Fast rates of convergence for learning problems

John Duchi

Outline

I Mean estimation and uniform laws

II Convexity

III Growth conditions and fast rates

Best rates from uniform laws

What do our uniform laws give us?

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ \widehat{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h; X_i) \right\}$$

and

$$\sup_{h \in \mathcal{H}} |\widehat{L}_n(h) - L(h)| \lesssim \frac{1}{\sqrt{n}} \text{ with high probability}$$

Best this can be?

$$L(\widehat{h}) - L(h^\star) = \widehat{L}_n(\widehat{h}) - \widehat{L}_n(h^\star) + L(\widehat{h}) - \widehat{L}_n(\widehat{h}) + \widehat{L}_n(h^\star) - \widehat{L}_n(h^\star)$$

The best rate from this approach

Central limit theorems: Consider the third error term involving $h^\star$:

$$\sqrt{n}\left(L(h^\star) - \widehat{L}_n(h^\star)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[L(h^\star) - \ell(h^\star; X_i)\right]$$
$$\stackrel{d}{\rightsquigarrow} \mathsf{N}\left(0, \mathrm{Var}(\ell(h^\star; X))\right)$$

Is this right?

Estimating a mean

Goal: We want to estimate $\theta^\star = \mathbb{E}[X]$, use loss

$$\ell(\theta; x) = \frac{1}{2}(\theta - x)^2$$

with risk

$$L(\theta) = \frac{1}{2}\mathbb{E}[(\theta - X)^2] = \frac{1}{2}\mathbb{E}[(\theta - \mathbb{E}[X] + \mathbb{E}[X] - X)^2]$$
$$= \frac{1}{2}(\theta - \mathbb{E}[X])^2 + \mathrm{Var}(X).$$

Gap in risks: Subtracting we have

$$L(\theta) - L(\theta^\star) = \frac{1}{2}(\theta - \theta^\star)^2 + \mathrm{Var}(X) - \mathrm{Var}(X) = \frac{1}{2}(\theta - \theta^\star)^2$$

Estimating a sub-Gaussian mean

Let X_i be independent σ^2-sub-Gaussian, so that

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i = \operatorname*{argmin}_\theta \widehat{L}_n(\theta)$$

and for $t \ge 0$ we have

$$\mathbb{P}(|\widehat{\theta}_n - \theta^\star| \ge t) \le 2\exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

Lemma

With probability at least $1-\delta$, we have

$$L(\widehat{\theta}_n) - L(\theta^\star) = \frac{1}{2}(\widehat{\theta}_n - \theta^\star)^2 \le C\frac{\sigma^2 \log\frac{1}{\delta}}{n}.$$

Uniform law for means?

$$\sup_{\theta \in \mathbb{R}} \left\{ \widehat{L}_n(\theta) - L(\theta) \right\} = +\infty$$

Convexity: heuristic and graphical explanation

Convexity: definitions

Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if

$$f(\lambda u + (1 - \lambda)v) \leq \lambda f(u) + (1 - \lambda) f(v)$$

for all $u, v \in \mathbb{R}^d$ and $\lambda \in [0, 1]$

Basic properties

A few properties of convex functions

- If $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable, then f is convex if and only if $f''(t) \geq 0$
- If $f : \mathbb{R}^m \to \mathbb{R}$ is convex and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, then $g(x) := f(Ax + b)$ is convex
- If f_1, f_2 are convex, then $f_1 + f_2$ is convex

Examples

Example

Logistic loss: $\phi(t) = \log(1 + e^{-t})$ and $\ell(\theta; x, y) = \log(1 + e^{-yx^T\theta}) = \phi(yx^T\theta)$

Example

Any norm $\|\cdot\|$.

Example

ℓ_1-regularized linear regression:

$$\frac{1}{2n} \|X\theta - y\|_2^2 + \lambda \|\theta\|_1 .$$

Convex functions have no local minima

Theorem

Let $\mathbb{B} = \{\theta : \|\theta\| \le 1\}$, $\mathbb{S} = \{\theta : \|\theta\| = 1\}$, *suppose* f *is convex and satisfies*

$$f(\theta) \ge f(\theta^\star) \quad \textit{for } \theta \in \theta^\star + \epsilon\mathbb{S}.$$

For $\theta \not\in \theta^\star + \epsilon\mathbb{B}$, *define*

$$\theta_\epsilon := \frac{\epsilon}{\|\theta - \theta^\star\|}\theta + \left(1 - \frac{\epsilon}{\|\theta - \theta^\star\|}\right)\theta^\star$$

Then

$$f(\theta) - f(\theta^\star) \ge \frac{\|\theta - \theta^\star\|}{\epsilon}\left[f(\theta_\epsilon) - f(\theta^\star)\right]$$

Proof of theorem

Note that $\theta_\epsilon \in \theta^\star + \epsilon\mathbb{S}$, so for $t = \frac{\epsilon}{\|\theta - \theta^\star\|} \le 1$,

$$\theta_\epsilon = \frac{\epsilon}{\|\theta - \theta^\star\|}\theta + \left(1 - \frac{\epsilon}{\|\theta - \theta^\star\|}\right)\theta^\star = t\theta + (1-t)\theta^\star,$$

and

$$f(\theta_\epsilon) \le tf(\theta) + (1-t)f(\theta^\star)$$

Convex loss functions

Suppose that we use a convex loss, i.e. $\ell(\theta; X)$ is convex in θ. Then

$$\widehat{L}_n(\theta) > \widehat{L}_n(\theta^\star) \text{ for all } \theta \in \theta^\star + \epsilon\mathbb{S}$$

implies that

$$\widehat{\theta}_n = \operatorname*{argmin}_\theta \widehat{L}_n(\theta) \text{ satisfies } \|\widehat{\theta}_n - \theta^\star\| \le \epsilon$$

