mathcal{Y}$
  - (CLASSIFICATION) $\mathcal{Y} = \{\pm 1\}$
  - (REGRESSION) $\mathcal{Y} = [-1, +1]$
  - (RANKING) $\mathcal{Y} = S_k$, group of k-permutations
- Want to learn a prediction function $f : \mathcal{X} \to \mathcal{Y}$
- Loss function: how bad is prediction $f(x)$ if "truth" is y

Predictions and Losses

- Learner/Statistician/Decision Maker chooses prediction function $f : \mathcal{X} \to \mathcal{Y}$
- Adversary/Nature/Environment produces examples $(x, y) \in \mathcal{X} \times \mathcal{Y}$
- Learner's loss $\ell(f(x), y)$
- Assume ℓ is bounded

Probabilistic Approach

- (x_t, y_t) are drawn from a stochastic process
- For instance, (x_t, y_t) i.i.d. from some distribution P
- Parametric case: $P = P_\theta$ with $\theta \in \Theta \subseteq \mathbb{R}^p$
- Distribution free or "agnostic" case: P arbitrary

Goal: Choose $\widehat{f}$ based on the sample $((x_t, y_t))_{t=1}^n$ to have small expected loss

$$\mathbb{E}_{x_{1:n}, y_{1:n}, x, y \sim P}\left[\ell(\widehat{f}(x), y)\right]$$

Empirical Risk Minimization

- Risk and empirical risk

$$L(f) = \mathbb{E}_{(x,y)\sim P}\left[\ell(f(x), y)\right] \qquad \widehat{L}(f) = \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t)$$

- Risk minimizer

$$f^\star = \operatorname*{argmin}_{f\in\mathcal{F}} \; L(f)$$

- Empirical risk minimizer (ERM)

$$\widehat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \; \widehat{L}(f)$$

- Excess risk

$$L(\widehat{f}) - L(f^\star)$$

Game Theoretic Approach

- **FOR** $t = 1$ to n
  - Adversary plays $x_t \in \mathcal{X}$
  - Learner plays $f_t \in \mathcal{F}$
  - Adversary plays $y_t \in \mathcal{Y}$
  - Learner suffers $\ell(f_t(x_t), y_t)$
- **ENDFOR**

- No assumption on data generating mechanism
- Want to "do well" on every sequence $(x_1, y_1), \ldots, (x_n, y_n)$

Goal: Tricky to define

Regret

- Measure learner's loss relative to some benchmark computed in hindsight
- (External) Regret

$$\sum_{t=1}^{n} \ell(f_t(x_t), y_t) - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)$$

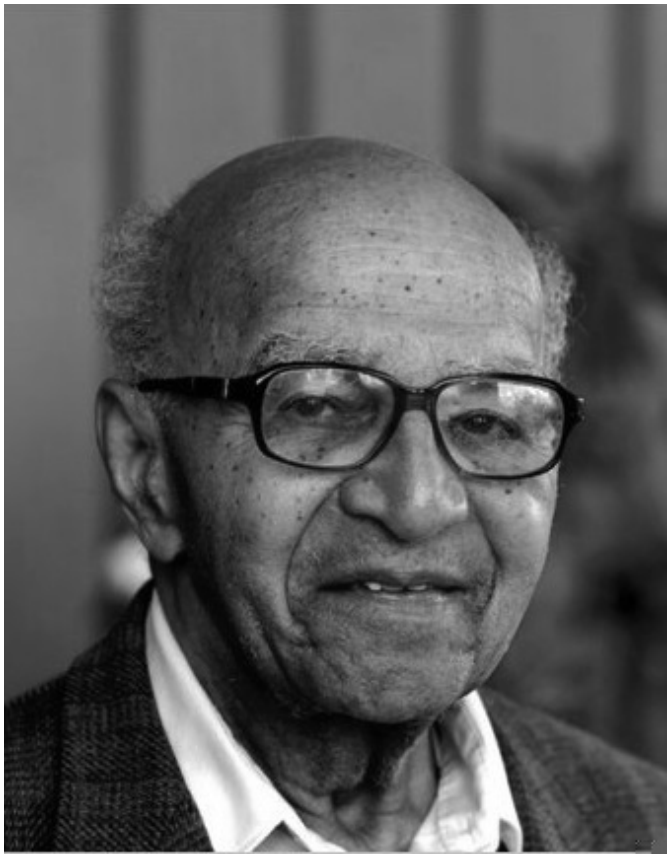
- Benchmark here is the best fixed decision in hindsight
- Many variants exist (switching regret, Φ-regret)

Why Study Regret?

- Lets us proceed with no assumptions on the data generating process
- Regret-minimizing algorithms perform well if data is i.i.d.
- Yields simple one-pass algorithms
- If players in a game follow regret-minimizing algorithms, the empirical distribution of play converges to an equilibrium
- Long history in Computer Science, Finance, Game Theory, Information Theory, and Statistics

Two pioneers

James Hannan (1922-2010)



David Blackwell (1919-2010)

Simplest Case: Finite Class of Functions

- $|\mathcal{F}| = K$
- HANNAN'S THEOREM. There is a (randomized) learner strategy for which

$$\text{(expected) regret} = o(n)$$

- "no-regret learning" or "Hannan consistency": when regret $= o(n)$

Multiple Discovery

- Originally proved by Hannan (1956)
- Blackwell (1956) showed how it follows from his approachability theorem
- Result has been proven many times since then:
  - Banos (1968)
  - Cover (1991)
  - Foster & Vohra (1993)
  - Vovk (1993)

Rest of the Talk

- Rademacher complexity and its sequential analog
- Fat-shattering dimension and its sequential analog
- Uniform martingale law of large numbers

Rademacher Complexity

- Recall ERM $\widehat{f}$, RM $f^\star$

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{L}(f) \qquad\qquad f^\star = \operatorname*{argmin}_{f \in \mathcal{F}} L(f)$$

- Easy to show

$$\mathbb{E}\left[L(\widehat{f}) - L(f^\star)\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} L(f) - \widehat{L}(f)\right]$$

- Symmetrization (ϵ_t's are Rademacher, i.e. symmetric Bernoulli)

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} L(f) - \widehat{L}(f)\right] \leq 2\,\mathbb{E}_{\epsilon_{1:n}, x_{1:n}, y_{1:n}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right]$$

Which Algorithm Should We Analyze?

