Near-Optimal Model Discrimination

arxiv.org/abs/2012.02901

Dmitrii M. Ostrovskii Mohamed Ndaoud
Adel Javanmard Meisam Razaviyayn

University of Southern California

Southern California Probability Symposium
December 11, 2021

Outline

- General problem formulation
- Linear models
- Extensions

General problem

Model discrimination task

- Let $z \in \mathcal{Z}$ be a random observation distributed according to $\mathbb{P}_0$ or $\mathbb{P}_1$.

- Let $\theta_0, \theta_1 \in \mathbb{R}^d$ be the **best-fit models** of z according to $\mathbb{P}_0, \mathbb{P}_1$, i.e.

$$\theta_k = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ L_k(\theta) := \mathbb{E}_{z \sim \mathbb{P}_k} \ell(\theta, z) \right\},$$

with strictly convex loss $\ell(\cdot, z) : \mathbb{R}^d \to \mathbb{R}$, population risks $L_0(\cdot), L_1(\cdot)$.

- **Statistician** has access to $\theta^* \in \{\theta_0, \theta_1\}$ (but **not** to $\bar{\theta} \in \{\theta_0, \theta_1\} \setminus \theta^*$), knows $\ell(\cdot, z)$, and observes **two** i.i.d. samples:

$$Z^0 = (z_1^0, ..., z_n^0) \sim \mathbb{P}_0^{\otimes n}, \quad Z^1 = (z_1^1, ..., z_n^1) \sim \mathbb{P}_1^{\otimes n}.$$

- **Task**: distinguish between the two hypotheses

$$\mathcal{H}_0 : \{\theta^* = \theta_0\}, \quad \mathcal{H}_1 : \{\theta^* = \theta_1\}.$$

Model discrimination task

- **Classical setup:** both θ_0, θ_1 known; one sample $Z \sim \mathbb{P}_\theta^{\otimes n}$ observed.

  Which $\theta \in \{\theta_0, \theta_1\}$ corresponds to the sample?
  Two simple hypotheses about θ.

- **Our setup:** we observe both samples but only one model $\theta^* \in \{\theta_0, \theta_1\}$.

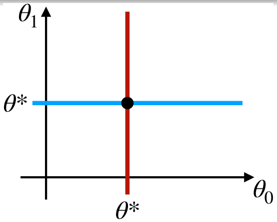
  Which $Z \in \{Z^0, Z^1\}$ corresponds to θ^*?
  Two composite hypotheses about (θ_0, θ_1).

- **Statistician** has access to $\theta^* \in \{\theta_0, \theta_1\}$ (but **not** to $\bar{\theta} \in \{\theta_0, \theta_1\} \setminus \theta^*$), knows $\ell(\cdot, z)$, and observes **two** i.i.d. samples:

$$Z^0 = (z_1^0, ..., z_n^0) \sim \mathbb{P}_0^{\otimes n}, \quad Z^1 = (z_1^1, ..., z_n^1) \sim \mathbb{P}_1^{\otimes n}.$$

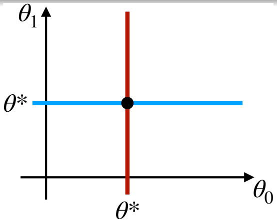
- **Task**: distinguish between the two hypotheses about $(\theta_0, \theta_1) \in \mathbb{R}^{2d}$:

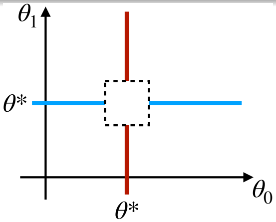
$$\mathcal{H}_0 : (\theta_0, \theta_1) \in \{\theta^*\} \times \{\theta \neq \theta^*\} \quad \text{vs.} \quad \mathcal{H}_1 : (\theta_0, \theta_1) \in \{\theta \neq \theta^*\} \times \{\theta^*\}$$



$$\mathcal{H}_0 : (\theta_0, \theta_1) \in \{\theta^*\} \times \{\theta \neq \theta^*\} \quad \text{vs.} \quad \mathcal{H}_1 : (\theta_0, \theta_1) \in \{\theta \neq \theta^*\} \times \{\theta^*\}$$

Separation and sample complexity





$$\mathcal{H}_0 : (\theta_0, \theta_1) \in \{\theta^*\} \times \{\theta \neq \theta^*\} \quad \text{vs.} \quad \mathcal{H}_1 : (\theta_0, \theta_1) \in \{\theta \neq \theta^*\} \times \{\theta^*\}$$

- **Separate** θ_0 and θ_1 to exclude the degenerate case $\theta_0 = \theta_1 = \theta^*$.

