On Fast Rates in Empirical Risk Minimization Beyond Least-Squares

Dmitrii M. Ostrovskii
http://ostrodmit.github.io

USC Epstein Seminar
October 9, 2019

Problem setup

Statistical learning problem

Given some **loss** $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, minimize the **population risk**:

$$\theta_* \in \operatorname*{Argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} L(\theta) := \mathbf{E}[\ell(X^\top \theta, Y)],$$

where expectation $\mathbf{E}[\cdot]$ is w.r.t. the unknown distribution $\mathcal{P}$ of $(X, Y) \in \mathbb{R}^d \times \mathcal{Y}$. Since $\mathcal{P}$ is unknown, θ_* can't be found; instead, it is estimated from **i.i.d. sample**:

$$(X_i, Y_i) \sim \mathcal{P}, \quad i \in \{1, ..., n\}.$$

- Random-design **classification**, $\mathcal{Y} = \{0, 1\}$, and **regression**, $\mathcal{Y} = \mathbb{R}$.
- Structure prediction problems with complex $\mathcal{Y}$ (graphs, word sequences, etc.)
- Performance of a candidate estimate $\widehat{\theta}$ measured by the **excess risk:**

$$L(\widehat{\theta}) - L(\theta_*),$$

that is, how well $\widehat{\theta}$ performs against the best model θ_* in terms of $\mathcal{P}$.

Goal

- **Empirical risk minimization:** replace $L(\theta)$ with **empirical risk**:

$$\widehat{\theta}_n \in \operatorname*{Argmin}_{\theta \in \mathbb{R}^d} \left\{ L_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(X_i^\top \theta, Y_i) \right\}.$$

  Also called M-**estimation** in statistics.

- Special case: conditional **quasi maximum likelihood estimator** (qMLE):

$$\ell(\eta, y) = -\log p_\eta(y)$$

  for some parametric family $\{p_\eta(y), \eta \in \mathbb{R}\}$, possibly not containing the true distribution $\mathcal{P}$ (i.e. misspecified model).

- Rich classical **asymptotic theory*** when dimension d is fixed, and $n \to \infty$.

Goal: extend the asymptotic theory to **finite-sample** setups.

*[Borovkov, 1998; van der Vaart, 1998; Lehmann and Casella, 2006].

Asymptotic theory

- **Local regularity assumptions**: $L(\theta)$ sufficiently smooth around θ_*, and

$$\mathbf{H}_* := \nabla^2 L(\theta_*) \succ 0.$$

- Fisher information matrix $\mathbf{G}_* := \mathbf{E}[\nabla_\theta \ell(X^\top \theta_*, Y)\, \nabla_\theta\, \ell(X^\top \theta_*, Y)^\top]$, and let

$$\mathbf{M}_* := \mathbf{H}_*^{-1/2} \mathbf{G}_* \mathbf{H}_*^{-1/2}.$$

$d_{\text{eff}} := \mathrm{Tr}(\mathbf{M}_*)$ is the **effective dimension**. In well-specified models,

$$\mathbf{G}_* = \mathbf{H}_* \quad \Longrightarrow \quad \mathbf{M}_* = \mathbf{I}_d \quad \Longrightarrow \quad d_{\text{eff}} = d.$$

Theorem. Assume that Θ is open, and ℓ is sufficiently regular (in particular, $\ell'''(\cdot,\cdot)$ is bounded in some neighborhood of θ_*). When $n \to \infty$,

$$\sqrt{n}\mathbf{H}_*^{1/2}(\widehat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, \mathbf{M}_*),$$

$$n\|\mathbf{H}_*^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \rightsquigarrow \mathcal{N}(0, \mathbf{M}_*)^2, \quad 2n(L(\widehat{\theta}_n) - L(\theta_*)) \rightsquigarrow \mathcal{N}(0, \mathbf{M}_*)^2.$$

As a result, with probability $\geq 1 - \delta$,

$$\left\{ L(\widehat{\theta}_n) - L(\theta_*),\, \|\mathbf{H}_*(\theta_n - \theta_*)\|^2 \right\} = O\left(\frac{d_{\text{eff}} \log(1/\delta)}{n} \right).$$

Asymptotic theory (cont.)

Analysis based on the observation that $\widehat{\theta}_n \to \theta_*$ (assume $d = 1$ for simplicity):

1. By Taylor's thm, for some $\bar{\theta}_n \in [\theta_*, \widehat{\theta}_n]$,

$$0 = \sqrt{n} L_n'(\widehat{\theta}_n) = \sqrt{n} L_n'(\theta_*) + \sqrt{n}(\widehat{\theta}_n - \theta_*) L_n''(\theta_*) + \frac{L_n'''(\bar{\theta}_n)}{2\sqrt{n}} [\sqrt{n}(\widehat{\theta}_n - \theta_*)]^2.$$

   Regrouping the terms,

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) = \frac{-\sqrt{n} L_n'(\theta_*)}{L_n''(\theta_*) + \frac{1}{2} L_n'''(\bar{\theta}_n)(\widehat{\theta}_n - \theta_*)}.$$

