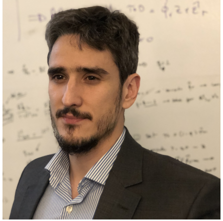
Non-Asymptotic Analysis of M-Estimation for Statistical Learning and Inference under Self-Concordance

Lang Liu

University of Washington

October 21, 2022

Collaborators



Carlos Cinelli

Zaid Harchaoui

@ COLT 2022
@ NeurIPS 2022 workshop on Score-Based Methods
Submitted @ AISTATS 2023

Maximum Likelihood Estimation

- **Data** $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$.
- **Parametric family** $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$.
- **Target parameter**

$$\theta_\star := \arg\min_{\theta \in \Theta} \Big\{ \mathbb{E}[-\log P_\theta(Z)] =: \mathbb{E}[\underbrace{\ell(\theta; Z)}_{\text{Loss function}}] =: \underbrace{L(\theta)}_{\text{Population risk}} \Big\}.$$

- **Maximum likelihood estimator** (MLE)

$$\theta_n := \arg\min_{\theta \in \Theta} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log P_\theta(Z_i) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i) =: \underbrace{L_n(\theta)}_{\text{Empirical risk}} \right\}.$$

Generalized Linear Models

- **Data** $Z := (X, Y) \in \mathcal{X} \times \mathcal{Y}$.
- **Sufficient statistic** $t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$.
- **Reference measure** μ on $\mathcal{Y}$.
- **Statistical model**

$$p_\theta(y \mid x) \sim \frac{\exp(\theta^\top t(x, y))}{\int \exp(\theta^\top t(x, \bar{y})) \mathrm{d}\mu(\bar{y})} \mathrm{d}\mu(y).$$

- **Loss function**

$$\ell(\theta; z) = -\theta^\top t(x, y) + \log \int \exp(\theta^\top t(x, \bar{y})) \mathrm{d}\mu(\bar{y}).$$

Example: Softmax Regression

- **Data space** $\mathcal{X} \subset \mathbb{R}^\tau$ and $\mathcal{Y} = \{1, \ldots, K\}$.
- **Statistical model**

$$p(y = k \mid x) \sim \frac{\exp(w_k^\top x)}{\sum_{j=1}^K \exp(w_j^\top x)}.$$

Example: Softmax Regression

- **Data space** $\mathcal{X} \subset \mathbb{R}^{\tau}$ and $\mathcal{Y} = \{1, \ldots, K\}$.
- **Statistical model**

$$p(y = k \mid x) \sim \frac{\exp(w_k^\top x)}{\sum_{j=1}^{K} \exp(w_j^\top x)}.$$

- Define $\theta^\top := (w_1^\top, \ldots, w_K^\top)$ and

$$t(x, y)^\top := (0_\tau^\top, \ldots, 0_\tau^\top, x^\top, 0_\tau^\top, \ldots, 0_\tau^\top).$$

  Then we have

$$p(y = k \mid x) \sim \frac{\exp(\theta^\top t(x, k))}{\sum_{y=1}^{K} \exp(\theta^\top t(x, y))}.$$

Example: Conditional Random Fields

- **Data space** $\mathcal{X} = \mathbb{X}^T$ and $\mathcal{Y} = \mathbb{Y}^T$.
- **Conditional random fields on a chain**

$$p(y \mid x) \propto \exp\left\{\sum_{t=1}^{T-1} \lambda_t f_t(x, y_t, y_{t+1}) + \sum_{t=1}^{T} \mu_t g_t(x, y_t)\right\} \mathrm{d}\mu(y).$$

Example: Conditional Random Fields

- **Data space** $\mathcal{X} = \mathbb{X}^T$ and $\mathcal{Y} = \mathbb{Y}^T$.
- **Conditional random fields on a chain**

$$p(y \mid x) \propto \exp\left\{\sum_{t=1}^{T-1} \lambda_t f_t(x, y_t, y_{t+1}) + \sum_{t=1}^{T} \mu_t g_t(x, y_t)\right\} \mathrm{d}\mu(y).$$

