Finite-sample Analysis of $M$-estimators using Self-concordance

Dmitrii M. Ostrovskii, Francis Bach

INRIA-CWI Workshop 2018
INRIA, Paris
Problem setup

Statistical learning problem

Given some loss \( \ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R} \), find a minimizer \( \theta^* \in \mathbb{R}^d \) of expected risk:

\[
\theta^* \in \operatorname{Argmin} L(\theta) := \mathbb{E}[\ell(Y, X^\top \theta)],
\]

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

\[
(X_1, Y_1), \ldots, (X_n, Y_n) \sim \mathcal{P} \quad (\text{i.i.d.})
\]

- \( Y \) only depends (non-linearly) on \( \eta = X^\top \theta \), a linear combination of inputs.
- Random-design classification, \( \mathcal{Y} = \{0, 1\} \), and regression, \( \mathcal{Y} = \mathbb{R} \).
- Performance of an estimate \( \hat{\theta} \) measured by excess risk \( L(\hat{\theta}) - L(\theta^*) \).
Goal

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

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

Also called $M$-estimation in statistics.

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

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

for some density $p_\eta(y)$ parametrized by $\eta$.

  • “Quasi”: the true distribution $P$ might **not** belong to this model.

Goal

Extend classical theory of qMLE, holding in the limit $n \to \infty$ with fixed $d$, to **finite-sample** setup.

• Encompass model **misspecification** and non-likelihood $M$-estimators.
Motivation 1: Classical asymptotic theory*

- **Local regularity assumptions:** $L(\theta)$ sufficiently smooth at $\theta_*$, and
  \[ H := \nabla^2 L(\theta_*) \succ 0. \]

- Gradient covariance $G := E[\nabla_\theta \ell(Y, X^T \theta_*) \nabla_\theta \ell(Y, X^T \theta_*)^T]$, and let
  \[ M := H^{-1/2} G H^{-1/2}. \]

  $d_{\text{eff}} := \text{tr}(M)$ is the **effective dimension**. In well-specified models:
  \[ G = H \Rightarrow M = I_d \Rightarrow d_{\text{eff}} = d. \]

- In the limit $n \to \infty$, Central Limit Theorem & Taylor Expansion give:
  \[ \sqrt{n} H^{-1/2}(\hat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, M), \]
  \[ n \|H^{1/2}(\hat{\theta}_n - \theta_*)\|^2 \rightsquigarrow \mathcal{N}(0, M)^2, \quad 2n(L(\hat{\theta}_n) - L(\theta_*)) \rightsquigarrow \mathcal{N}(0, M)^2. \]

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

*[Borovkov, 1998; van der Vaart, 1998; Lehmann and Casella, 2006].
Motivation 2: Random-design linear regression, I

- Gaussian model $Y = \mathcal{N}(X^\top \theta, \sigma^2)$ leads to quadratic loss and risk:
  \[
  \ell(Y, X^\top \theta) = \frac{1}{2\sigma^2} (Y - X^\top \theta)^2,
  \]
  \[
  L(\theta) - L(\theta_*) = \frac{1}{2} \|H^{1/2}(\theta - \theta_*)\|^2,
  \]
  \[
  L_n(\theta) - L_n(\theta_*) = \frac{1}{2} \|H_n^{1/2}(\theta - \theta_*)\|^2 + \langle \nabla L_n(\theta_*), \theta - \theta_* \rangle \text{ zero-mean}
  \]

- In particular, at any $\theta$ we have $\nabla^2 L(\theta) \equiv H$ and $\nabla^2 L_n(\theta) \equiv H_n$ with
  \[
  H = E[XX^\top], \quad H_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top.
  \]

**Theorem T**: Estimation of a sample covariance matrix [Vershynin, 2010]

Assume $H^{-1/2}X$ is subgaussian, i.e., has tails lighter than $\mathcal{N}(\mu, I_d)$, and

