UNIVERSITY *of* WASHINGTON

Likelihood Score under Generalized Self-Concordance

Lang Liu, Zaid Harchaoui

Department of Statistics, University of Washington

Overview

- Establish **non-asymptotic bounds** on the normalized likelihood score whose tail behavior is governed by an **effective dimension**.
- Obtain finite-sample **confidence bound** for the maximum likelihood estimator and analysis for **Rao's score test**.
- Allow the loss to be **generalized self-concordance** and the model to be **mis-specified**.

Maximum Likelihood Estimation

Problem. Let $Z \sim \mathbb{P}$ and $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$.

$$\theta_\star = \arg\min_{\theta\in\Theta}\left\{ \underbrace{L(\theta)}_{\text{population risk}} := \mathbb{E}_{Z\sim\mathbb{P}}[-\log p_\theta(Z)] \right\}.$$

When the model is **well-specified**, i.e., $\mathbb{P} = P_{\theta_0}$, assume $\theta_\star = \theta_0$.

Empirical risk minimization. Given an i.i.d. sample $\{Z_i\}_{i=1}^n \sim \mathbb{P}$, the **maximum likelihood estimator** is

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

$S_n(\theta) := -\nabla L_n(\theta)$ is the **likelihood score**.

Generalized Linear Models

Generalized linear models (GLM). Let $Z := (X, Y) \in \mathcal{X} \times \mathcal{Y}$.

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

- $t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ **sufficient statistic**.
- μ **reference measure** on $\mathcal{Y}$, e.g., Lebesgue/counting measure.

Example: softmax regression. $\mathcal{X} \subset \mathbb{R}^\tau$ and $\mathcal{Y} = \{1, \dots, K\}$.

$$p(y = k \mid x) \propto \exp(w_k^\top x) \propto \exp\left(\theta^\top t(x,k)\right),$$

where $\theta^\top := (w_1^\top, \dots, w_K^\top)$ and $t(x,y)^\top := (0_\tau^\top, \dots, 0_\tau^\top, x^\top, 0_\tau^\top, \dots, 0_\tau^\top)$.

Normalized Likelihood Score

Classical asymptotic theory. We are interested in the **normalized score** $\tilde{S}_n(\theta) := H_n(\theta)^{-1/2}S_n(\theta)$ where $H_n(\theta) := \nabla^2 L_n(\theta)$.

$$\sqrt{n}S_n(\theta_\star) := \frac{1}{\sqrt{n}}\sum_{i=1}^n[-\nabla_\theta \log p_\theta(Z_i)] \to_d \mathcal{N}(0, G(\theta_\star))$$

$$H_n(\theta_\star) := \frac{1}{n}\sum_{i=1}^n\left[-\nabla_\theta^2 \log p_\theta(Z_i)\right] \to_p H(\theta_\star) := \nabla^2 L(\theta)$$

with $G(\theta) := \mathbb{E}_{Z\sim P}[\nabla_\theta \log p_\theta(Z)\nabla_\theta \log p_\theta(Z)^\top]$. Therefore,

$$\sqrt{n}\tilde{S}_n(\theta_\star) \to \mathcal{N}_d(0, H(\theta_\star)^{-1/2}G(\theta_\star)H(\theta_\star)^{-1/2}).$$

Our non-asymptotic theory. With high probability,

$$n\left\|\tilde{S}_n(\theta_\star)\right\|^2 = S_n(\theta_\star)^\top H_n(\theta_\star)^{-1}S_n(\theta_\star) \lesssim d_\star$$