2. When $n \to \infty$, using the regularity of $L(n)$,

$$L_n''(\theta_*) \to L''(\theta_*).$$

3. We have $\widehat{\theta}_n \to \theta_*$ due to Cramér (1946). Since $L_n'''(\bar{\theta}_n)$ is bounded,

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \approx \frac{-\sqrt{n} L_n'(\theta_*)}{L''(\theta_*)},$$

   Note that $L''(\theta_*) = \mathbf{H}_*$, and $\sqrt{n} L_n'(\theta_*)$ converges to $\mathcal{N}(0, \mathbf{G}_*)$ by CLT. ∎

Simple case: least squares

Model $Y = \mathcal{N}(X^\top\theta, \sigma^2)$ leads to $\ell(X^\top\theta, Y) = \frac{1}{2\sigma^2}(Y - X^\top\theta)^2$, **quadratic risks:**

$$L(\theta) - L(\theta_*) = \frac{1}{2}\|\mathbf{H}^{1/2}(\theta - \theta_*)\|^2,$$
$$L_n(\theta) - L_n(\theta_*) = \frac{1}{2}\|\mathbf{H}_{n}^{1/2}(\theta - \theta_*)\|^2 + \langle\nabla L_n(\theta_*), \theta - \theta_*\rangle$$

- In particular, at any θ we have $\nabla^2 L(\theta) \equiv \mathbf{H}$ and $\nabla^2 L_n(\theta) \equiv \mathbf{H}_n$ with
$$\mathbf{H} = \mathbf{E}[XX^\top], \quad \mathbf{H}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top.$$
$\mathbf{H}_n$ converges to $\mathbf{H}$ when $n \to \infty$. Moreover, there is a finite-sample result:

Simple case: least squares

Model $Y = \mathcal{N}(X^\top\theta, \sigma^2)$ leads to $\ell(X^\top\theta, Y) = \frac{1}{2\sigma^2}(Y - X^\top\theta)^2$, **quadratic risks:**

$$L(\theta) - L(\theta_*) = \frac{1}{2}\|\mathbf{H}^{1/2}(\theta - \theta_*)\|^2,$$
$$L_n(\theta) - L_n(\theta_*) = \frac{1}{2}\|\mathbf{H}_n^{1/2}(\theta - \theta_*)\|^2 + \langle \nabla L_n(\theta_*), \theta - \theta_* \rangle$$

- In particular, at any θ we have $\nabla^2 L(\theta) \equiv \mathbf{H}$ and $\nabla^2 L_n(\theta) \equiv \mathbf{H}_n$ with

$$\mathbf{H} = \mathbf{E}[XX^\top], \quad \mathbf{H}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top.$$

$\mathbf{H}_n$ converges to $\mathbf{H}$ when $n \to \infty$. Moreover, there is a finite-sample result:

Theorem [Vershynin, 2010]

Assume $X - \mathbf{E}[X]$ has subgaussian moment growth in all directions: for $\mu = \mathbf{E}[X]$,

$$\mathbf{E}^{1/p}[\langle X - \mu, u\rangle^p] \lesssim \sqrt{p}\mathbf{E}^{1/2}[\langle X - \mu, u\rangle^2], \quad \forall u \in \mathbb{R}^d,$$

Whenever $n \gtrsim d + \log(1/\delta)$, w.p. $\geq 1 - \delta$ it holds:

$$(1 - \varepsilon)\mathbf{H} \preccurlyeq \mathbf{H}_n \preccurlyeq (1 + \varepsilon)\mathbf{H},$$

$$\text{where } \varepsilon \lesssim \sqrt{\frac{d + \log(1/\delta)}{n}}.$$

Simple case: least squares (cont.)

Theorem (folklore, see [Hsu et al., 2012])

Assume $X - \mu$ is subgaussian, and the noise $\xi = Y - X^\top \theta_*$ is subgaussian. Let

$$n \gtrsim d + \log(1/\delta).$$

Then w.p. $\geq 1 - \delta$,

$$L(\widehat{\theta}_n) - L(\theta_*) = \|\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \frac{d_{\mathrm{eff}} \log^2(1/\delta)}{n}.$$

Simple case: least squares (cont.)

Theorem (folklore, see [Hsu et al., 2012])

Assume $X - \mu$ is subgaussian, and the noise $\xi = Y - X^\top \theta_*$ is subgaussian. Let

$$n \gtrsim d + \log(1/\delta).$$

Then w.p. $\geq 1 - \delta$,

$$L(\widehat{\theta}_n) - L(\theta_*) = \|\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \frac{d_{\text{eff}} \log^2(1/\delta)}{n}.$$

Proof sketch:

1. Since $\nabla L_n(\widehat{\theta}_n) = 0$, we have $\|\mathbf{H}_n^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 = \|\mathbf{H}_n^{-1/2} \nabla L_n(\theta_*)\|^2$.
2. By Vershynin's matrix concentration result, $\frac{1}{2}\mathbf{H} \preccurlyeq \mathbf{H}_n \preccurlyeq 2\mathbf{H}$, whence

$$\begin{aligned} L(\widehat{\theta}_n) - L(\theta_*) = \tfrac{1}{2}\|\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \|\mathbf{H}_n^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 &= \|\mathbf{H}_n^{-1/2} \nabla L_n(\theta_*)\|^2 \\ &\lesssim \|\mathbf{H}^{-1/2} \nabla L_n(\theta_*)\|^2. \end{aligned}$$