- Define $\theta^\top := (\lambda_1, \ldots, \lambda_{T-1}, \mu_1, \ldots, \mu_T)$ and

$$t(x, y)^\top := (f_1(x, y_1, y_2), \ldots, f_{T-1}(x, y_{T-1}, y_T), g_1(x, y_1), \ldots, g_T(x, y_T)).$$

Then we have

$$p(y \mid x) \sim \frac{\exp(\theta^\top t(x, y))}{\int \exp(\theta^\top t(x, \bar{y}))\mathrm{d}\mu(\bar{y})} \mathrm{d}\mu(y).$$

Related Work: Asymptotic Theory†

Well-specified model: $P \in \mathcal{P}_\Theta$

$$\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, H_\star^{-1}),$$

where $H_\star := H(\theta_\star) := \nabla^2 L(\theta_\star)$.

† Cramér '46, Huber '74, Ibragimov and Has'minskii '81, van der Vaart '00.

Related Work: Asymptotic Theory[†]

Well-specified model: $P \in \mathcal{P}_\Theta$

$$\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, H_\star^{-1}),$$

where $H_\star := H(\theta_\star) := \nabla^2 L(\theta_\star)$.

Mis-specified model: $P \notin \mathcal{P}_\Theta$

$$\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, H_\star^{-1} G_\star H_\star^{-1}),$$

where $G_\star := G(\theta_\star) := \mathbb{E}[\nabla \ell(\theta_\star; Z) \nabla \ell(\theta_\star; Z)^\top]$.

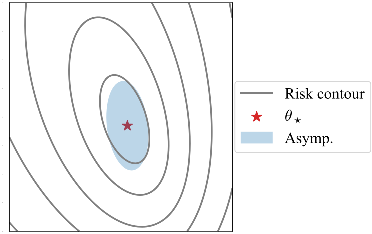
[†] Cramér '46, Huber '74, Ibragimov and Has'minskii '81, van der Vaart '00.

Asymptotic Confidence Set

- **Asymptotic normality** $\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, \Sigma)$.
- **Consistent estimator** $\Sigma_n \to_p \Sigma$.
- **Slutsky's lemma** $n\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \to_d \chi_d^2$.

Asymptotic Confidence Set

- **Asymptotic normality** $\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, \Sigma)$.
- **Consistent estimator** $\Sigma_n \to_p \Sigma$.
- **Slutsky's lemma** $n\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \to_d \chi_d^2$.
- **Asymptotic confidence set** $\{\theta : \|\Sigma_n^{-1/2}(\theta_n - \theta)\|_2^2 \leq q_{\chi_d^2}(1-\delta)/n\}$
  - Asymptotically tight.
  - Valid for $n \to \infty$ and fixed d.



Related Work: Non-Asymptotic Theory

Specific models

- Gaussian regression (Baraud ’04).
- Ridge regression (Hsu et al ’14).
- Logistic regression (Bach ’10).

Related Work: Non-Asymptotic Theory

Specific models

- Gaussian regression (Baraud '04).
- Ridge regression (Hsu et al '14).
- Logistic regression (Bach '10).

General approaches

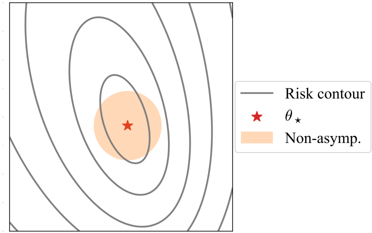
- Empirical process (Spokoiny '12).
- Convex optimization (Ostrovskii and Bach '21).

Non-Asymptotic Confidence Set under Strong Convexity

- **Excess risk** $L(\theta_n) - L(\theta_\star) \le O(n^{-1})$ with high probability.
- **Taylor's theory** $\|H(\bar{\theta})^{1/2}(\theta_n - \theta_\star)\|_2^2 \le O(n^{-1})$ with high probability.
- **Strong convexity** $H(\theta) \succeq \lambda I$ implies $\|\theta_n - \theta_\star\|_2^2 \le O((n\lambda)^{-1})$ with high probability.

