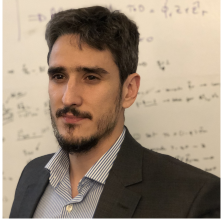
Orthogonal Statistical Learning with Self-Concordant Loss

Lang Liu, Carlos Cinelli, Zaid Harchaoui

University of Washington

Team

Lang Liu



Carlos Cinelli

Zaid Harchaoui

Motivating Example: Average Treatment Effect

Average Treatment Effect (ATE)

- **Data**: $D \in \{0, 1\}$ treatment, $X \in \mathbb{R}^p$ features, $Y \in \mathbb{R}$ outcome.
- **ATE**: $\theta_0 := \mathbb{E}\left[\mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X]\right]$.
- **Nuisance**: $g_{0,k} : X \mapsto \mathbb{E}[Y \mid D = k, X]$ for $k \in \{0, 1\}$.

Motivating Example: Average Treatment Effect

Average Treatment Effect (ATE)

- **Data**: $D \in \{0, 1\}$ treatment, $X \in \mathbb{R}^p$ features, $Y \in \mathbb{R}$ outcome.
- **ATE**: $\theta_0 := \mathbb{E}\left[\mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X]\right]$.
- **Nuisance**: $g_{0,k} : X \mapsto \mathbb{E}[Y \mid D = k, X]$ for $k \in \{0, 1\}$.
- **Challenge**: existence of a high (possibly infinite) dimensional nuisance.
- **Remedy**: orthogonal statistical learning and double/debiased machine learning, e.g.,
  - Chernozhukov et al. ’18
  - Mackey et al. ’18
  - Foster and Syrgkanis ’20

Orthogonal Statistical Learning

Orthogonal statistical learning (OSL)

- **Data**: $\mathcal{D} := \{Z_1, \ldots, Z_{2n}\}$ i.i.d. sample from $\mathbb{P}$.
- **Target parameter**: $\theta \in \Theta \subset \mathbb{R}^d$.
- **Nuisance**: $g \in (\mathcal{G}, \|\cdot\|_{\mathcal{G}})$
- **Loss**: $\ell_z : \Theta \times \mathcal{G} \to \mathbb{R}_+$.
- **Risk**: $L(\theta, g) := \mathbb{E}_{Z \sim \mathbb{P}}[\ell_Z(\theta, g)]$.
- **Goal**: assuming a true nuisance g_0, want to estimate

$$\theta_\star := \arg\min_{\theta \in \Theta} L(\theta, g_0).$$

Orthogonal Statistical Learning

OSL meta-algorithm

- **Sample splitting**: $\mathcal{D}_1 := \{Z_1, \ldots, Z_n\}$ and $\mathcal{D}_2 := \{Z_{n+1}, \ldots, Z_{2n}\}$.
- **Nuisance parameter**: outputs $\hat{g}$ based on $\mathcal{D}_2$.
- **Target parameter**: outputs $\hat{\theta}$ by minimizing

$$\min_{\theta \in \Theta} L_n(\theta, \hat{g}) := \frac{1}{n} \sum_{i=1}^{n} \ell_{Z_i}(\theta, \hat{g}).$$

- **Excess risk**: $\mathcal{E}(\hat{\theta}, g_0) := L(\hat{\theta}, g_0) - L(\theta_\star, g_0)$.

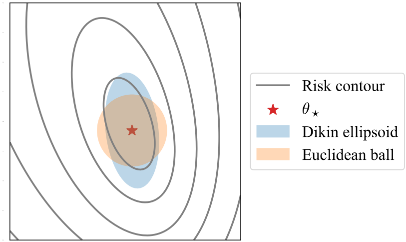
Localization and Dikin Ellipsoid

Assumption (Localization)

There exists $N > 0$ such that for all $n > N$, we have $\hat{\theta} \in \Theta_{\theta_\star}$ and $\hat{g} \in \mathcal{G}_{g_0}$.

Dikin ellipsoid

- **Hessian**: $H(\theta, g) := \nabla^2_\theta L(\theta, g)$ and $H_\star := H(\theta_\star, g_0)$.
- **Dikin ellipsoid**: $\Theta_{\theta_\star, r} := \{\theta \in \Theta : \|\theta - \theta_\star\|_{H_\star} := \|H_\star^{1/2}(\theta - \theta_\star)\|_2 < r\}$.