3. $\mathbf{H}^{-1/2} \nabla L_n(\theta_*) = \frac{1}{n}\sum_{i=1}^n \mathbf{H}^{-1/2} \xi_i X_i$ is the average of i.i.d. zero-mean random vectors. ■

Towards the general case

- Analysis above were simplified by the “automatic” **localization** of $\widehat{\theta}_n$ near θ_*.
  - In the asymptotic setup, we used LLN and a local bound on $L_n'''(\theta)$.
  - For least squares, localization is “automatic” because $L_n'''(\cdot) \equiv 0$. The argument only required Taylor expansion at θ_* and convergence of $\mathbf{H}_n$.
- Generally, risk is not quadratic, and Hessians are not constant.

$$\nabla^2 L(\theta) = \mathbf{H}(\theta), \quad \nabla^2 L_n(\theta) = \mathbf{H}_n(\theta).$$

  To extend the argument, we must localize $\widehat{\theta}_n$ to the right neighborhood of θ_*.
- Such localization is naturally done via **self-concordance**.
  - Introduced in [Nesterov and Nemirovski, 1994] in the context of interior-point methods.
  - Brought to statistics in [Bach, 2010] to study logistic regression.

Self-concordant losses

We always assume that $\ell(\eta, y)$ is convex in η (can be relaxed to quasi-convexity).

Definition. $\ell(\eta, y)$ is **self-concordant (SC)** if for any $(\eta, y) \in \mathbb{R} \times \mathcal{Y}$ it holds

$$|\ell_\eta'''(\eta, y)| \le [\ell_\eta''(\eta, y)]^{3/2}.$$

Self-concordant losses

We always assume that $\ell(\eta, y)$ is convex in η (can be relaxed to quasi-convexity).

Definition. $\ell(\eta, y)$ is **self-concordant (SC)** if for any $(\eta, y) \in \mathbb{R} \times \mathcal{Y}$ it holds

$$|\ell'''_\eta(\eta, y)| \le [\ell''_\eta(\eta, y)]^{3/2}.$$

- This definition is homogeneous in η. The next one is not:

Definition. $\ell(\eta, y)$ is **pseudo self-concordant (PSC)** if instead it holds

$$|\ell'''_\eta(\eta, y)| \le \ell''_\eta(\eta, y).$$

- **PSC** losses are somewhat more common than **SC** ones.
- However, obtaining optimal rate for **PSC** losses requires larger sample size.

Sub-optimal result

Recall

$$d_{\text{eff}} = \text{Tr}[\mathbf{H}(\theta_*)^{-1/2}\,\mathbf{G}(\theta_*)\,\mathbf{H}(\theta_*)^{-1/2}],$$

and we have the Hessian map $\theta \mapsto \mathbf{H}(\theta)$ given by

$$\mathbf{H}(\theta) := \mathbf{E}[\ell''(X^\top\theta, Y)XX^\top].$$

We see that $\mathbf{H}(\theta) = \mathbf{E}[\widetilde{X}(\theta)\widetilde{X}(\theta)^\top]$ for *curved design* $\widetilde{X}(\theta) := [\ell''(X^\top\theta, Y)]^{1/2}X$.

Theorem 1 [Ostrovskii and Bach, 2018])

Assume that $\ell(\eta, y)$ is **SC**, and that $\widetilde{X}(\theta_*)$ and $\nabla_\theta \ell(X^\top\theta_*, Y)$ are subgaussian. Whenever

$$n \gtrsim \max\left\{d + \log(1/\delta), d_{\text{eff}}\, d \log(1/\delta)\right\},$$

w.p. $\geq 1 - \delta$ it holds

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\mathbf{H}(\theta_*)^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \frac{d_{\text{eff}} \log(1/\delta)}{n}.$$

- ☺ Distribution conditions are **local**, i.e., concern only θ_*.
- ☹ Large sample size $n = O(d_{\text{eff}}\, d)$.

Key observation

Given $\mathbf{H}(\theta) = \nabla^2 L(\theta)$, consider **Dikin ellipsoids** of $L(\theta)$ at θ_0:

$$\Theta(\theta_0, r) := \{\theta : \|\mathbf{H}(\theta_0)^{1/2}(\theta - \theta_0)\|^2 \le r^2\}.$$

[Nesterov and Nemirovski, 1994]: $c\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}(\theta) \preccurlyeq C\mathbf{H}(\theta_*)$ for any $\theta \in \Theta(\theta_*, 1)$.

Localization lemma. Assume the following two events hold:

1. $c\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq C\mathbf{H}(\theta_*)$ uniformly over $\theta \in \Theta(\theta_*, r)$ with some $r \lesssim 1$.
2. $\|\mathbf{H}(\theta_*)^{-1/2}\nabla L_n(\theta_*)\|^2 \lesssim r^2$.