Non-Asymptotic Confidence Set under Strong Convexity

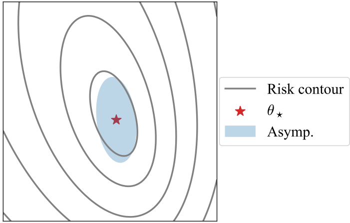
- **Excess risk** $L(\theta_n) - L(\theta_\star) \leq O(n^{-1})$ with high probability.
- **Taylor's theory** $\|H(\bar{\theta})^{1/2}(\theta_n - \theta_\star)\|_2^2 \leq O(n^{-1})$ with high probability.
- **Strong convexity** $H(\theta) \succeq \lambda I$ implies $\|\theta_n - \theta_\star\|_2^2 \leq O((n\lambda)^{-1})$ with high probability.
- **Non-asymptotic confidence set** $\{\theta : \|\theta_n - \theta\|_2^2 \leq O((n\lambda)^{-1})\}$.
  - Conservative.
  - Valid for all n and d.



Asymptotic and Non-Asymptotic Confidence Sets

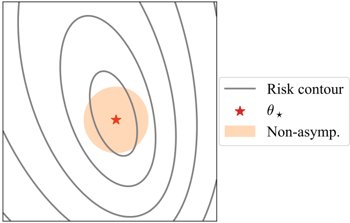
Asymptotic theory

- $n\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \to_d \chi_d^2$.
- Slutsky's Lemma.
- Asymptotically tight.
- Valid for $n \to \infty$ and fixed d.



Non-asymptotic theory

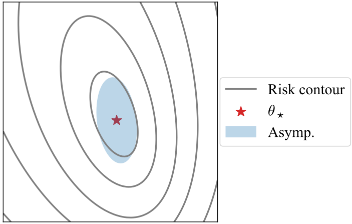
- $\|\theta_n - \theta_\star\|_2^2 \le O\big((n\lambda)^{-1}\big)$.
- Strong convexity.
- Conservative.
- Valid for all n and d.



Asymptotic and Non-Asymptotic Confidence Sets

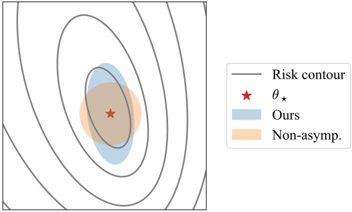
Asymptotic theory

- $n\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \to_d \chi_d^2$.
- Slutsky's Lemma.
- Asymptotically tight.
- Valid for $n \to \infty$ and fixed d.



Our contribution

- $\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \le O(n^{-1})$.
- Self-concordance.
- Conservative.
- Valid for $n > O(d + d_\star)$.



Non-Asymptotic Theory under Self-Concordance

Non-asymptotic theory: with high probability,

$$\underbrace{\nabla L(\theta_\star)(\theta_n - \theta_\star)}_{0} + \frac{1}{2}(\theta_n - \theta_\star)^\top H(\bar{\theta})(\theta_n - \theta_\star) = \underbrace{L(\theta_n) - L(\theta_\star)}_{\text{Excess risk}} \leq O(n^{-1}).$$

Non-Asymptotic Theory under Self-Concordance

Non-asymptotic theory: with high probability,

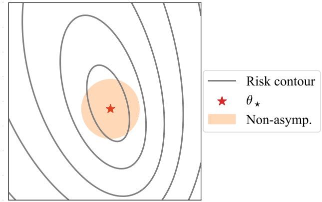
$$\underbrace{\nabla L(\theta_\star)(\theta_n - \theta_\star)}_{0} + \frac{1}{2}(\theta_n - \theta_\star)^\top H(\bar\theta)(\theta_n - \theta_\star) = \underbrace{L(\theta_n) - L(\theta_\star)}_{\text{Excess risk}} \le O(n^{-1}).$$

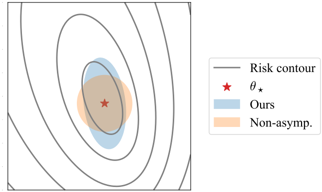
Strong convexity $H(\theta) \succeq \lambda I$

$$\lambda \|\theta_n - \theta_\star\|_2^2 \le O(n^{-1}).$$

Self-Concordance $H(\bar\theta) \approx H_n(\theta_n)$

$$\|H_n(\theta_n)^{1/2}(\theta_n - \theta_\star)\|_2^2 \le O(n^{-1}).$$

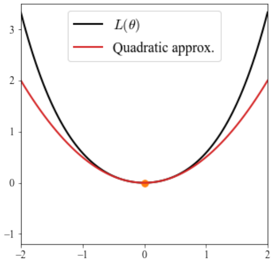




Strong Convexity versus Self-Concordance

Strong convexity

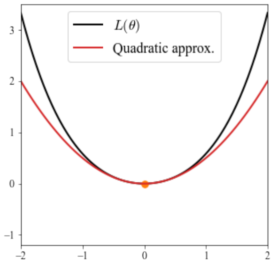
- Globally lower bounded Hessian.
- No control on how Hessian varies.