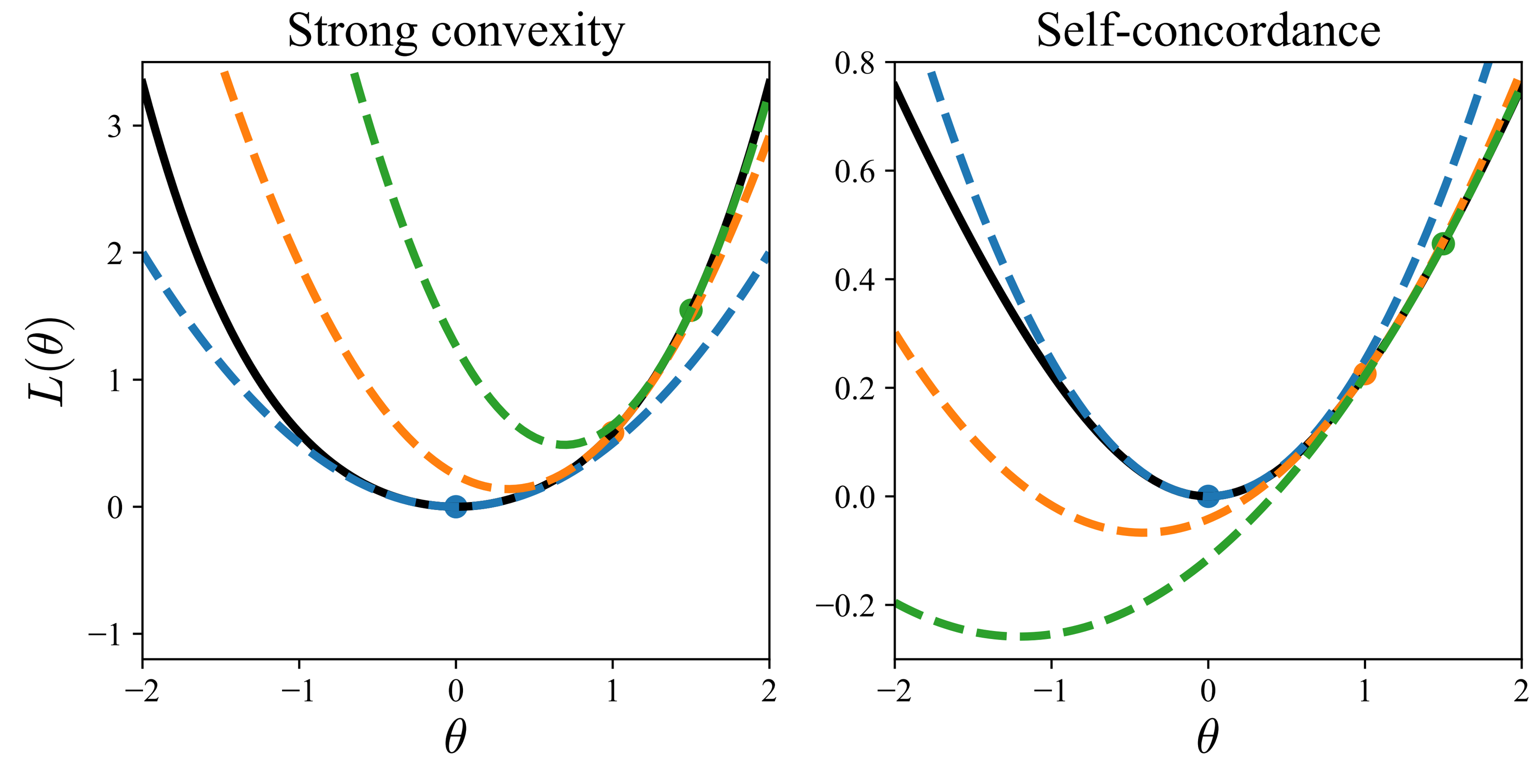
whenever $n \gtrsim \log d$, where $d_\star$ is the **effective dimension** given by

$$d_\star := \mathbf{Tr}\left(H(\theta_\star)^{-1/2}G(\theta_\star)H(\theta_\star)^{-1/2}\right).$$

- **Well-specified** model $\to d_\star = d$.
- **Mis-specified** model $\to d_\star$ may be much smaller than d.

Self-Concordance

	Strong convexity	Self-concordance
Hessian lower bound	Global	Local
Hessian varying rate	No control	Slow



Main Results

Estimation bound. With high probability,

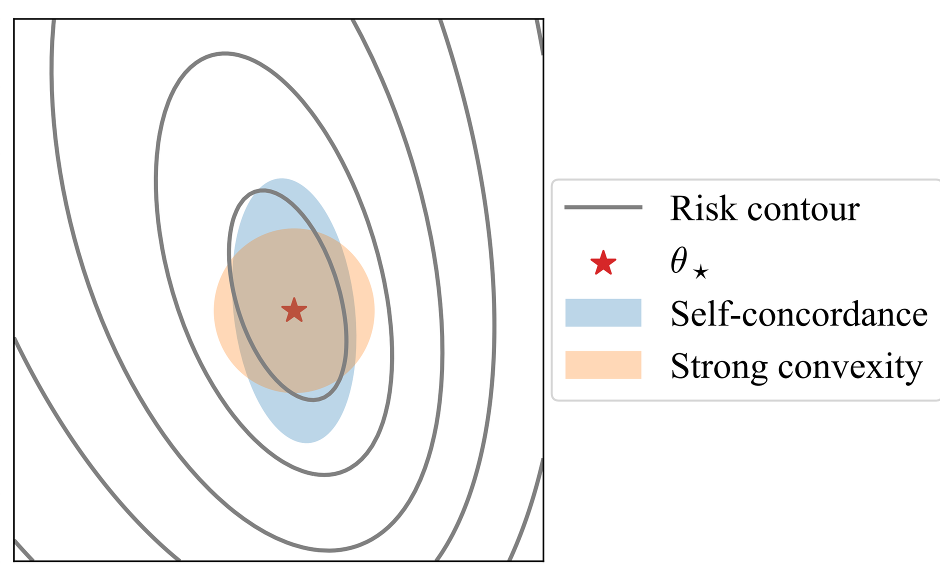
$$n\left\|H_\star^{1/2}(\theta_n - \theta_\star)\right\|^2 \lesssim d_\star, \quad \text{whenever } n \gtrsim d + d_\star.$$

- Asymptotic theory $n\|H_\star^{1/2}(\theta_n - \theta_\star)\|^2 \to_d \chi^2_{d_\star}$.
- Characterize the **critical sample size**.