Then, $\widehat{\theta}_n$ belongs to $\Theta(\theta_*, r)$, and the excess risk bound of Theorem 1 holds.

Key observation

Given $\mathbf{H}(\theta) = \nabla^2 L(\theta)$, consider **Dikin ellipsoids** of $L(\theta)$ at θ_0:

$$\Theta(\theta_0, r) := \{\theta : \|\mathbf{H}(\theta_0)^{1/2}(\theta - \theta_0)\|^2 \le r^2\}.$$

[Nesterov and Nemirovski, 1994]: $c\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}(\theta) \preccurlyeq C\mathbf{H}(\theta_*)$ for any $\theta \in \Theta(\theta_*, 1)$.

Localization lemma. Assume the following two events hold:

1. $c\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq C\mathbf{H}(\theta_*)$ uniformly over $\theta \in \Theta(\theta_*, r)$ with some $r \lesssim 1$.
2. $\|\mathbf{H}(\theta_*)^{-1/2}\nabla L_n(\theta_*)\|^2 \lesssim r^2$.

Then, $\widehat{\theta}_n$ belongs to $\Theta(\theta_*, r)$, and the excess risk bound of Theorem 1 holds.

- Indeed, by definition of $\widehat{\theta}_n$, $L_n(\widehat{\theta}_n) \le L_n(\theta_*)$. Assume $\widehat{\theta}_n \notin \Theta(\theta_*, r)$.
- Pick $\bar{\theta}_n \in [\theta_*, \widehat{\theta}_n]$ on the **boundary** of $\Theta(\theta_*, r)$. By cvxty, $L_n(\bar{\theta}_n) \le L_n(\theta_*)$,
$$0 \ge L_n(\bar{\theta}_n) - L_n(\theta_*) = \langle \nabla L_n(\theta_*), \bar{\theta}_n - \theta_* \rangle + \|\mathbf{H}_n(\theta')^{1/2}(\bar{\theta}_n - \theta_*)\|^2$$
for some $\theta' \in [\theta_*, \bar{\theta}_n]$.
- Using **1.**, we have $\|\mathbf{H}_n(\theta')^{1/2}(\bar{\theta}_n - \theta_*)\|^2 \gtrsim \|\mathbf{H}(\theta_*)^{1/2}(\bar{\theta}_n - \theta_*)\|^2 = r^2$.
- Hence, $\langle \nabla L_n(\theta_*), \bar{\theta}_n - \theta_* \rangle \gtrsim r^2$. By Cauchy-Schwarz, this contradicts **2.** ∎

Localization: recap

- Once we guaranteed localization $\widehat{\theta}_n \in \Theta(\theta_*, r)$ with $r \lesssim 1$, we can repeat the analysis for least squares, since $L_n(\cdot)$ is quadratic on $\Theta(\theta_*, r)$, and

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\mathbf{H}(\theta_*)^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \frac{d_{\text{eff}} \log(1/\delta)}{n}.$$

Localization: recap

- Once we guaranteed localization $\widehat{\theta}_n \in \Theta(\theta_*, r)$ with $r \lesssim 1$, we can repeat the analysis for least squares, since $L_n(\cdot)$ is quadratic on $\Theta(\theta_*, r)$, and

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\mathbf{H}(\theta_*)^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \frac{d_{\text{eff}} \log(1/\delta)}{n}.$$

- I.e., we need n to be large enough to guarantee **1** and **2**. In particular, for **2**,

$$\|\mathbf{H}_n^{-1/2}(\theta_*)\nabla L_n(\theta_*)\|^2 \lesssim r^2,$$

which leads to the second threshold for n:

$$n \gtrsim \frac{1}{r^2} d_{\text{eff}} \log(1/\delta).$$

Localization: recap

- Once we guaranteed localization $\widehat{\theta}_n \in \Theta(\theta_*, r)$ with $r \lesssim 1$, we can repeat the analysis for least squares, since $L_n(\cdot)$ is quadratic on $\Theta(\theta_*, r)$, and

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\mathbf{H}(\theta_*)^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \frac{d_{\mathrm{eff}} \log(1/\delta)}{n}.$$

- I.e., we need n to be large enough to guarantee **1** and **2**. In particular, for **2**,

$$\|\mathbf{H}_n^{-1/2}(\theta_*)\nabla L_n(\theta_*)\|^2 \lesssim r^2,$$

which leads to the second threshold for n:

$$n \gtrsim \frac{1}{r^2} d_{\mathrm{eff}} \log(1/\delta).$$

- **Now the question is:**

For which r can we ensure $c\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq C\mathbf{H}(\theta_*)$ uniformly on $\Theta(\theta_*, r)$?

  - We can afford $r = O(1/\sqrt{d})$ using self-concordance.
  - We can push this to $r = O(1)$ if we try hard enough!

Self-concordance at play:

For which radius r can we guarantee $c\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq C\mathbf{H}(\theta_*)$ on $\Theta(\theta_*, r)$?