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

Then, with probability at least $1 - \delta$ it holds:

\[
0.5H \preceq H_n \preceq 2H.
\]
**Theorem 0**: Finite-sample risk bound for linear regression [Hsu et al., 2012]

Assume that $H^{-1/2}X$ and $G^{-1/2}\nabla \ell_\theta(Y, X^\top \theta_*)$ are subgaussian, and

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

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

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

**Proof sketch:**

1. Since $\nabla L_n(\hat{\theta}_n) = 0$, we have $\|H_n^{1/2}(\hat{\theta}_n - \theta_*)\|^2 = \|H_n^{-1/2}\nabla L_n(\theta_*)\|^2$.
2. Combining with Theorem T,

$$L(\hat{\theta}_n) - L(\theta_*) = \frac{1}{2}\|H^{1/2}(\hat{\theta}_n - \theta_*)\|^2 \leq 2\|H^{-1/2}\nabla L_n(\theta_*)\|^2;$$

3. Since $G^{-1/2}\nabla L_n(\theta_*)$ is the average of $n$ i.i.d. subgaussian vectors,

$$\|H^{-1/2}\nabla L_n(\theta_*)\|^2 \lesssim \frac{d_{\text{eff}} \log(1/\delta)}{n}.$$
Towards the general case

• Generally, risk is not quadratic, and Hessians are not constant:

\[
\nabla^2 L(\theta) = H(\theta), \quad \nabla^2 L_n(\theta) = H_n(\theta).
\]

• To extend the previous argument, we must control the precision of local quadratic approximation of \( L_n(\theta) \) and \( L(\theta) \) around \( \theta_* \).

• We exploit self-concordance, a concept introduced in [Nesterov and Nemirovski, 1994] in the theory of interior-point methods, and brought to the statistical learning context in [Bach, 2010] to study logistic regression.
Self-concordant losses

We always assume that $\ell(y, \eta)$ is convex in the second argument.

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

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

- While the above definition is homogeneous in $\eta$, the next one is not:

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

$$|\ell'''_{\eta}(y, \eta)| \leq C\ell''_{\eta}(y, \eta).$$

- **PSC** losses are somewhat more common than **SC** ones.
- However, we will see that obtaining optimal rate for **PSC** losses requires somewhat larger sample size.
Example 1: Generalized linear models

Conditional negative log-likelihood of $Y$ given $\eta = X^T \theta$ in the form

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

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

$$a(\eta) = \log \int_Y e^{y\eta + b(y)} dy.$$

This defines the density $p_\eta(y) \propto e^{y\eta + b(y)}$ such that $a(\eta) = E_{p_\eta}[Y]$, and

$$\ell^{(s)}_{\eta}(y, \eta) = a^{(s)}(\eta) = E_{p_\eta}[(Y - E_{p_\eta} Y)^s], \quad s \geq 2.$$

**SC/PSC** specify a relation between 2nd and 3rd central moments of $p_\eta(\cdot)$

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

$$\left|a'''(\eta)\right| \leq E_{p_\eta} |(Y - E_{p_\eta} [Y])^3| \leq E_{p_\eta} [(Y - E_{p_\eta} [Y])^2] = a''(\eta).$$

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

  $\varphi''(t)$ discontinuous at $\pm 1$.

**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 $[-1, 1]$:

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

Recall that in the general case, we have the Hessian process $H(\theta)$, given by

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

where $\tilde{X}(\theta) := [\ell''(Y, X^\top \theta)]^{1/2}X$ is the curvature-scaled design.