- Obvious analogue of ERM is "follow-the-leader" or "fictitious play":

$$f_{t+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{s=1}^{t} \ell(f(x_s), y_s)$$

- Does not enjoy good regret bound
- Lack of a generic regret-minimizing strategy is a problem
- Directly attack minimax regret

Minimax Regret

Minimax regret:

$$V_n := \min_{\substack{\text{Learner} \\ \text{strategies}}} \max_{\substack{\text{Adversary} \\ \text{strategies}}} \mathbb{E}\left[\sum_{t=1}^{n} \ell(f_t(x_t), y_t) - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right]$$

Theorem (Rakhlin, Sridharan, Tewari (2010))

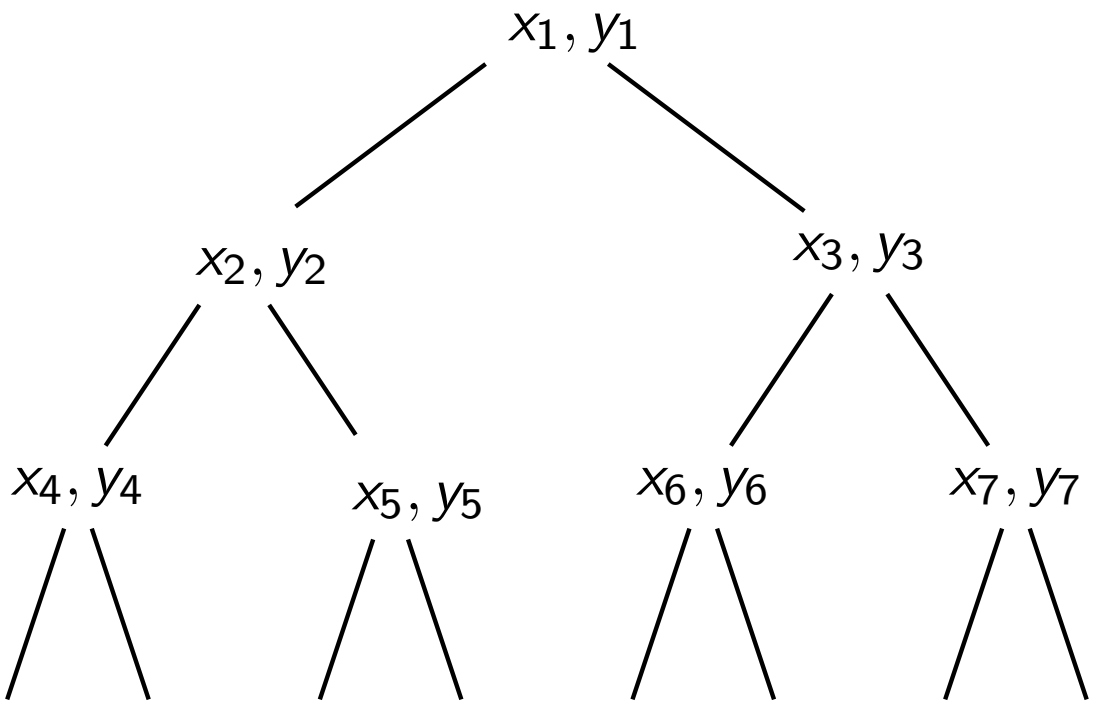
$$V_n \leq 2\,\mathcal{R}_n^{\text{seq}}$$

Important precursor: Abernethy et al. (2009)

Sequential Rademacher Complexity

$$\mathcal{R}_n^{\text{seq}} := \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell\big(f\big(\mathbf{x}(\epsilon_{1:t-1})\big), \mathbf{y}(\epsilon_{1:t-1})\big) \right]$$

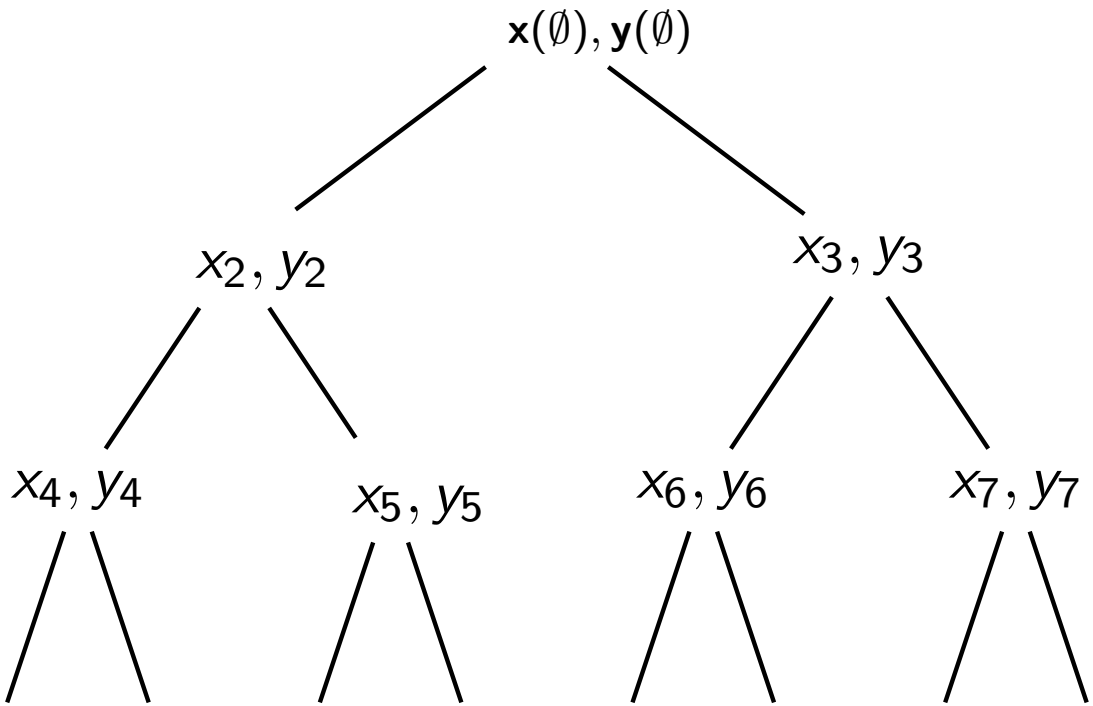
Tree $\mathbf{x}, \mathbf{y}$



Sequential Rademacher Complexity

$$\mathcal{R}_n^{\text{seq}} := \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell\big(f\big(\mathbf{x}(\epsilon_{1:t-1})\big), \mathbf{y}(\epsilon_{1:t-1})\big) \right]$$