1. Recall that $\mathbf{H}(\theta)$ and $\mathbf{H}_n(\theta)$ are the population and empirical 2nd-moment matrices of $\widetilde{X}(\theta) = \sqrt{\ell''_\eta(X^\top\theta, Y)}X$. If it is subgaussian, Vershynin gives
$$c\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta_*) \preccurlyeq C\mathbf{H}(\theta_*) \quad \text{w}.h.p.$$
whenever $n \gtrsim K^4(d + \log(1/\delta))$.
2. Integrating $|\ell'''_\eta(\eta, y)| \le [\ell''(\eta, y)]^{3/2}$ from $\eta_* = X^\top\theta_*$ to $\eta = X^\top\theta$,
$$\frac{1}{(1 + [\ell''(\eta_*, Y)]^{\frac{1}{2}}|\eta - \eta_*|)^2} \le \frac{\ell''(\eta, Y)}{\ell''(\eta_*, Y)} \le \frac{1}{(1 - [\ell''(\eta_*, Y)]^{\frac{1}{2}}|\eta - \eta_*|)^2},$$
$$\frac{1}{(1 + |\langle\widetilde{X}(\theta_*), \theta - \theta_*\rangle|)^2} \le \frac{\ell''(X^\top\theta, Y)}{\ell''(X^\top\theta_*, Y)} \le \frac{1}{(1 - |\langle\widetilde{X}(\theta_*), \theta - \theta_*\rangle|)^2}.$$
3. The ratio is bounded when $|\langle\widetilde{X}(\theta_*), \theta - \theta_*\rangle| \lesssim 1$, i.e., by Cauchy-Schwarz,
$$\underbrace{\|\mathbf{H}(\theta_*)^{-1/2}\widetilde{X}(\theta_*)\|}_{\approx\sqrt{d}} \cdot \underbrace{\|\mathbf{H}(\theta_*)^{1/2}(\theta - \theta_*)\|}_{r} \lesssim 1 \;\Rightarrow\; \boxed{r \lesssim \frac{1}{\sqrt{d}}}. \;\blacksquare$$

Improved result

Theorem 2 [Ostrovskii and Bach, 2018]

Assume $\ell(\eta, y)$ is **SC**, $\nabla_\theta \ell(X^\top \theta_*, Y)$ is subgaussian, and $\widetilde{X}(\theta)$ is K-subgaussian at any $\theta \in \Theta(\theta_*, 1)$. Whenever

$$n \gtrsim \max\left\{d \log(ed/\delta), d_{\text{eff}} \log(1/\delta)\right\},$$

w.p. $\geq 1 - \delta$ it holds

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\mathbf{H}(\theta_*)^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \frac{d_{\text{eff}} \log(1/\delta)}{n}.$$

- Sample size $n \gtrsim d_{\text{eff}}\, d$ in Theorem 1 is due to small radius $r = O(1/\sqrt{d})$ (rather than $r = O(1)$) in which sample Hessians are uniformly approximated.
- We need to ensure that $\mathbf{H}_n(\theta) \approx \mathbf{H}(\theta_*)$ w.h.p. uniformly over $\theta \in \Theta(\theta_*, 1)$.
- This could be done by showing first that $L(\theta)$ and $L_n(\theta)$ are **SC** on $\Theta(\theta_*, 1)$.

Self-concordance of population risk

- Given $\theta_0 = \theta_*$ and any $\theta_1 \in \Theta(\theta_*, r)$, consider

$$\phi(t) = L(\theta_t), \quad \text{where} \quad \theta_t = (1-t)\theta_0 + t\theta_1.$$

It suffices to ensure

$$|\phi'''(t)| \lesssim [\varphi''(t)]^{3/2}.$$

- By simple algebra, $\phi^{(p)}(t) = \langle \ell_\eta^{(p)}(X^\top \theta_t, Y) \cdot X, \Delta \rangle$ with $\Delta = \theta_1 - \theta_0$. Then,

$$\begin{aligned} |\phi'''(t)| &\le \mathbf{E}[|\ell_\eta'''(X^\top \theta_t, Y)| \cdot |\langle X, \Delta \rangle|^3] \\ &\le \mathbf{E}[[\ell_\eta''(Y, X^\top \theta_t)]^{3/2} \cdot |\langle X, \Delta \rangle|^3] && [\text{by } \mathbf{SC}] \\ &= \mathbf{E}[|\langle \widetilde{X}(\theta_t), \Delta \rangle|^3], \end{aligned}$$

while $[\phi''(t)]^{3/2} = \mathbf{E}^{3/2}[|\langle \widetilde{X}(\theta_t), \Delta \rangle|^2]$. But $u \mapsto u^{3/2}$ is convex, not concave!

- The bound follows by noting that $\widetilde{X}(\theta_t)$ is subgaussian when $\theta_t \in \Theta(\theta_*, 1)$, and comparing its 2nd and 3rd moments along direction Δ.