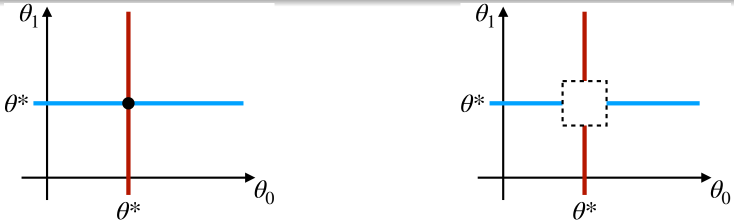
$$\mathcal{H}_0 : (\theta_0, \theta_1) \in \{\theta^*\} \times \bar{\Theta}_0 \quad \text{vs.} \quad \mathcal{H}_1 : (\theta_0, \theta_1) \in \bar{\Theta}_1 \times \{\theta^*\}$$

- "Prediction-wise" separation:

$$\Delta_0 := L_0(\theta_1) - L_0(\theta_0) > 0, \quad \Delta_1 := L_1(\theta_0) - L_1(\theta_1) > 0.$$

- Implicitly choose $\bar{\Theta}_0, \bar{\Theta}_1$ according to this separation assumption.

Separation and sample complexity



$$\mathcal{H}_0 : (\theta_0, \theta_1) \in \{\theta^*\} \times \{\theta \neq \theta^*\} \quad \text{vs.} \quad \mathcal{H}_1 : (\theta_0, \theta_1) \in \{\theta \neq \theta^*\} \times \{\theta^*\}$$

- **Separate** θ_0 and θ_1 to exclude the degenerate case $\theta_0 = \theta_1 = \theta^*$.

$$\mathcal{H}_0 : (\theta_0, \theta_1) \in \{\theta^*\} \times \bar{\Theta}_0 \quad \text{vs.} \quad \mathcal{H}_1 : (\theta_0, \theta_1) \in \bar{\Theta}_1 \times \{\theta^*\}$$

- "Prediction-wise" separation:

$$\Delta_0 := L_0(\theta_1) - L_0(\theta_0) > 0, \quad \Delta_1 := L_1(\theta_0) - L_1(\theta_1) > 0.$$

- Implicitly choose $\bar{\Theta}_0, \bar{\Theta}_1$ according to this separation assumption.

Characterize the **sample complexity** of distinguishing between $\mathcal{H}_0$ and $\mathcal{H}_1$ with fixed error probabilities of both types (say $2/3$) in terms of Δ_0, Δ_1, d.

Well-specified linear models

Linear model setup

Well-specified linear model: $z = (x, y) \in \mathbb{R}^{d+1}$, $\ell(\theta, z) = \frac{1}{2}(x^\top \theta - y)^2$, and

$$\mathbb{P}_k : \; x \sim \mathcal{N}(0, \boldsymbol{\Sigma}_k), \;\; y = x^\top \theta_k + \xi \;\text{ with } \xi \sim \mathcal{N}(0,1) \;\text{ for }\; k \in \{0,1\}.$$

- Write $Z_k = (X_k; Y_k)$, where $X_k \in \mathbb{R}^{n \times d}$ and $Y_k \in \mathbb{R}^n$ for $k \in \{0,1\}$.
- Covariances $\boldsymbol{\Sigma}_k$ and their estimates: $\widehat{\boldsymbol{\Sigma}}_k := \frac{1}{n} X_k^\top X_k$.
- Population and empirical ranks: $r_k = \operatorname{rank}(\boldsymbol{\Sigma}_k)$ and $\widehat{r}_k = \operatorname{rank}(\widehat{\boldsymbol{\Sigma}}_k)$.
- Separations and their empirical counterparts:

$$\Delta_k = \tfrac{1}{2}\|\theta_1 - \theta_0\|_{\boldsymbol{\Sigma}_k}^2 = \tfrac{1}{2}\|\boldsymbol{\Sigma}_k^{1/2}(\theta_1 - \theta_0)\|^2,$$
$$\widehat{\Delta}_k = \tfrac{1}{2}\|\theta_1 - \theta_0\|_{\widehat{\boldsymbol{\Sigma}}_k}^2 = \tfrac{1}{2n}\|X_k(\theta_1 - \theta_0)\|^2.$$