Strong Convexity versus Self-Concordance

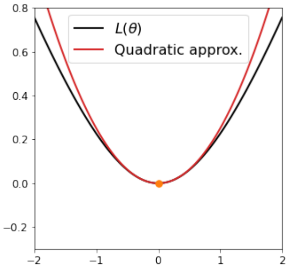
Strong convexity

- Globally lower bounded Hessian.
- No control on how Hessian varies.



Self-concordance

- No global lower bound.
- Slowly varying Hessian.



Self-Concordance

Define $\mathrm{D}f(x)[u] := \frac{\mathrm{d}}{\mathrm{d}t} f(x + tu)|_{t=0}$ and $\mathrm{D}^2 f(x)[u, u] := \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x + tu)|_{t=0}$.

Definition 1 (Nesterov and Nemirovskii ’94)

Let f be closed and convex. We say f is *self-concordant* with parameter $R > 0$ if

$$\left|\mathrm{D}^3 f(x)[u, u, u]\right| \le R \left|\mathrm{D}^2 f(x)[u, u]\right|^{3/2}.$$

Self-Concordance

Define $\mathrm{D}f(x)[u] := \frac{\mathrm{d}}{\mathrm{d}t} f(x + tu)|_{t=0}$ and $\mathrm{D}^2 f(x)[u, u] := \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x + tu)|_{t=0}$.

Definition 1 (Nesterov and Nemirovskii '94)

Let f be closed and convex. We say f is *self-concordant* with parameter $R > 0$ if

$$\left|\mathrm{D}^3 f(x)[u, u, u]\right| \le R \left|\mathrm{D}^2 f(x)[u, u]\right|^{3/2}.$$

- Newton's method.
- Interior point methods.
- Most non-quadratic loss functions are not self-concordant.

Pseudo Self-Concordance

Definition 2 (Bach '10)

Let f be closed and convex. We say f is *pseudo self-concordant* with parameter $R > 0$ if

$$\left|D^3 f(x)[u, u, u]\right| \leq R\|u\|_2 D^2 f(x)[u, u].$$

- **GLMs** with $\|t(x, y)\| \leq M$ are pseudo self-concordant with $R = 2M$.

Pseudo Self-Concordance

Definition 2 (Bach '10)

Let f be closed and convex. We say f is *pseudo self-concordant* with parameter $R > 0$ if

$$\left|\mathrm{D}^3 f(x)[u, u, u]\right| \leq R\|u\|_2 \mathrm{D}^2 f(x)[u, u].$$

- **GLMs** with $\|t(x, y)\| \leq M$ are pseudo self-concordant with $R = 2M$.
- **Hessian approximation**:

$$e^{-R\|y-x\|_2} \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq e^{R\|y-x\|_2} \nabla^2 f(x).$$

- **Localization**: $x_\star := \arg\min_x f(x)$ satisfies

$$\|x_\star - x\|_{\nabla^2 f(x)} \lesssim \|\nabla f(x)\|_{\nabla^2 f(x)^{-1}},$$

where $\|u\|_A := \sqrt{u^\top A u}$.

Effective Dimension

Effective dimension $d_\star := \mathbf{Tr}(\Omega_\star) := \mathbf{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2})$

- **Well-specified model**: $d_\star = d$.
- **Mis-specified model**:
  - Problem-specific characterization of the complexity of Θ.
  - $\sqrt{n} H_\star^{1/2}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, \Omega_\star)$.

Effective Dimension

Effective dimension $d_\star := \mathbf{Tr}(\Omega_\star) := \mathbf{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2})$

- **Well-specified model**: $d_\star = d$.
- **Mis-specified model**:
  - Problem-specific characterization of the complexity of Θ.
  - $\sqrt{n} H_\star^{1/2}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, \Omega_\star)$.