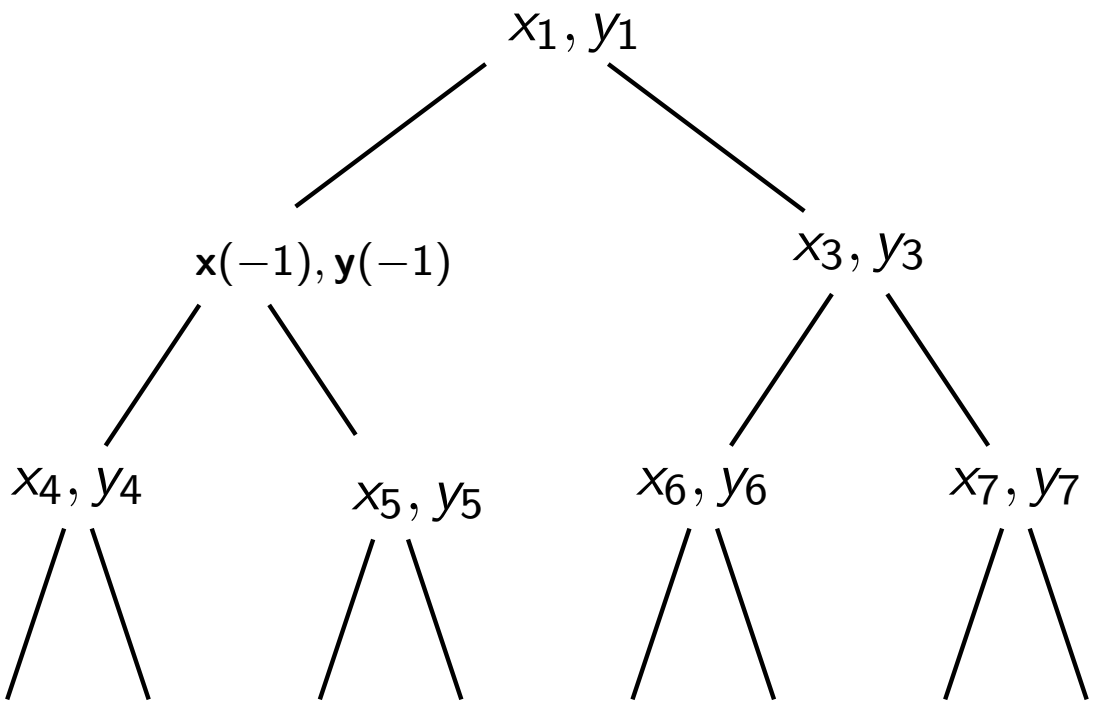
Tree $\mathbf{x}, \mathbf{y}$



Sequential Rademacher Complexity

$$\mathcal{R}_n^{\text{seq}} := \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell\big(f\big(\mathbf{x}(\epsilon_{1:t-1})\big), \mathbf{y}(\epsilon_{1:t-1})\big) \right]$$

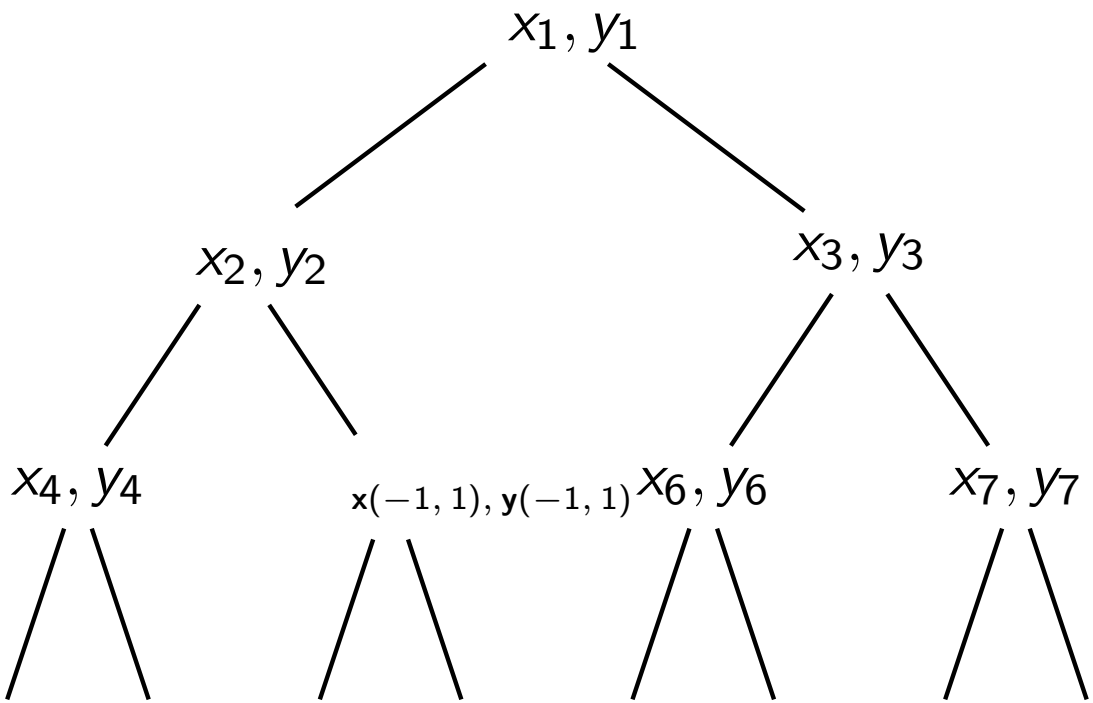
Tree $\mathbf{x}, \mathbf{y}$



Sequential Rademacher Complexity

$$\mathcal{R}_n^{\text{seq}} := \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell\big(f\big(\mathbf{x}(\epsilon_{1:t-1})\big), \mathbf{y}(\epsilon_{1:t-1})\big) \right]$$