Basic test based on the prediction error of θ^* under $\mathcal{H}_0$ and $\mathcal{H}_1$:

$$\mathbb{1}\left\{\|Y_0 - X_0\theta^*\|^2 - n \geqslant \|Y_1 - X_1\theta^*\|^2 - n\right\}.$$

Let $\xi_k = Y_k - X_k\theta_k \sim \mathcal{N}(0, \boldsymbol{I}_n)$ be the noises. Under $\mathcal{H}_0 : \theta^* = \theta_0$, one has

$$\begin{aligned}\text{LHS} &= \|\xi_0\|^2 - n,\\ \text{RHS} &= \|\xi_1\|^2 - n - 2\langle \xi_1, X_1(\theta_0 - \theta_1)\rangle + \|X_1(\theta_1 - \theta_0)\|^2.\end{aligned}$$

- Thus, $\mathbb{E}[\text{LHS}] = 0$ and $\mathbb{E}[\text{RHS}|X_1] = \|X_1(\theta_1 - \theta_0)\|^2 = n\widehat{\Delta}_1$, where

  $$\widehat{\Delta}_1 = \tfrac{1}{n}\|X_1(\theta_0 - \theta_1)\|^2 = \|\theta_0 - \theta_1\|^2_{\widehat{\boldsymbol{\Sigma}}_1}$$

  is the empirical counterpart of $\Delta_1 = \|\theta_1 - \theta_0\|^2_{\boldsymbol{\Sigma}_1}$.
- This motivates the basic test: type-I error $\iff$ "fluctuations $\geqslant n\Delta_1$."

Basic test: analysis

$$\mathbb{1}\left\{\|Y_0 - X_0\theta^*\|^2 - n \geqslant \|Y_1 - X_1\theta^*\|^2 - n\right\}.$$

More precisely, LHS $\sim \chi_n^2 - n$ and RHS$|X_1 \sim \chi_n^2 - n + 2\mathcal{N}(0, n\widehat{\Delta}_1) + n\widehat{\Delta}_1$.

- Basic tail inequalities for Gaussian and χ^2 laws:

  $\mathbb{P}[\mathcal{N}(0,1) \geqslant u] \leqslant \exp(-u^2)$, $\quad \mathbb{P}[|\chi_s^2 - s| \geqslant v] \lesssim \exp(-c\min\{v, v^2/s\})$.

- Bound for the (conditional over X_0, X_1) type-I error:

$$\begin{aligned}\mathbb{P}_I &= \mathbb{P}[\text{fluctuations} \geqslant n\widehat{\Delta}_1] \\ &\leqslant \mathbb{P}\left[\chi_n^2 - n \geqslant \frac{n\widehat{\Delta}_1}{3}\right] + \mathbb{P}\left[n - \chi_n^2 \geqslant \frac{n\widehat{\Delta}_1}{3}\right] + \mathbb{P}\left[\mathcal{N}(0, n\widehat{\Delta}_1) \geqslant \frac{n\widehat{\Delta}_1}{6}\right] \\ &\lesssim \exp\left(-\frac{cn^{\cancel{2}}\widehat{\Delta}_1^2}{\cancel{n}}\right) + \exp(-cn\widehat{\Delta}_1).\end{aligned}$$

- Thus, error prob. of both types at most $\exp(-cn\min\{\Delta, \Delta^2\})$, where

$$\Delta := \min\{\Delta_0, \Delta_1\}.$$

If $\Delta \lesssim 1$: term $\exp(-cn\Delta^2)$ dominates $\Rightarrow$ $O(1/\Delta^2)$ sample complexity.

Improved test

Idea: reduce χ^2 fluctuations by projecting the residuals on signal spaces.