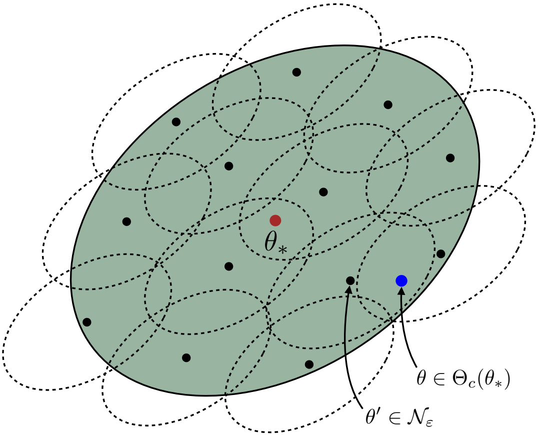
Self-concordance of population risk

- Given $\theta_0 = \theta_*$ and any $\theta_1 \in \Theta(\theta_*, r)$, consider
$$\phi(t) = L(\theta_t), \quad \text{where} \quad \theta_t = (1-t)\theta_0 + t\theta_1.$$
It suffices to ensure
$$|\phi'''(t)| \lesssim [\varphi''(t)]^{3/2}.$$
- By simple algebra, $\phi^{(p)}(t) = \langle \ell_\eta^{(p)}(X^\top \theta_t, Y) \cdot X, \Delta \rangle$ with $\Delta = \theta_1 - \theta_0$. Then,
$$\begin{aligned} |\phi'''(t)| &\le \mathbf{E}[|\ell_\eta'''(X^\top \theta_t, Y)| \cdot |\langle X, \Delta\rangle|^3] \\ &\le \mathbf{E}[[\ell_\eta''(Y, X^\top \theta_t)]^{3/2} \cdot |\langle X, \Delta\rangle|^3] && [\text{by } \mathbf{SC}] \\ &= \mathbf{E}[|\langle \widetilde{X}(\theta_t), \Delta\rangle|^3], \end{aligned}$$
while $[\phi''(t)]^{3/2} = \mathbf{E}^{3/2}[|\langle \widetilde{X}(\theta_t), \Delta\rangle|^2]$. But $u \mapsto u^{3/2}$ is convex, not concave!
- The bound follows by noting that $\widetilde{X}(\theta_t)$ is subgaussian when $\theta_t \in \Theta(\theta_*, 1)$, and comparing its 2nd and 3rd moments along direction Δ.

Self-concordance of empirical risk?

Unfortunately, this argument fails for $L_n(\cdot)$, where we compare *empirical moments.*

Correct approach: covering Dikin ellipsoid



1. We have already proved that $\mathbf{H}(\theta) \approx \mathbf{H}(\theta_*)$ on $\Theta(\theta_*, 1)$.
2. On the other hand, we can use the earlier integration argument to approximate $\mathbf{H}_n(\theta)$ by $\mathbf{H}_n(\theta')$ in a small ellipsoid $\Theta(\theta', 1/\sqrt{d})$.
3. Now cover $\Theta(\theta_*, 1)$ by $\Theta(\theta_0, 1/\sqrt{d})$ with θ_0 in the epsilon-net $\mathcal{N}_\varepsilon$. Control uniform deviations of $\mathbf{H}_n(\theta')$ from $\mathbf{H}(\theta')$, $\theta' \in \mathcal{N}_\varepsilon$. Costs extra $\log(d)$. ■

Pseudo self-concordant losses

- Because of the "wrong" power of ℓ'' in **PSC**, we need an extra condition:

$$\mathbf{\Sigma} := \mathbf{E}[XX^\top] \preceq \rho \mathbf{H}(\theta_*).$$

  for some $\rho > 0$. Standard assumption in logistic regression [Bach, 2010].

- The radius r of the Dikin ellipsoid in which we can control $\mathbf{H}_n(\theta)$ shrinks by $1/\sqrt{\rho}$, hence the critical sample size increases by ρ.
- While worst-case bounds on ρ can be exponentially bad [Hazan et al., 2014], this is not the case when the distribution of X is reasonable. E.g., we show

$$\rho \lesssim \|\theta_*\|_{\mathbf{\Sigma}}^{3/2}$$

  in logistic regression with $X \sim \mathcal{N}(0, \mathbf{\Sigma})$ and arbitrary $\mathbf{\Sigma}$.

Example 1: Generalized linear models

Conditional negative log-likelihood of y given $\eta = x^\top \theta$ in the form

$$\ell(\eta, y) = -y\eta + a(\eta) - b(y),$$

where $a(\eta)$ is called the **cumulant**, and is given by

$$a(\eta) = \log \int_{\mathcal{Y}} e^{y\eta + b(y)} \mathrm{d}y.$$

This defines the density $p_\eta(y) \propto e^{y\eta+b(y)}$ such that $a(\eta) = \mathbf{E}_{p_\eta}[y]$.
SC/**PSC** relate 2nd and 3rd central moments w.r.t. $p_\eta(\cdot)$.