A picture of how we achieve fast rates

Growth and smoothness

Let us fix some radius $r > 0$, and assume

$$\textbf{[Growth]} \qquad L(\theta) \geq L(\theta^\star) + \frac{\lambda}{2} \|\theta - \theta^\star\|^2 \ \text{ for } \|\theta - \theta^\star\| \leq r$$

and that

$$\textbf{[Smoothness]} \qquad \ell(\cdot; x) \ \text{ is } M\text{-Lipschitz on } \{\theta : \|\theta - \theta^\star\| \leq r\}$$

Example (linear regression with bounded x)

Fast rates under growth conditions

Theorem

Let the conditions on growth and smoothness hold, define

$$\Theta_\epsilon := \{\theta \in \Theta \mid \|\theta - \theta^\star\| \le \epsilon\}$$

and the localized Rademacher complexity

$$R_n(\Theta_\epsilon) := \mathbb{E}\left[\sup_{\theta \in \Theta_\epsilon} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left[\ell(\theta; X_i) - \ell(\theta^\star; X_i)\right] \right|\right]$$

Fix $t \ge 0$ *and choose any* $\epsilon \le r$ *such that*

$$\frac{\lambda \epsilon^2}{2} \ge 2 R_n(\Theta_\epsilon) + \sqrt{2} \frac{M}{\sqrt{n}} t \cdot \epsilon.$$

Then

$$\mathbb{P}\left(\|\widehat{\theta} - \theta\| \ge \epsilon\right) \le 2 e^{-nt^2}$$

Bounding the local Rademacher complexity

Under the conditions of the theorem, for $\Theta \subset \mathbb{R}^d$ we have

$$\|\widehat{\theta} - \theta\| \le C \frac{M}{\lambda\sqrt{n}} \left(\sqrt{d} + t\right) \quad \text{w.p. } \ge 1 - 2e^{-nt^2}.$$

Bounding the local Rademacher complexity

Under the conditions of the theorem, for $\Theta \subset \mathbb{R}^d$ we have

$$L(\widehat{\theta}) - L(\theta^\star) \le \frac{O(\text{stuff}) \log \frac{1}{\delta}}{n} \quad \text{w.p. } \ge 1 - \delta.$$

Multiclass classification

Suppose we have multiclass logistic loss for $\theta = [\theta_1 \ \cdots \ \theta_k]$, $\theta_l \in \mathbb{R}^d$, $y \in \{1, \ldots, k\}$, $\|x\|_2 \le M$

$$\ell(\theta; x, y) = \log \left(\sum_{l=1}^{k} \exp\left(x^T (\theta_l - \theta_y) \right) \right).$$

Then

$$R_n(\Theta_\epsilon) \lesssim \frac{M\sqrt{dk}}{\sqrt{n}} \epsilon$$

Proof of theorem

Part 1: Consider the event $\widehat{L}_n(\theta) \le \widehat{L}_n(\theta^\star)$ for some $\theta \in \Theta_\epsilon$, which implies

$$\left(\widehat{L}_n(\theta) - L(\theta)\right) - \left(\widehat{L}_n(\theta^\star) - L(\theta^\star)\right) \le -\frac{\lambda}{2}\epsilon^2$$

Proof of theorem

Part 2: Consider *localized excess risk* for $\theta \in \Theta_\epsilon$

$$\sum_{i=1}^{n} [(\ell(\theta; X_i) - \ell(\theta^\star; X_i)) - (L(\theta) - L(\theta^\star))]$$

and get (for all $t \geq 0$)

$$\mathbb{P}\left(\sup_{\theta \in \Theta_\epsilon} \left|\widehat{L}_n(\theta) - L(\theta) - (\widehat{L}_n(\theta^\star) - L(\theta^\star))\right| \geq 2R_n(\Theta_\epsilon) + t\right) \leq 2\exp\left(-\frac{nt^2}{2M^2\epsilon^2}\right)$$

Proof of theorem

Part 3: Implications: $\|\widehat{\theta} - \theta^\star\| \ge \epsilon$

$$\Rightarrow \widehat{L}_n(\theta) - \widehat{L}_n(\theta^\star) \le 0 \ \text{ for some } \theta \in \Theta_\epsilon$$

$$\Rightarrow \sup_{\theta \in \Theta_\epsilon} \left| \widehat{L}_n(\theta) - L(\theta) - (\widehat{L}_n(\theta^\star) - L(\theta^\star)) \right| \ge \frac{\lambda \epsilon^2}{2}$$

$$\Rightarrow \sup_{\theta \in \Theta_\epsilon} \left| \widehat{L}_n(\theta) - L(\theta) - (\widehat{L}_n(\theta^\star) - L(\theta^\star)) \right| \ge 2R_n(\Theta_\epsilon) + \sqrt{2}\frac{M}{\sqrt{n}}\, t \cdot \epsilon$$

Reading and bibliography

1. P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities.
   Annals of Statistics, 33(4):1497–1537, 2005
2. A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*.
   Springer, New York, 1996 (Ch. 3.2–3.4)
3. S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances.
   ESAIM: Probability and Statistics, 9:323–375, 2005 (§5.3)
4. A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*.
   SIAM and Mathematical Programming Society, 2009 (Ch. 5.3)