		Poly-Poly	**Poly-Exp**	**Exp-Poly**	**Exp-Exp**
Eigendecay	$G_\star$	$i^{-\alpha}$	$i^{-\alpha}$	$e^{-\mu i}$	$e^{-\mu i}$
	$H_\star$	$i^{-\beta}$	$e^{-\nu i}$	$i^{-\beta}$	$e^{-\nu i}$
Ratio	$d_\star / d$	$d^{(\beta-\alpha)\vee(-1)}$	$d^{-\alpha} e^{\nu d}$	d^{-1}	1 if $\mu = \nu$ d^{-1} if $\mu > \nu$ $d^{-1} e^{(\nu-\mu)d}$ if $\mu < \nu$

Main Results

Theorem 3 (Informal)

*Under the **pseudo self-concordance** assumption and other assumptions, whenever*

$$n \gtrsim O(d + d_\star),$$

with probability at least $1 - \delta$*, the MLE* θ_n *uniquely exists and satisfies*

$$n \|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim d_\star + \|\Omega_\star\|_2 \log(1/\delta).$$

Main Results

Theorem 3 (Informal)

*Under the **pseudo self-concordance** assumption and other assumptions, whenever*

$$n \gtrsim O(d + d_\star),$$

with probability at least $1 - \delta$, the MLE θ_n uniquely exists and satisfies

$$n \|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim d_\star + \|\Omega_\star\|_2 \log(1/\delta).$$

- Well-specified model $d_\star = d$ and $\|\Omega_\star\|_2 = 1$.
- Recall $n \|\theta_n - \theta_\star\|_{H_\star}^2 \to_d \chi_{d_\star}^2$.
- Characterize the critical sample size.

Main Results

Proof Sketch

- **Pseudo self-concordance**: $\ell(\cdot; z)$ pseudo self-concordant implies L_n as well.
- **Localization**: $\|\theta_n - \theta_\star\|^2_{H_n(\theta_\star)} \lesssim \|\nabla L_n(\theta_\star)\|^2_{H_n(\theta_\star)^{-1}}$.
- **Matrix concentration**: $H_\star/2 \preceq H_n(\theta_\star) \preceq 2H_\star$, which implies

$$\|\theta_n - \theta_\star\|^2_{H_\star} \lesssim \|\nabla L_n(\theta_\star)\|^2_{H_\star^{-1}} .$$

- **Quadratic form of sub-Gaussian vectors**:

$$n\|\theta_n - \theta_\star\|^2_{H_\star} \lesssim d_\star + \|\Omega_\star\|_2 \log(1/\delta).$$

Main Results

Confidence bound

- Approximate $H_\star$ by $H_n(\theta_n)$ (**Hessian approximation + matrix concentration**).
- Approximate $G_\star$ by $G_n(\theta_n)$ (**Lipschitz property of the second moment**).
- Estimators $\Omega_n(\theta_n) := H_n(\theta_n)^{-1/2} G_n(\theta_n) H_n(\theta_n)^{-1/2}$ and $d_n := \mathbf{Tr}\left(\Omega_n(\theta_n)\right)$.

Theorem 4 (Informal)

*Under the **pseudo self-concordance** assumption and other assumptions, whenever*

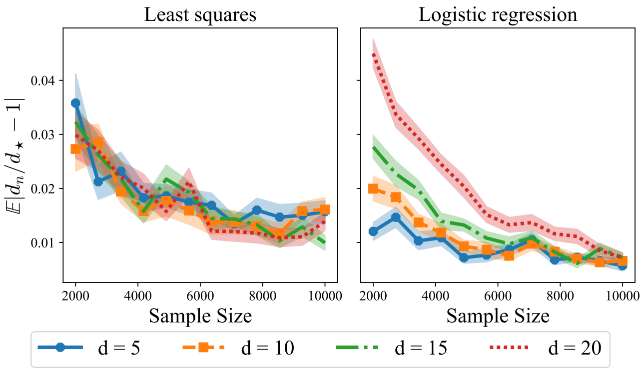
$$n \gtrsim O(d \log n + d_\star),$$

with probability at least $1 - \delta$*, the MLE* θ_n *uniquely exists and satisfies*

$$n \left\| \theta_n - \theta_\star \right\|^2_{H_n(\theta_n)} \lesssim d_n + \left\| \Omega_n(\theta_n) \right\|_2 \log\left(1/\delta\right).$$