PSC: **Logistic regression** since ($\mathcal{Y} = \{0, 1\}$), and

$$|a'''(\eta)| = |\mathbf{E}_{p_\eta}(y - \mathbf{E}_{p_\eta}[y])^3| \le \mathbf{E}_{p_\eta}[(y - \mathbf{E}_{p_\eta}[y])^2] = a''(\eta).$$

PSC: **Poisson regression:** $Y \sim \text{Poisson}(e^\eta)$, then $a(\eta) = \exp(\eta)$.
SC: **Exponential-response model:** $Y \sim \text{Exp}(\eta)$, $\eta > 0$, $a(\eta) = -\log(\eta)$.

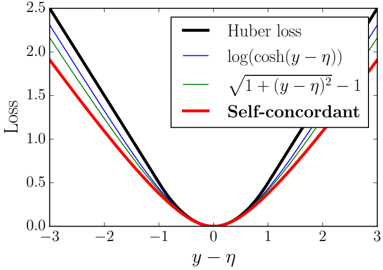
Example 2: Robust estimation

Loss $\ell(y,\eta) = \varphi(y-\eta)$ with $\varphi(t)$ convex, even, 1-Lipschitz, and $\varphi''(0) = 1$.

- **Huber loss**

$$\varphi(t) = \begin{cases} t^2/2, & |t| \le 1, \\ \tau t - 1/2, & |t| > 1. \end{cases}$$

$\varphi''(t)$ discontinuous at ± 1.



PSC: **Pseudo-Huber losses:** $\varphi(t) = \log\cosh(t)$, $\varphi(t) = \sqrt{1+t^2} - 1$.

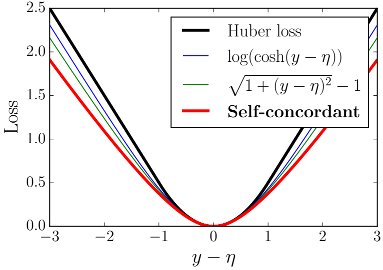
Example 2: Robust estimation

Loss $\ell(y,\eta) = \varphi(y-\eta)$ with $\varphi(t)$ convex, even, 1-Lipschitz, and $\varphi''(0) = 1$.

- **Huber loss**

$$\varphi(t) = \begin{cases} t^2/2, & |t| \le 1, \\ \tau t - 1/2, & |t| > 1. \end{cases}$$

$\varphi''(t)$ discontinuous at ± 1.



PSC: **Pseudo-Huber losses:** $\varphi(t) = \log\cosh(t)$, $\varphi(t) = \sqrt{1+t^2} - 1$.

SC: **Fenchel dual of the log-barrier** $\phi(u) = -\log(1-u^2)/2$ on $[-1,1]$:

$$\varphi(t) = \frac{1}{2}\left[\sqrt{1+4t^2} - 1 + \log\left(\frac{\sqrt{1+4t^2}-1}{2t^2}\right)\right].$$

Conclusion

We use self-concordance – a concept from optimization – to obtain asymptotically near-optimal rates in finite-sample statistical regime $n = \widetilde{O}(d)$.

Behind the scenes:

- high-dimensional setup and ℓ_1-regularization.
- non-parametric setup, ℓ_2-regularization. Interesting interplay of SC with the source and capacity conditions, see [Marteau-Ferey et al., 2019b].
- Quasi-Newton algorithms [Marteau-Ferey et al., 2019a].

Perspectives:

- Extension to heavy-tailed distributions.
- Extension to (generalized) Bayesian estimators and Gibbs-ERM.
- Other use cases: covariance matrix estimation ($\log\det$ loss), EM algorithm.

Thank you!

Extension to heavy-tailed distributions

- Our results crucially rely on the existence of an estimator $\widehat{\mathbf{\Sigma}}$ of covariance matrix $\mathbf{\Sigma}$ such that w.p. $\geq 1 - \delta$ it holds

$$(1-\varepsilon)\mathbf{\Sigma} \preccurlyeq \widehat{\mathbf{\Sigma}} \preccurlyeq (1+\varepsilon)\mathbf{\Sigma}$$

  with relative error

$$\varepsilon \lesssim \sqrt{\frac{d + \log(1/\delta)}{n}},$$

  when the underlying distribution is subgaussian.

- The first step towards extending them to heavy-tailed distributions is to construct a a covariance estimator with similar properties in this case.
- In our joint work with Alessandro Rudi, we "almost" achieve this goal.[Ostrovskii and Rudi, 2019]:

$$\varepsilon \lesssim \sqrt{\frac{d \cdot \log(1/\delta)}{n}}.$$

- Recent work [Mendelson and Zhivotovskiy, 2018] suggests this can be improved.

References I

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414.

Borovkov, A. A. (1998). *Mathematical statistics*. Gordon and Breach Science Publishers.

Hazan, E., Koren, T., and Levy, K. Y. (2014). Logistic regression: tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). Random design analysis of ridge regression. *The Journal of Machine Learning Research*, 23(9):1–24.

Koltchinskii, V. and Lounici, K. (2014). Concentration inequalities and moment bounds for sample covariance operators. *arXiv:1405.2468*.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Marteau-Ferey, U., Bach, F., and Rudi, A. (2019a). Globally convergent newton methods for ill-conditioned generalized self-concordant losses. *arXiv preprint arXiv:1907.01771*.