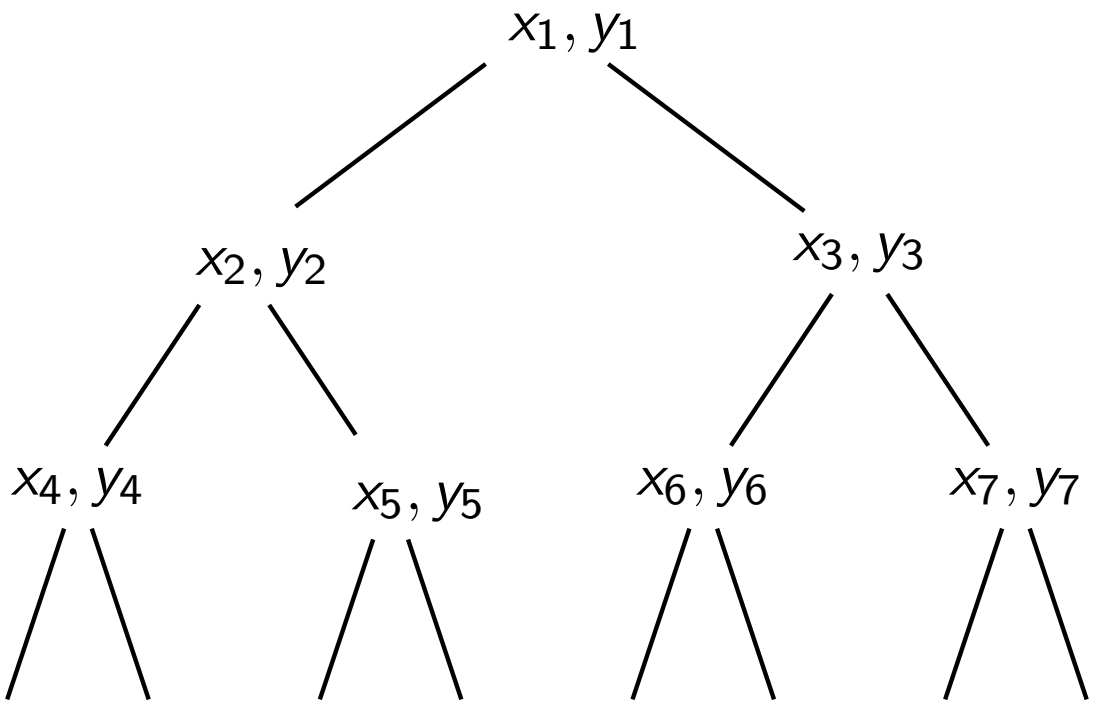
Tree $\mathbf{x}, \mathbf{y}$



Sequential Rademacher Complexity

$$\mathcal{R}_n^{\text{seq}} := \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell\big(f\big(\mathbf{x}(\epsilon_{1:t-1})\big), \mathbf{y}(\epsilon_{1:t-1})\big) \right]$$

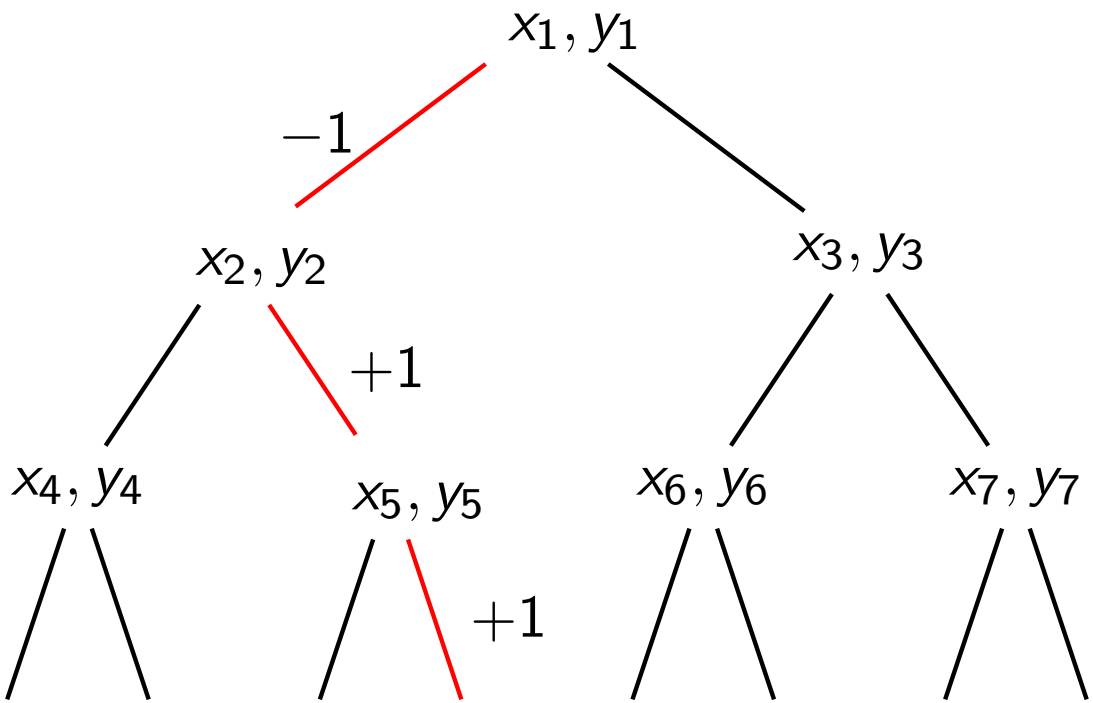
Tree $\mathbf{x}, \mathbf{y}$



Sequential Rademacher Complexity

$$\mathcal{R}_n^{\text{seq}} := \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell\big(f\big(\mathbf{x}(\epsilon_{1:t-1})\big), \mathbf{y}(\epsilon_{1:t-1})\big) \right]$$

Tree $\mathbf{x}, \mathbf{y}$



Rademacher Complexity: Classical vs. Sequential

$$\mathcal{R}_n(\ell \circ \mathcal{F}) := \mathbb{E}_{\epsilon_{1:n}, x_{1:n}, y_{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t) \right]$$

$$\mathcal{R}_n^{\mathsf{seq}}(\ell \circ \mathcal{F}) := \sup_{\mathbf{x}, \mathbf{y}} \mathbb{E}_{\epsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(\mathbf{x}(\epsilon_{1:t-1})), \mathbf{y}(\epsilon_{1:t-1})) \right]$$

- Sequences $x_{1:n}, y_{1:n}$ replaced by tree $\mathbf{x}, \mathbf{y}$
- Expectation over sequences $x_{1:n}, y_{1:n}$ replaced by supremum over trees $\mathbf{x}, \mathbf{y}$

Seq. Rademacher Complexity: Properties

- (INCLUSION) If $\mathcal{F} \subseteq \mathcal{F}'$ then

$$\mathcal{R}_n^{\mathsf{seq}}(\ell \circ \mathcal{F}) \leq \mathcal{R}_n^{\mathsf{seq}}(\ell \circ \mathcal{F}')$$

- (SCALING) If $c \in \mathbb{R}$ then

$$\mathcal{R}_n^{\mathsf{seq}}(c\ell \circ \mathcal{F}) = |c| \cdot \mathcal{R}_n^{\mathsf{seq}}(\ell \circ \mathcal{F})$$

- (TRANSLATION) If $\ell' = \ell + h$ then

$$\mathcal{R}_n^{\mathsf{seq}}(\ell \circ \mathcal{F}) = \mathcal{R}^{\mathsf{seq}}(\ell' \circ \mathcal{F})$$

Seq. Rademacher Complexity: Properties

- (INCLUSION) If $\mathcal{F} \subseteq \mathcal{F}'$ then

$$\mathcal{R}_n^{\text{seq}}(\ell \circ \mathcal{F}) \leq \mathcal{R}_n^{\text{seq}}(\ell \circ \mathcal{F}')$$

- (SCALING) If $c \in \mathbb{R}$ then

$$\mathcal{R}_n^{\text{seq}}(c\ell \circ \mathcal{F}) = |c| \cdot \mathcal{R}_n^{\text{seq}}(\ell \circ \mathcal{F})$$