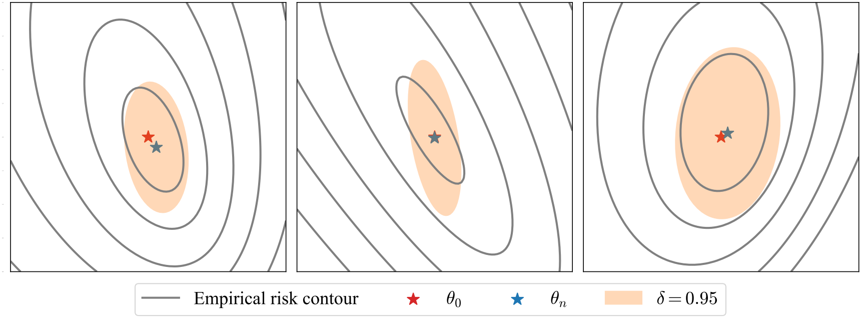
Numerical Illustration: Approximation of the Effective Dimension

- **Least squares**: $X \sim \mathcal{N}(0, I_d)$ and $Y = \theta_0^\top X + \mathcal{N}(0, 1)$.
- **Logistic regression**: $X \sim \mathcal{N}(0, I_d)$ and $\mathbb{P}(Y = 1) = \sigma(\theta_0^\top X)$.



Numerical Illustration: Shape of the Confidence Set

- **Logistic regression**: $X \sim \mathcal{N}(0, \Sigma)$ and $\mathbb{P}(Y = 1) = \sigma(\theta_0^\top X)$.



Extension: Goodness of Fit Testing

Goodness of fit testing

$$\mathbf{H}_0 : \theta_\star = \theta_0 \leftrightarrow \mathbf{H}_1 : \theta_\star \neq \theta_0.$$

Test	Test statistic	$\theta_\star = \theta_0$	$\theta_\star = \theta_0 + \omega(n^{-1/2})$	$\theta_\star = \theta_0 + O(n^{-1/2})$
Rao's score	$\|\nabla \ell_n(\theta_0)\|^2_{H_n(\theta_0)^{-1}}$	$O(d/n)$	$1 - o(1)$	$O(1)$
Likelihood ratio	$2[\ell_n(\theta_0) - \ell_n(\theta_n)]$	$O(d/n)$	$1 - o(1)$	$O(1)$
Wald	$\|\theta_n - \theta_0\|^2_{H_n(\theta_n)}$	$O(d/n)$	$1 - o(1)$	$O(1)$

Extension: Semi-Parametric Estimation

- **Nuisance parameter** $g_0 \in (\mathcal{G}, \|\cdot\|_{\mathcal{G}})$.
- **Population risk** $L(\theta, g) := \mathbb{E}[\ell(\theta, g; Z)]$.
- **Two-step learning procedure based on sample-splitting**[‡]
  - Obtain a nonparametric estimator $\hat{g}$ on one sub-sample.
  - Estimate $\theta_\star$ via empirical risk minimization on another sub-sample:

$$\theta_n = \arg\min_{\theta \in \Theta} L_n(\theta, \hat{g}).$$

Example 5 (Robinson '88)

Let Y outcome, D treatment, and X control. Consider

$$Y = D\theta_\star + g_0(X) + U.$$

[‡]Chernozhukov et al '18, Foster and Syrgkanis '20.

Extension: Semi-Parametric Estimation

Theorem 6 (Informal)

Under the ***pseudo self-concordance*** *and other assumptions, with probability at least* $1 - \delta$*,*

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim \frac{d_\star}{n} \log(1/\delta) + \|\hat{g} - g_0\|_{\mathcal{G}}^2 .$$

- If g_0 is p-smooth, it can be estimated at rate $O(n^{-p/(2p+d)})$.
- The term $\|\hat{g} - g_0\|_{\mathcal{G}}^2$ **cannot** achieve the $O(n^{-1})$ rate.