**Theorem 1:** Finite-sample excess risk bound for self-concordant losses

Assume that the loss is SC, and $G^{-1/2}\nabla\ell_\theta(Y, X^\top \theta_*)$ and $H(\theta_*)^{-1/2}\tilde{X}(\theta_*)$ are subgaussian. Whenever

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

with probability $1 - \delta$ it holds

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

- Distribution conditions are local (only at $\theta_*$);
- Large sample complexity – scaling as the product $O(d_{\text{eff}} d)$. 
Given \( H(\theta) = \nabla^2 L(\theta) \), consider **Dikin ellipsoids** of \( L(\theta) \) at \( \theta_0 \):

\[
\Theta(\theta_0, r) := \{ \theta : \|H(\theta_0)^{1/2}(\theta - \theta_0)\|^2 \leq r^2 \}.
\]

**Key Observation.** Suppose that \( H_n(\theta) \approx H_n(\theta_*) \) w.h.p. for any \( \theta \in \Theta(\theta_*, r) \). Then, \( \hat{\theta}_n \in \text{Argmin} L_n(\theta) \) can be localized to \( \Theta(\theta_*, r) \) once

\[
\|H(\theta_*)^{-1/2} \nabla L_n(\theta_*)\|^2 \lesssim r^2,
\]

**Proof sketch:**

- Indeed, by definition of \( \hat{\theta}_n \), \( L_n(\hat{\theta}_n) \leq L_n(\theta_*) \). Assume \( \hat{\theta}_n \notin \Theta_n(\theta_*, r) \).
- Pick \( \bar{\theta}_n \in [\theta_*, \hat{\theta}_n] \) on the border of \( \Theta_n(\theta_*, r) \). Still, \( L_n(\bar{\theta}_n) \leq L_n(\theta_*) \).

\[
0 \geq L_n(\bar{\theta}_n) - L_n(\theta_*) \approx \langle \nabla L_n(\theta_*) , \bar{\theta}_n - \theta_* \rangle + \left\| H_n(\theta_*)^{1/2}(\bar{\theta}_n - \theta_*) \right\|^2 \approx r^2 \text{ (by Theorem T)}.
\]

- By Cauchy-Schwarz, we arrive at \( \|H(\theta_*)^{-1/2} \nabla L_n(\theta_*)\|^2 \gtrsim r^2 \).

**Contradiction!**
Once $\hat{\theta}_n$ has been localized to the neighborhood of $\theta_*$ where $L_n(\theta)$ is quadratic, we can mimic the argument for linear regression.

Localization is guaranteed once

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

which leads to the second threshold for $n$:

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

Now the question is:

What is the radius $r$ of the Dikin ellipsoid in which $H_n(\theta) \approx H_n(\theta_*)$?

- Short answer: we can afford $r^2 \approx 1/d$ using self-concordance.
What is the radius $r$ of the Dikin ellipsoid in which $H_n(\theta) \approx H_n(\theta^*)$?

1. Recall that

$$H_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell''(Y_i, X_i^\top \theta) X_i X_i.$$ 

2. Integrating $|\ell'''(y, \eta)| \leq [\ell''(y, \eta)]^{\frac{3}{2}}$ from $\eta^* = X^\top \theta^*$ to $\eta = X^\top \theta$,

$$\frac{1}{(1 + [\ell''(y, \eta^*)])^{\frac{1}{2}} |\eta - \eta^*|^{2}} \leq \frac{\ell''(y, \eta)}{\ell''(y, \eta^*)} \leq \frac{1}{(1 - [\ell''(y, \eta^*)])^{\frac{1}{2}} |\eta - \eta^*|^{2}},$$

$$\frac{1}{(1 + |\langle \tilde{X}(\theta^*), \theta - \theta^* \rangle|)}^{2} \leq \frac{\ell''(Y, X^\top \theta)}{\ell''(Y, X^\top \theta^*)} \leq \frac{1}{(1 - |\langle \tilde{X}(\theta^*), \theta - \theta^* \rangle|)}^{2}.$$ 

3. The ratio is bounded if $|\langle \tilde{X}(\theta^*), \theta - \theta^* \rangle| \leq c < 1$, i.e., by Cauchy-Schwarz,

$$\left\| H(\theta^*)^{-1/2} \tilde{X}(\theta^*) \right\| \cdot \left\| H(\theta^*)^{1/2}(\theta - \theta^*) \right\| \leq c \Rightarrow r \gtrsim \frac{1}{\sqrt{d}}.$$