- (TRANSLATION) If $\ell' = \ell + h$ then

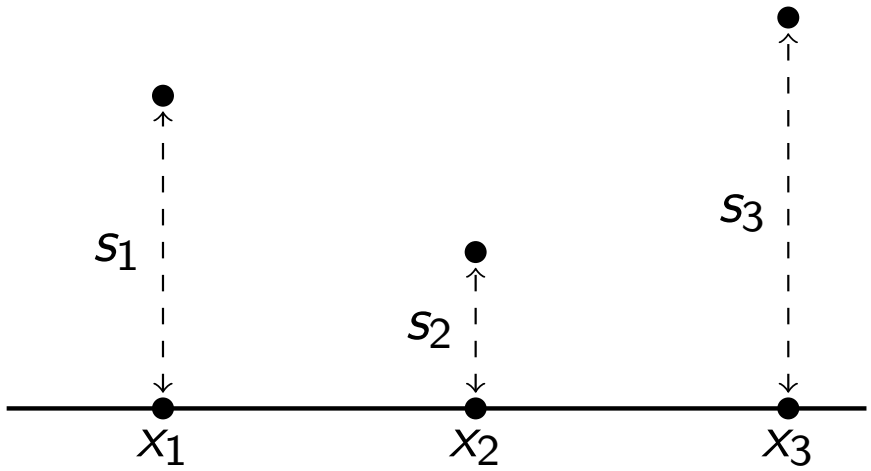
$$\mathcal{R}_n^{\text{seq}}(\ell \circ \mathcal{F}) = \mathcal{R}^{\text{seq}}(\ell' \circ \mathcal{F})$$

Using these and other properties, possible to bound seq. Rademacher complexity of decision trees, neural networks, etc.

Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

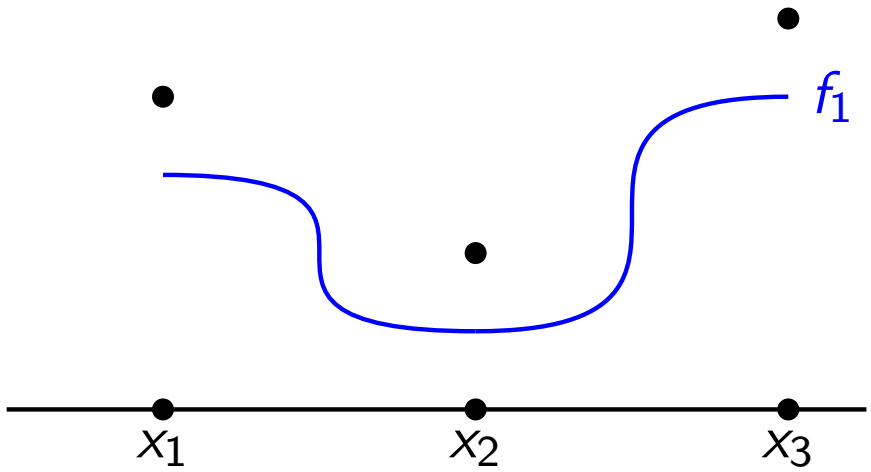
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

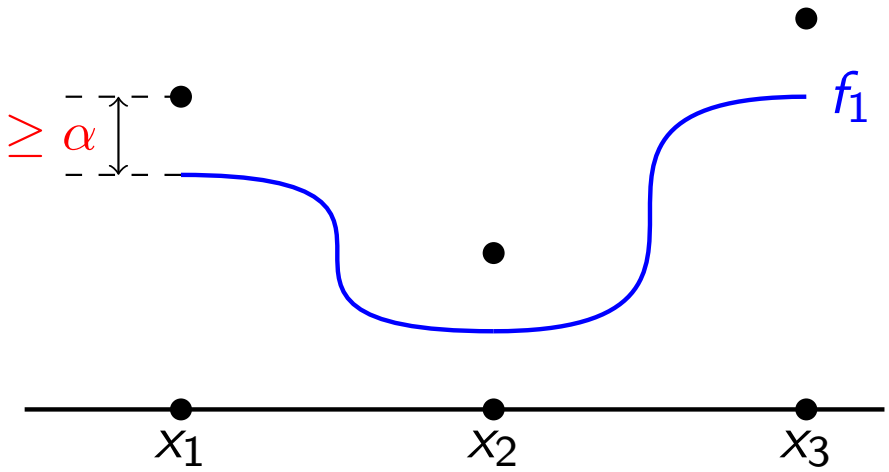
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

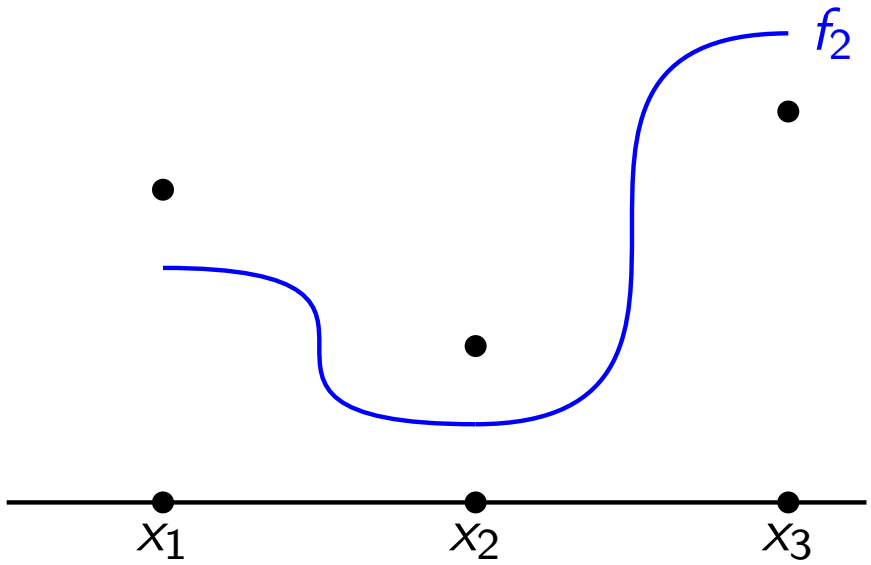
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

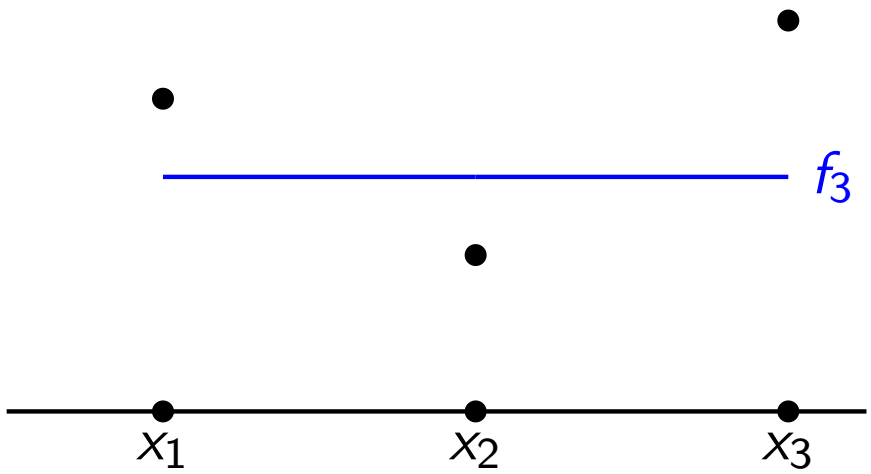
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