Effective Dimension

Effective dimension

- **Score**: $S_z(\theta, g) := \nabla_\theta \ell_z(\theta, g)$ and $S(\theta, g) := \mathbb{E}[S_Z(\theta, g)] = \nabla_\theta L(\theta, g)$.
- **Covariance**: $\Sigma(\theta, g) := \mathrm{Cov}(S_Z(\theta, g))$ and $\Sigma_\star := \Sigma(\theta_\star, g_0)$.
- **Effective dimension**: $d_\star := \sup_{g \in \mathcal{G}_{g_0}} \mathrm{Tr}(H_\star^{-1/2} \Sigma(\theta_\star, g) H_\star^{-1/2})$.

Effective Dimension

Effective dimension

- **Score**: $S_z(\theta, g) := \nabla_\theta \ell_z(\theta, g)$ and $S(\theta, g) := \mathbb{E}[S_Z(\theta, g)] = \nabla_\theta L(\theta, g)$.
- **Covariance**: $\Sigma(\theta, g) := \mathrm{Cov}(S_Z(\theta, g))$ and $\Sigma_\star := \Sigma(\theta_\star, g_0)$.
- **Effective dimension**: $d_\star := \sup_{g \in \mathcal{G}_{g_0}} \mathrm{Tr}(H_\star^{-1/2} \Sigma(\theta_\star, g) H_\star^{-1/2})$.
  - Well-specified model—$d_\star = d$.
  - Mis-specified model—problem-specific characterization of the complexity of Θ.
  - E.g., Huber '67, Ostrovskii and Bach '21.

Main Result

Theorem (Informal)

Under suitable assumptions, the OSL estimator $\hat{\theta}$ *has excess risk, with probability at least* $1 - \delta$,

$$\mathcal{E}(\hat{\theta}, g_0) \lesssim \frac{e^R}{\kappa^2} \left[K_1^2 \log(1/\delta) \frac{d_\star}{n} + \beta_2^2 \|\hat{g} - g_0\|_{\mathcal{G}}^4 \right]$$

whenever $n \gtrsim \max\{N, (K_2^2 + \sigma_H^2) d^2\}$.

Main Result

Theorem (Informal)

Under suitable assumptions, the OSL estimator $\hat{\theta}$ *has excess risk, with probability at least* $1-\delta$,

$$\mathcal{E}(\hat{\theta}, g_0) \lesssim \frac{e^R}{\kappa^2}\left[K_1^2 \log(1/\delta)\frac{d_\star}{n} + \beta_2^2 \|\hat{g} - g_0\|_{\mathcal{G}}^4\right]$$

whenever $n \gtrsim \max\{N, (K_2^2 + \sigma_H^2)d^2\}$.

Remark

Foster and Syrgkanis (2020) obtained the rate, with $\lambda_\star := \inf_\theta \lambda_{\min}(H(\theta, g_0))$,

$$O\left(\frac{d}{\lambda_\star^2}\frac{d}{n} + \frac{d}{\lambda_\star^2}\|\hat{g} - g_0\|_{\mathcal{G}}^4\right).$$

Main Result

Theorem (Simplified)

Under suitable assumptions, the OSL estimator $\hat{\theta}$ has excess risk, with probability at least $1-\delta$,

$$\mathcal{E}(\hat{\theta}, g_0) \lesssim O\left(\frac{1}{\lambda_\star}\frac{d_\star}{n} + \frac{1}{\lambda_\star}\|\hat{g} - g_0\|_{\mathcal{G}}^4\right)$$

whenever $n \gtrsim \max\{N, (K_2^2 + \sigma_H^2)d^2\}$.

Remark

Foster and Syrgkanis (2020) obtained the rate, with $\lambda_\star := \inf_\theta \lambda_{\min}(H(\theta, g_0))$,

$$O\left(\frac{d}{\lambda_\star^2}\frac{d}{n} + \frac{d}{\lambda_\star^2}\|\hat{g} - g_0\|_{\mathcal{G}}^4\right).$$

Main Result

Table: In their simplified version, our bound scales as $O(d_\star/n)$ and Foster and Syrgkanis's bound scales as $O(d'/n)$ where $d' := d^2/\lambda_\star$. We compare them in different regimes of eigendecays.