**Improved result**

**Theorem 2:** Improved sample complexity for self-concordant losses

Assume the loss is $\text{SC}$, $G^{-1/2}\nabla \ell_\theta(Y, X^\top \theta_*)$ is subgaussian, and $H(\theta)^{-1/2}\tilde{X}(\theta)$ is subgaussian in the unit Dikin ellipsoid of $L(\theta)$ at $\theta_*$:

$$\Theta(\theta_*, 1) = \{\theta : \|H(\theta_*)^{1/2}(\theta - \theta_*)\| \leq 1\}.$$  

Then for $(\star)$ it is sufficient that

$$n \gtrsim d \log(d/\delta) \lor d_{\text{eff}} \log(1/\delta),$$

**Main idea:**

- Sample complexity $n \gtrsim d_{\text{eff}} d$ in Theorem 1 is due to Hessian approximation in the small Dikin ellipsoid with $r = O(1/\sqrt{d})$ rather than $r = O(1)$.

- We need to prove that $H_n(\theta) \approx H_n(\theta_*)$ for $\theta \in \Theta(\theta_*, 1)$. To do this, we combine self-concordance with a **covering argument**.
Covering the Dikin ellipsoid

1. It is rather easy to prove first that $\mathbf{H}(\theta)$ is near-constant on $\Theta(\theta_*, 1)$.
2. By SC, $\mathbf{H}_n(\theta)$ is near-constant in smaller ellipsoids $\Theta(\theta, 1/\sqrt{d})$.
3. Now cover $\Theta(\theta_*, 1)$ by $\Theta(\theta, 1/\sqrt{d})$ with $\theta$ in the epsilon-net $\mathcal{N}_\varepsilon$, and control uniform deviations $\mathbf{H}_n(\theta)$ from $\mathbf{H}(\theta)$ on $\mathcal{N}_\varepsilon$. OK since $\log |\mathcal{N}_\varepsilon| = O(d \log d)$. 
Because of the “incorrect” power of $\ell''$ in PSC, we need an extra condition:

$$E[XX^\top] \leq \rho E[\ell''(Y, X^\top \theta^*_0)XX^\top].$$

for some $\rho > 0$. This condition is standard in logistic regression [Bach, 2010].

- We obtain similar results, but with $\rho$ times worse sample complexity.
- Worst-case bounds on $\rho$ can be exponentially bad [Hazan et al., 2014]. However, this is not the case in practice [Bach, 2010].
We use self-concordance – a concept from optimization – to obtain statistical results – near-optimal rates in finite-sample regimes in some statistical models.

Perspectives:

- Regularized estimators.
- Iterative algorithms: stochastic approximation, Quasi-Newton, ...
- Other models: covariance matrix estimation with log det loss, ...

Thank you!


• Once $\hat{\theta}_n$ is in neighborhood of $\theta_*$ where $L_n(\theta)$ is quadratic, we’re done:

$$L_n(\hat{\theta}_n) - L_n(\theta_*) \lesssim \|H_n(\theta_*)^{1/2}(\hat{\theta}_n - \theta_*)\|^2 \lesssim \|H_n^{-1/2}(\theta_*)\nabla L_n(\theta_*)\|^2;$$

by Theorem T, as long as $n \geq d + \log(1/\delta)$,

$$\|H_n^{-1/2}(\theta_*)\nabla L_n(\theta_*)\|^2 \approx \|H^{-1/2}(\theta_*)\nabla L_n(\theta_*)\|^2 \lesssim \frac{d_{\text{eff}} \log(1/\delta)}{n}.$$

Similarly for $L(\hat{\theta}_n) - L(\theta_*)$.

• Localization is guaranteed once $\|H_n^{-1/2}(\theta_*)\nabla L_n(\theta_*)\|^2 \lesssim r^2$, which leads to the second threshold for $n$:

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

• **Now the question is:**

*What is the radius $r$ of the Dikin ellipsoid in which $H_n(\theta) \approx H_n(\theta_*)$?*