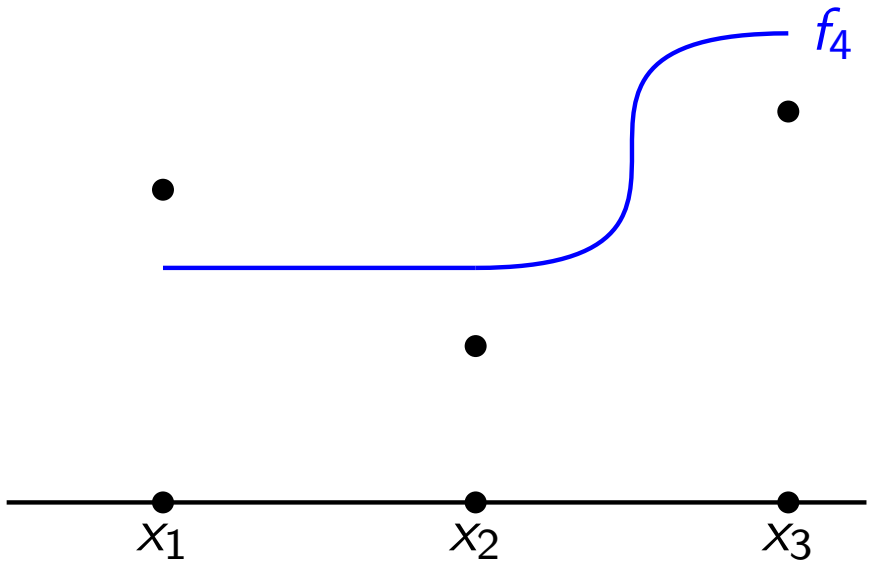
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

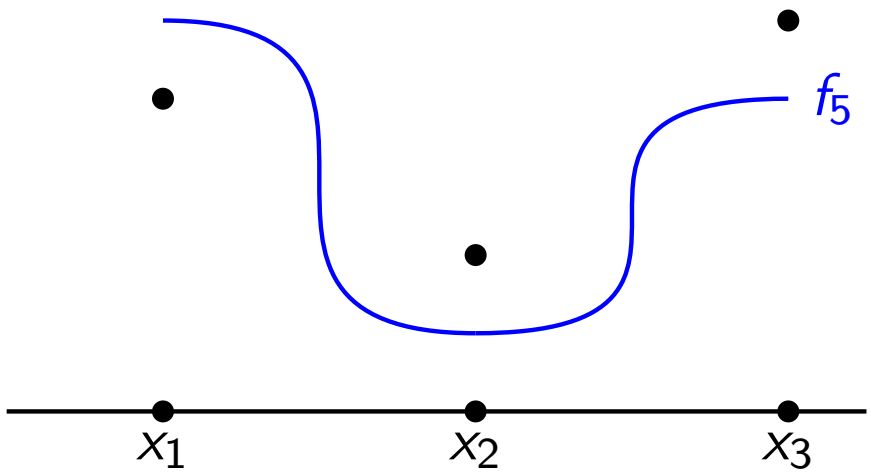
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

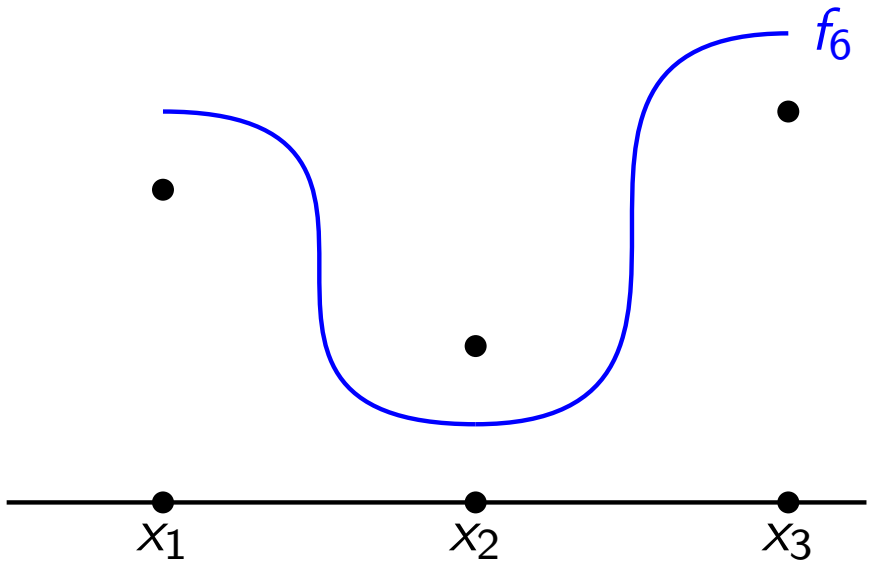
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

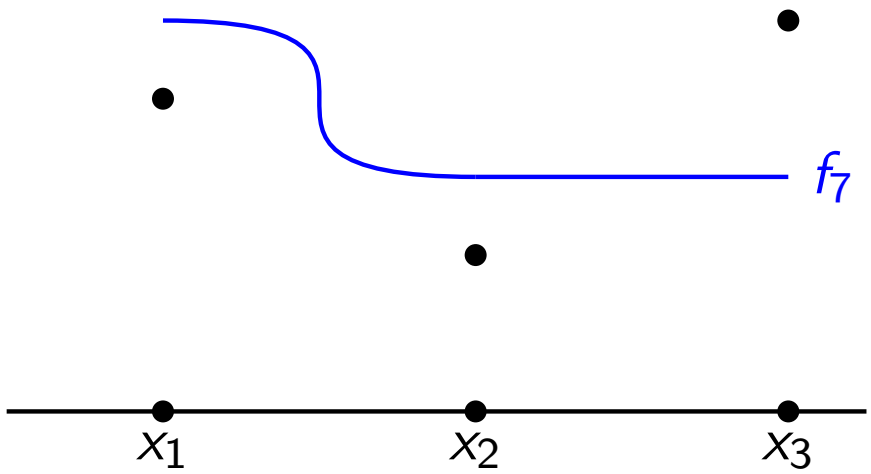
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