Test for linear model

$$\widehat{T} = \mathbb{1}\left\{\|\mathbf{\Pi}_{X_0}[Y_0 - X_0\theta^*]\|^2 - \widehat{r}_0 \geqslant \|\mathbf{\Pi}_{X_1}[Y_1 - X_1\theta^*]\|^2 - \widehat{r}_1\right\},$$

where $\mathbf{\Pi}_X := X(X^\top X)^\dagger X^\top$ is the projector on signal space $\mathrm{col}(X) \subseteq \mathbb{R}^n$.

- Recall that $\widehat{r}_k := \mathrm{rank}(\widehat{\mathbf{\Sigma}}_k)$ and $\widehat{\mathbf{\Sigma}} = \frac{1}{n}X^\top X$, hence indeed

  $\dim(\mathrm{col}(X)) = \mathrm{Tr}(\mathbf{\Pi}_X) = \mathrm{Tr}[(X^\top X)^\dagger X^\top X] = \mathrm{rank}(X^\top X) = \mathrm{rank}(\widehat{\mathbf{\Sigma}})$.

Improved test: analysis

Test for linear model

$$\widehat{T} = \mathbb{1}\left\{\|\mathbf{\Pi}_{X_0}[Y_0 - X_0\theta^*]\|^2 - \widehat{r}_0 \geqslant \|\mathbf{\Pi}_{X_1}[Y_1 - X_1\theta^*]\|^2 - \widehat{r}_1\right\},$$

where $\mathbf{\Pi}_X := X(X^\top X)^\dagger X^\top$ is the projector on signal space $\mathrm{col}(X) \subseteq \mathbb{R}^n$.

- For this test, under $\mathcal{H}_0$, we have

$$\mathrm{LHS}|X_0 \sim {\chi_{\widehat{r}_0}}^2 - \widehat{r}_0, \quad \mathrm{RHS}|X_1 \sim {\chi_{\widehat{r}_1}}^2 - \widehat{r}_1 + 2\mathcal{N}(0, n\widehat{\Delta}_1) + n\widehat{\Delta}_1.$$

- Smaller χ^2 fluctuations since $\widehat{r}_k \overset{a.s.}{=} \min\{r_k, n\} \leqslant n$. Type-I error prob.:

$$\mathbb{P}\left[{\chi_{\widehat{r}_0}}^2 - \widehat{r}_0 \geqslant \frac{n\widehat{\Delta}_1}{3}\right] + \mathbb{P}\left[\widehat{r}_1 - {\chi_{\widehat{r}_1}}^2 \geqslant \frac{n\widehat{\Delta}_1}{3}\right] + \mathbb{P}\left[\mathcal{N}(0, n\widehat{\Delta}_1) \geqslant \frac{n\widehat{\Delta}_1}{6}\right]$$

$$\lesssim \exp\left(-\frac{cn^2\widehat{\Delta}_1^2}{\widehat{r}_0}\right) + \exp\left(-\frac{cn^2\widehat{\Delta}_1^2}{\widehat{r}_1}\right) + \exp(-cn\widehat{\Delta}_1).$$

Theorem. Denoting $r_{\max} := \max\{r_0, r_1\}$, we have

$$\max\{P_I, P_{II}\} \lesssim \exp\left(-c\min\left\{n\Delta, \frac{n^2\Delta^2}{\min\{n, r_{\max}\}}\right\}\right).$$

Error probability bound

Theorem. Denoting $r_{\max} := \max\{r_0, r_1\}$, we have

$$\max\{P_I, P_{II}\} = \exp\left(-c \min\left\{n\Delta, \frac{n^2\Delta^2}{\min\{n, r_{\max}\}}\right\}\right).$$

Improved test: sample complexity

Error probability bound

Theorem. Denoting $r_{\max} := \max\{r_0, r_1\}$, we have

$$\max\{P_I, P_{II}\} = \exp\left(-c \min\left\{n\Delta, \frac{n^2\Delta^2}{\min\{n, r_{\max}\}}\right\}\right).$$

Sample complexity bound

Lemma Assume $\Delta \lesssim 1$. Then $\log(\max\{P_I, P_{II}\}) \lesssim -1$ is equivalent to

$$\boxed{n \gtrsim \min\left\{\frac{1}{\Delta^2}, \frac{\sqrt{r_{\max}}}{\Delta}\right\}.}$$