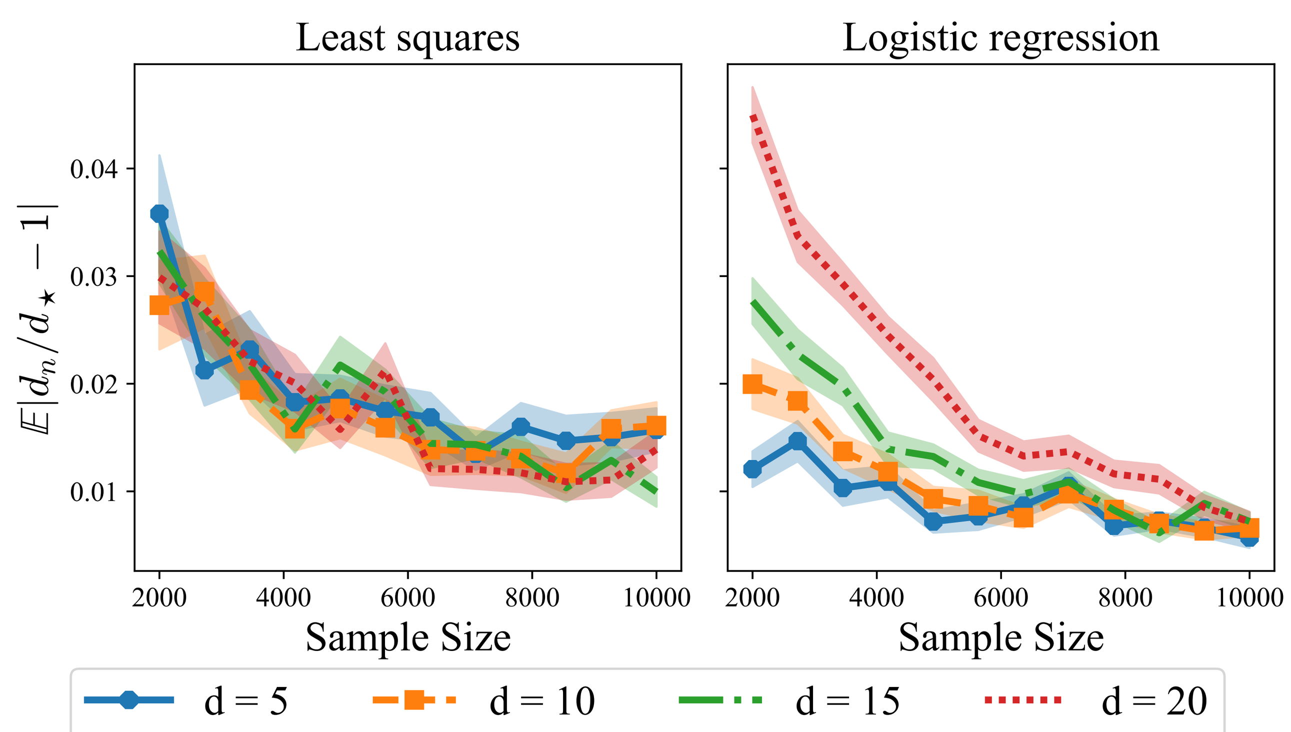
Confidence bound. With high probability,

$$n\left\|H_n(\theta_n)^{1/2}(\theta_n - \theta_\star)\right\|^2 \lesssim d_n, \quad \text{whenever } n \gtrsim d\log n + d_\star.$$

- Approximate $H(\theta_\star)$ and $G(\theta_\star)$ by $H_n(\theta_n)$ and $G_n(\theta_n)$.



How well does d_n approximate $d_\star$?



Goodness of fit testing. Assume a well-specified model $\mathbb{P} = P_{\theta_\star}$.

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

Rao's **score statistic** $T_n := \|H_n(\theta_0)^{-1/2}S_n(\theta_0)\|^2$.

- If $\theta_\star = \theta_0$ then $T_n = O(d/n) \to$ **critical value** $t_n = O(d/n)$.
- If $\theta_\star = \theta_0 + \omega(n^{-1/2})$ then **asymptotic power one**.
- If $\theta_\star = \theta_0 + O(n^{-1/2})$ then **asymptotically bounded power**.

Presented at NeurIPS 2022 Workshop on Score-Based Methods.