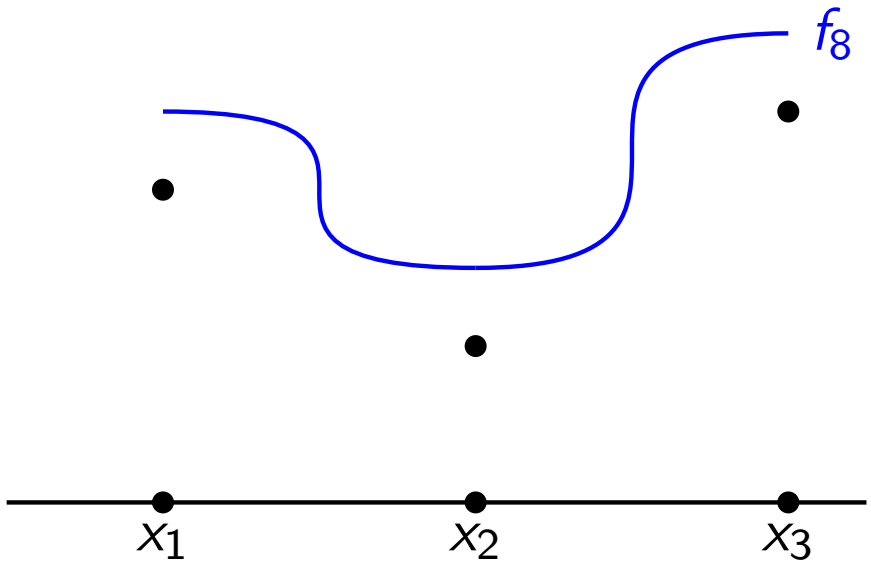
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

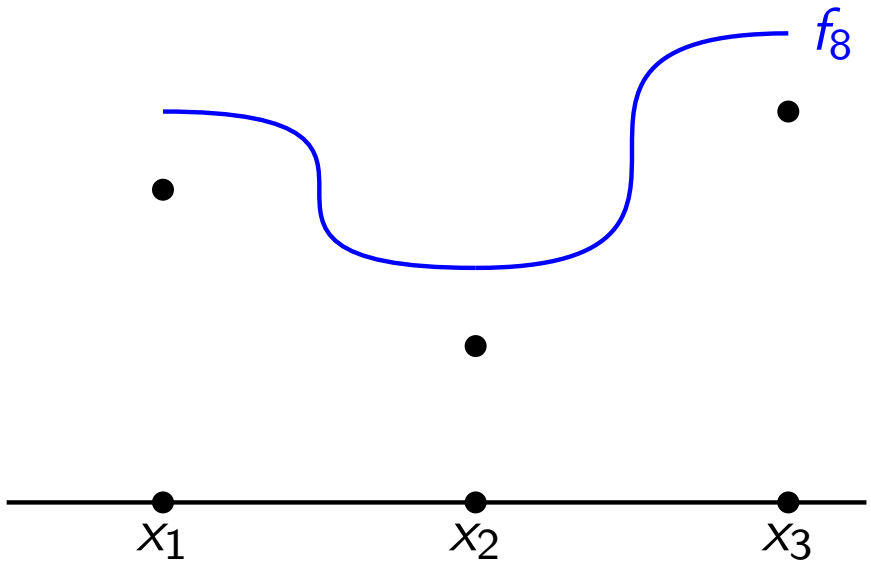
$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$



Regression: Fat Shattering Dimension

- $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [-1, +1]$
- $x_{1:n}$ is α-shattered by $\mathcal{F}$, if there exists thresholds $s_{1:n}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t(f(x_t) - s_t) \geq \alpha$$

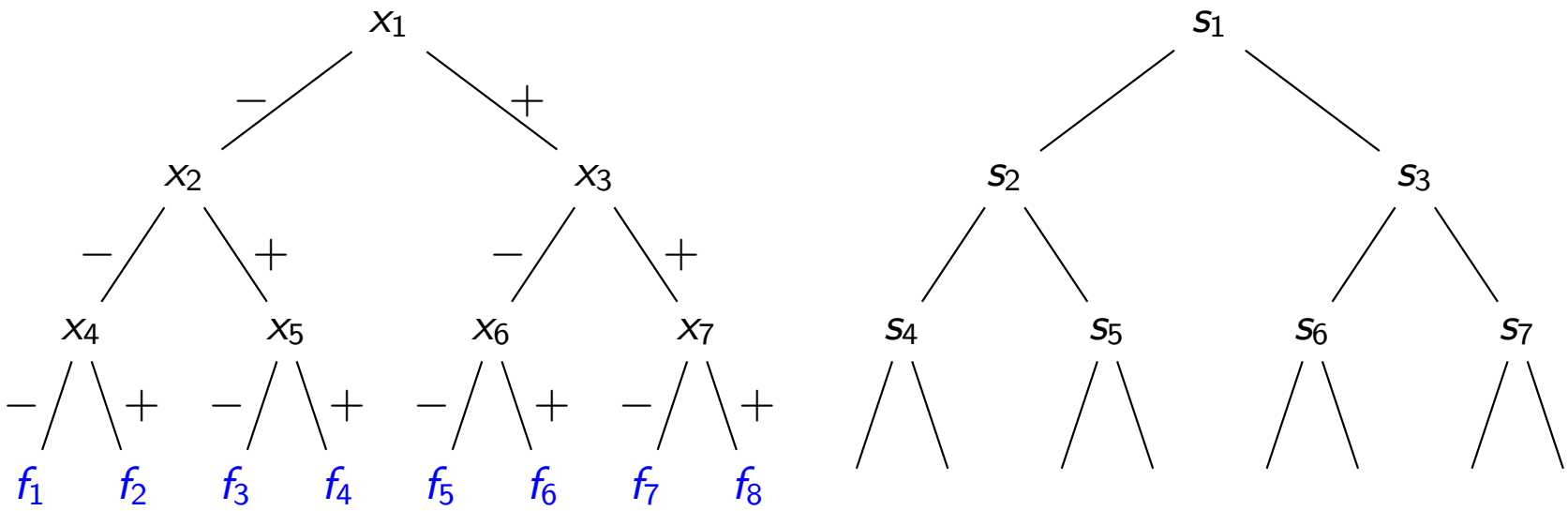


- The fat shattering dimension of $\mathcal{F}$ at scale α is the length of the longest sequence $x_{1:n}$ that is α-shattered by $\mathcal{F}$

Regression: Seq. Fat Shattering Dimension

- Tree $\mathbf{x}$ is α-shattered by $\mathcal{F}$, if there exists a threshold tree $\mathbf{s}$ such that for all $\epsilon_{1:n} \in \{\pm 1\}^n$

$$\exists f \in \mathcal{F},\ \forall t \in \{1, \ldots, n\},\ \epsilon_t \cdot \left(f(\mathbf{x}(\epsilon_{1:t-1}) - \mathbf{s}(\epsilon_{1:t-1}) \geq \alpha\right.$$