References II

Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. (2019b). Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance.

Mendelson, S. and Zhivotovskiy, N. (2018). Robust covariance estimation under $L_4 - L_2$ norm equivalence. *arXiv:1809.10462*.

Nesterov, Y. and Nemirovski, A. S. (1994). *Interior-point polynomial algorithms in convex programming*. Society of Industrial and Applied Mathematics.

Ostrovskii, D. and Bach, F. (2018). Finite-sample analysis of M-estimators using self-concordance. *arXiv preprint arXiv:1810.06838*.

Ostrovskii, D. and Rudi, A. (2019). Affine invariant covariance estimation for heavy-tailed distributions. In *COLT*.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*.

Covariance estimation problem

Estimate the covariance matrix $\mathbf{\Sigma} = \mathbf{E}[XX^\top]$ from i.i.d. copies $X_1, ..., X_n$ of $X \in \mathbb{R}^d$.

- Sample covariance estimator:

$$\widetilde{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top.$$

- Relative spectral-norm guarantee: when X is subgaussian,

$$\frac{\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{\|\mathbf{\Sigma}\|} \lesssim \sqrt{\frac{\mathtt{r}(\mathbf{\Sigma})\log(1/\delta)}{n}} \quad \text{with probability} \quad \geq 1-\delta,$$

where $\mathtt{r}(\mathbf{\Sigma}) = \frac{\mathrm{Tr}(\mathbf{\Sigma})}{\|\mathbf{\Sigma}\|}$ is effective rank [Koltchinskii and Lounici, 2014].

- Due to affine equivariance, this gives the guarantee

$$\left(1 - \sqrt{\frac{d\log(1/\delta)}{n}}\right)\mathbf{\Sigma} \preccurlyeq \widetilde{\mathbf{\Sigma}} \preccurlyeq \left(1 + \sqrt{\frac{d\log(1/\delta)}{n}}\right)\mathbf{\Sigma}.$$

Heavy-tailed distributions

$$\frac{\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|}{\|\boldsymbol{\Sigma}\|} \lesssim \sqrt{\frac{\mathbf{r}(\boldsymbol{\Sigma}) \log(d/\delta)}{n}}$$

$$\left(1 - \sqrt{\frac{d \log(1/\delta)}{n}}\right) \boldsymbol{\Sigma} \preccurlyeq \widetilde{\boldsymbol{\Sigma}} \preccurlyeq \boldsymbol{\Sigma} \left(1 + \sqrt{\frac{d \log(1/\delta)}{n}}\right).$$

- The second guarantee is more useful in some applications (random-design linear regression, noisy PCA).
- Both require light-tailed assumptions on X, i.e. $\widetilde{\boldsymbol{\Sigma}}$ is not robust.
- Minsker (2014) proposes an estimator with a spectral-norm guarantee for heavy-tailed distributions (4th moment):

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{Min}} = \frac{1}{n} \sum_{i=1}^{n} \tau(\|X_i\|) X_i X_i^{\top}.$$

where $\tau(x)$ is the truncation map. Breaks affine equivariance!

Main idea

- Minsker (2014) proposes an estimator with a spectral-norm guarantee for **heavy-tailed distributions** (4th moment):

$$\widehat{\Sigma}^{\mathrm{Min}} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|X_i\|) X_i X_i^\top.$$

  where $\tau(x)$ is the truncation map.

- In fact, the desired $\preccurlyeq$ guarantee would hold for

$$\widehat{\Sigma}^{*} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|\Sigma^{-1/2} X_i\|) X_i X_i^\top,$$

  but it is unavailable, as $\Sigma^{-1/2}X_i$'s are not observed.

Start with $\widehat{\Sigma}_0 = \widehat{\Sigma}^{\mathrm{Min}}$, and imitate $\widehat{\Sigma}^*$ iteratively:

$$\widehat{\Sigma}_{t+1} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \tau(\|\widehat{\Sigma}_t^{-1/2} X_i\|) X_i X_i^\top,$$

Actual estimator

$$\widehat{\boldsymbol{\Sigma}}_{t+1} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \tau(\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2} X_i\|) X_i X_i^\top,$$

- **Idea 1: sample splitting**. Separate the sample $X_1, ..., X_n$ into batches, and use the new batch to compute $\widehat{\boldsymbol{\Sigma}}_{t+1}$.
- **Idea 2: Iterative regularization**. Replace $\widehat{\boldsymbol{\Sigma}}_t^{-1/2}$ with $(\widehat{\boldsymbol{\Sigma}}_t + \lambda_t \mathbf{I})^{-1/2}$, where $\lambda_t = 2^{-t}\|\boldsymbol{\Sigma}\|$. Convergence in

$$O(\log(\mathrm{cond}(\boldsymbol{\Sigma}))) \;\; \text{iterations},$$

  where $\mathrm{cond}(\boldsymbol{\Sigma})$ is the condition number of $\boldsymbol{\Sigma}$.
- Similar complexity as for the sample covariance estimator!
- Relative error

$$\varepsilon = O\left(\frac{d\log(1/\delta)}{n}\right).$$