	Eigendecay $\Sigma_\star$	Eigendecay $H_\star$	Ratio $d'/d_\star$
Poly-Poly	$i^{-\alpha}$	$i^{-\beta}$	$d^{(\alpha+1)\wedge(\beta+2)}$
Poly-Exp	$i^{-\alpha}$	$e^{-\nu i}$	$d^{1\wedge(3-\alpha)}$
Exp-Poly	$e^{-\mu i}$	$i^{-\beta}$	$d^{\beta+2}$
Exp-Exp	$e^{-\mu i}$	$e^{-\nu i}$	$de^{\nu d}$ if $\mu = \nu$ $d^2 e^{\nu d}$ if $\mu > \nu$ $d^2 e^{\mu d}$ if $\mu < \nu$

Proof Sketch

By Taylor's theorem,

$$\mathcal{E}(\hat{\theta}, g_0) = L(\hat{\theta}, g_0) - L(\theta_\star, g_0) = S(\theta_\star, g_0)^\top(\hat{\theta} - \theta_\star) + \|\hat{\theta} - \theta_\star\|^2_{H(\bar{\theta}, g_0)}/2 \lesssim \|\hat{\theta} - \theta_\star\|^2_{H_\star}.$$

Proof Sketch

By Taylor's theorem,

$$\mathcal{E}(\hat\theta, g_0) = L(\hat\theta, g_0) - L(\theta_\star, g_0) = S(\theta_\star, g_0)^\top(\hat\theta - \theta_\star) + \|\hat\theta - \theta_\star\|^2_{H(\bar\theta, g_0)}/2 \lesssim \|\hat\theta - \theta_\star\|^2_{H_\star}.$$

By Taylor's theorem again,

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

It follows that

$$\mathcal{E}(\hat\theta, g_0) \lesssim \|\hat\theta - \theta_\star\|^2_{H_\star} \lesssim \frac{d_\star}{n} + \|\hat g - g_0\|^4_{\mathcal{G}}.$$

Proof Sketch

By Taylor's theorem,

$$\mathcal{E}(\hat{\theta}, g_0) := L(\hat{\theta}, g_0) - L(\theta_\star, g_0) = S(\theta_\star, g_0)^\top (\hat{\theta} - \theta_\star) + \|\hat{\theta} - \theta_\star\|^2_{H(\bar{\theta}, g_0)}/2 \lesssim \|\hat{\theta} - \theta_\star\|^2_{H_\star}.$$

By Taylor's theorem again,

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

Missing steps

- Control $S_n(\theta_\star, g)$ for every $g \in \mathcal{G}_{g_0}$.
- Relate $H_n(\theta, g)$ to $H(\theta, g)$ and then to $H(\theta_\star, g_0)$ for every $(\theta, g) \in \Theta_{\theta_\star} \times \mathcal{G}_{g_0}$.

Assumptions

Step 1: Relate $S_n(\theta_\star, g)$ to $S(\theta_\star, g)$ and then to $S(\theta_\star, g_0) = 0$.

- Sub-Gaussian score.
- Neyman orthogonal score.

Assumptions

Step 1: Relate $S_n(\theta_\star, g)$ to $S(\theta_\star, g)$ and then to $S(\theta_\star, g_0) = 0$.

- Sub-Gaussian score.
- Neyman orthogonal score.

Step 2: Relate $H_n(\theta, g)$ to $H(\theta, g)$ and then to $H(\theta_\star, g_0)$.

- Matrix Bernstein.
- Pseudo self-concordance.

Assumptions

Step 1: Relate $S_n(\theta_\star, g)$ to $S(\theta_\star, g)$ and then to $S(\theta_\star, g_0) = 0$.

- Sub-Gaussian score.
- Neyman orthogonal score.

Step 2: Relate $H_n(\theta, g)$ to $H(\theta, g)$ and then to $H(\theta_\star, g_0)$.

- Matrix Bernstein.
- Pseudo self-concordance.

Theorem (Informal)

Under assumptions above, with probability at least $1 - \delta$,

$$\mathcal{E}(\hat{\theta}, g_0) \lesssim \frac{e^R}{\kappa^2} \left[K_1^2 \log(1/\delta) \frac{d_\star}{n} + \beta_2^2 \|\hat{g} - g_0\|_{\mathcal{G}}^4 \right]$$

whenever $n \gtrsim \max\{N, (K_2^2 + \sigma_H^2) d^2\}$.

Summary

- Novel non-asymptotic bound for the OSL estimator.
- Assume pseudo self-concordance rather than strong convexity.
- The bound depends on the effective dimension instead of d.
- It improves previous work at least by a factor of d.

Paper