- The sequential fat shattering dimension of $\mathcal{F}$ at scale α is the depth of the deepest tree $\mathbf{x}$ that is α-shattered by $\mathcal{F}$

Learnability

- Regression with squared loss:
  $f : \mathcal{X} \to [-1, +1], (x, y) \in \mathcal{X} \times [-1, +1]$

$$\ell(f(x), y) = (y - f(x))^2$$

- Probabilistic setting

$$\mathbb{E}\left[L(\widehat{f}) - L(f^\star)\right] \to 0$$

- Game theoretic setting

$$\frac{V_n}{n} \to 0$$

Uniform Law of Large Numbers

- Fix class $\mathcal{F}$ of bounded real valued functions
- IID setting: If $X_1, X_2, \ldots$ are iid, do we have

$$\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{t=1}^{n} f(X_t) - \mathbb{E}\left[f(X)\right] \right) \to 0$$

  with convergence being uniform over all distributions?
- Martingale setting: If $X_1, X_2, \ldots$ is an arbitrary stochastic process, do we have

$$\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{t=1}^{n} (f(X_t) - \mathbb{E}\left[f(X_t) | X_{1:t-1}\right]) \right) \to 0$$

  with convergence being uniform over all distributions?

Four-way Equivalence: Classical Case

Several refs; see below

The following are equivalent (for a class $\mathcal{F}$ of bounded real valued functions)

- $\mathcal{F}$ is learnable in the iid setting under squared loss
- $\text{fat}_\alpha(\mathcal{F}) < \infty$ for all $\alpha > 0$
- $\mathcal{R}_n(\mathcal{F}) \to 0$
- Uniform law of large numbers holds for $\mathcal{F}$

Kearns, Schapire (1994); Bartlett, Long, Williamson (1996); Alon, Ben-David, Cesa-Bianchi, Haussler (1997); Mendelson (2002)

Fourfold Equivalence: Sequential Case

Rakhlin, Sridharan, Tewari (2010, 2014a, 2014b)

The following are equivalent (for a class $\mathcal{F}$ of bounded real valued functions)

- $\mathcal{F}$ is learnable in the online regression setting under squared loss
- $\mathrm{sfat}_\alpha(\mathcal{F}) < \infty$ for all $\alpha > 0$
- $\mathcal{R}_n^{\mathrm{seq}}(\mathcal{F}) \to 0$
- Uniform martingale law of large numbers holds for $\mathcal{F}$

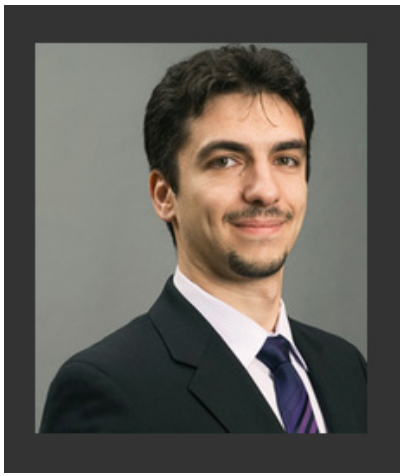
Summary

- Online learning framework uses game-theoretic, not probabilistic, foundations for prediction problems
- Complexity measures such as Rademacher complexity and fat-shattering dimension have natural sequential analogs
- Sequential complexity measures characterize function classes for which uniform martingale LLN holds

Gratitude and Thanks

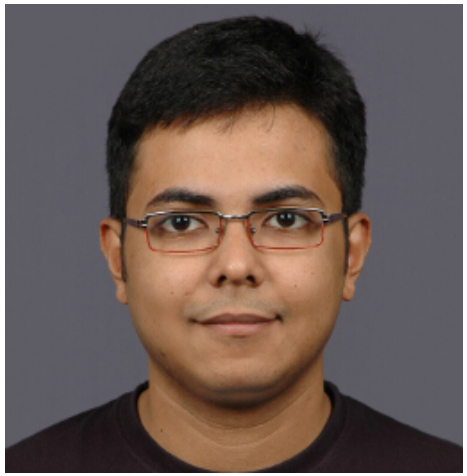
Thanks to the workshop organizers and CIMAT for an excellent workshop!

Thanks to my co-authors:



Alexander (Sasha) Rakhlin

Statistics, UPenn



Karthik Sridhran

Computer Science, Cornell

Thank You!

References (other than my own work)

- Hannan, J. (1957). Approximation to Bayes risk in repeated play. Contributions to the Theory of Games, 3, 97-139. URL
- Blackwell, D. (1954). Controlled random walks. In Proceedings of the International Congress of Mathematicians (Vol. 3, pp. 336-338). URL
- Banos, A. (1968). On pseudo-games. The Annals of Mathematical Statistics, 1932-1945. URL
- Cover, T. M. (1991). Universal portfolios. Mathematical finance, 1(1), 1-29. URL
- Foster, D. P., & Vohra, R. V. (1993). A randomization rule for selecting forecasts. Operations Research, 41(4), 704-709. URL
- Vovk, V. G. (1990). Aggregating strategies. In Proc. Third Workshop on Computational Learning Theory (pp. 371-383). Morgan Kaufmann. (unavailable online)

References (other than my own work)

- P. L. Bartlett, P. M. Long, and R. C. Williamson. (1996). Fat-shattering and the learnability of real-valued functions. J. Comput. Syst. Sci. 52, 3 (June 1996), 434-452. URL
- M. J. Kearns and R. E. Schapire. (1994). Efficient distribution-free learning of probabilistic concepts. J. Comput. Syst. Sci. 48, 3 (June 1994), 464-497. URL
- N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. J. ACM 44, 4 (July 1997), 615-631. URL
- S. Mendelson. (2002). Rademacher averages and phase transitions in Glivenko-Cantelli classes. IEEE Trans. Inf. Theory 48, 1, 251-263. URL

References to my own work

- K. Sridharan, A. Tewari. Convex games in Banach spaces, in *COLT* 2010. URL
- A. Rakhlin, K. Sridharan, A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability, in *NIPS* 2010. URL
- A. Rakhlin, K. Sridharan, A. Tewari. Online learning: Beyond regret, in *COLT* 2011. URL
- A. Rakhlin, K. Sridharan, A. Tewari. Online learning: Stochastic, constrained, and smoothed adversaries, in *NIPS* 2011. URL
- A. Rakhlin, K. Sridharan, A. Tewari. Sequential complexities and uniform martingale laws of large numbers, Probab. Theory Related Fields, 2014a. URL
- A. Rakhlin, K. Sridharan, A. Tewari. Online learning via sequential complexities, JMLR, 2014b. to appear. URL