Extension: Semi-Parametric Estimation

Neyman orthogonality (Neyman '79)

$$\mathrm{D}_g \nabla_\theta L(\theta_\star, g_0)[g - g_0] = 0.$$

Theorem 7 (Informal)

Under the ***pseudo self-concordance****, Neyman orthogonality, and other assumptions, with probability at least* $1 - \delta$*,*

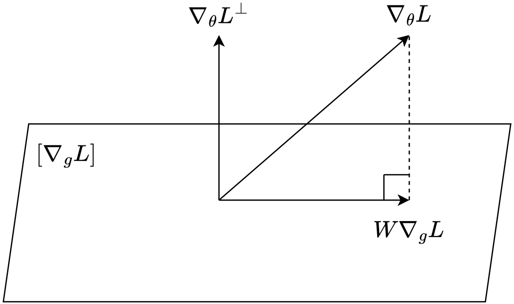
$$\|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim \frac{d_\star}{n} \log(1/\delta) + \|\hat{g} - g_0\|_{\mathcal{G}}^4 .$$

- If g_0 is p-smooth, it can be estimated at rate $O(n^{-p/(2p+d)})$.
- The term $\|\hat{g} - g_0\|_{\mathcal{G}}^4$ **can** achieve the $O(n^{-1})$ rate as long as $p \geq d/2$.

Extension: Semi-Parametric Estimation

Neyman orthogonality (Neyman '79)

$$\mathrm{D}_g \nabla_\theta L(\theta_\star, g_0)[g - g_0] = 0.$$



Summary

- Non-asymptotic bounds for the M-estimator under **self-concordance**.
- **Finite-sample counterpart** of the asymptotic confidence set.
- Characterize the **critical sample size** enough to enter the asymptotic regime.
- Extension to **goodness-of-fit testing** and **semi-parametric estimation**.

Follow-up work with Jillian and Krishna

IFDS
Institute for Foundations of Data Science

Partially Linear Model

Let Y outcome, D treatment, and X control. Consider

$$\begin{aligned} Y &= D\theta_0 + \alpha_0(X) + U \\ D &= \beta_0(X) + V. \end{aligned}$$

- Partialling out the effect of X

$$Y = (D - \beta_0(X))\theta_0 + \gamma_0(X) + U.$$

- Reparameterization $g_0 = (\beta_0, \gamma_0)$.
- Neyman orthogonal risk

$$L(\theta, g) := \mathbb{E}\left[(Y - \gamma(X) - (D - \beta(X))\theta)^2\right].$$

Proof Sketch for the OSL Estimation Bound

By Taylor's theorem,

$$
\begin{aligned}
0 &\geq L_n(\theta_n, \hat{g}) - L_n(\theta_\star, \hat{g}) \\
&= \nabla_\theta L_n(\theta_\star, \hat{g})^\top (\theta_n - \theta_\star) + \|\theta_n - \theta_\star\|^2_{H_n(\bar{\theta}, \hat{g})}/2 \\
&= [\nabla_\theta L_n(\theta_\star, \hat{g}) - \nabla_\theta L(\theta_\star, \hat{g})]^\top (\theta_n - \theta_\star) + \nabla_\theta L(\theta_\star, \hat{g})^\top (\theta_n - \theta_\star) + \|\theta_n - \theta_\star\|^2_{H_n(\bar{\theta}, \hat{g})}/2 \\
&\geq \|\nabla_\theta L_n(\theta_\star, \hat{g}) - \nabla_\theta L(\theta_\star, \hat{g})\|_{H_\star^{-1}} \|\theta_n - \theta_\star\|_{H_\star} + \nabla_\theta L(\theta_\star, \hat{g})^\top (\theta_n - \theta_\star) + \|\theta_n - \theta_\star\|^2_{H_n(\bar{\theta}, \hat{g})}/2 \\
&\gtrsim -\left[\sqrt{d_\star/n} + \|\hat{g} - g_0\|^2_{\mathcal{G}}\right] \|\theta_n - \theta_\star\|_{H_\star} + \|\theta_n - \theta_\star\|^2_{H_\star}.
\end{aligned}
$$