Error probability bound

Theorem. Denoting $r_{\max} := \max\{r_0, r_1\}$, we have

$$\max\{P_I, P_{II}\} = \exp\left(-c \min\left\{n\Delta, \frac{n^2\Delta^2}{\min\{n, r_{\max}\}}\right\}\right).$$

Sample complexity bound

Lemma Assume $\Delta \lesssim 1$. Then $\log(\max\{P_I, P_{II}\}) \lesssim -1$ is equivalent to

$$\boxed{n \gtrsim \min\left\{\frac{1}{\Delta^2}, \frac{\sqrt{r_{\max}}}{\Delta}\right\}.}$$

Proof:

1. Prove: $n\Delta \gtrsim \min\left\{\frac{1}{\Delta}, \sqrt{r_{\max}}\right\} \iff n\Delta \min\left\{1, \frac{n\Delta}{\min\{n, r_{\max}\}}\right\} \gtrsim 1$ if $\Delta \lesssim 1$.
2. The second condition reads $n\Delta \gtrsim \max\left\{1, \min\left\{\frac{1}{\Delta}, \frac{r_{\max}}{n\Delta}\right\}\right\}$, or equivalently $n\Delta \gtrsim \min\left\{\frac{1}{\Delta}, \max\left\{1, \frac{r_{\max}}{n\Delta}\right\}\right\}$ by using $\Delta \lesssim 1$ and treating all possible cases.
3. It remains to verify that $n\Delta \gtrsim \sqrt{r_{\max}}$ if and only if $n\Delta \gtrsim \max\left\{1, \frac{r_{\max}}{n\Delta}\right\}$. $\square$

Comparison

Basic test

$$\mathbb{1}\left\{\|Y_0 - X_0\theta^*\|^2 - n \geqslant \|Y_1 - X_1\theta^*\|^2 - n\right\},$$

$$\textbf{Sample complexity:}\ \ n = O\left(\frac{1}{\Delta^2}\right).$$

Improved test

$$\mathbb{1}\left\{\|\mathbf{\Pi}_{X_0}[Y_0 - X_0\theta^*]\|^2 - \widehat{r}_0 \geqslant \|\mathbf{\Pi}_{X_1}[Y_1 - X_1\theta^*]\|^2 - \widehat{r}_1\right\}.$$

$$\textbf{Sample complexity:}\ \ n = O\left(\min\left\{\frac{1}{\Delta^2}, \frac{\sqrt{r_{\max}}}{\Delta}\right\}\right).$$

Note: $\widehat{r}_k \overset{a.s.}{=} \min\{r_k, n\}$ and $\mathbf{\Pi}_{X_k}$ projects on $\mathrm{col}(X_k) \subset \mathbb{R}^n$ of dimension $\widehat{r}_k$. Thus, the tests coincide when $n \leqslant \min\{r_0, r_1\}$. In fact, a **phase transition:**

- **Well-separated:** $\Delta \gtrsim \frac{1}{\sqrt{r_{\max}}}$. Sample complexity $n = O(1/\Delta^2) \lesssim r_{\max}$.
- **Ill-separated:** $\Delta \ll \frac{1}{\sqrt{r_{\max}}}$. Sample complexity $\gg r_{\max} \Rightarrow$ projections.

Interpretation via least-squares

Recall the normal equations for the least-squares estimates $\widehat{\theta}_0, \widehat{\theta}_1$ of θ_0, θ_1:

$$\widehat{\boldsymbol{\Sigma}}_0 \widehat{\theta}_0 = \tfrac{1}{n} X_0^\top Y_0, \quad \widehat{\boldsymbol{\Sigma}}_1 \widehat{\theta}_1 = \tfrac{1}{n} X_1^\top Y_1.$$

This allows to rewrite the squared norms of the projected residuals:

$$\begin{aligned}
\|\boldsymbol{\Pi}_X [Y - X\theta^*]\|^2 &= (Y - X\theta^*)^\top \boldsymbol{\Pi}_X (Y - X\theta^*) \\
&= (X^\top Y - X^\top X\theta^*)^\top (X^\top X)^\dagger (X^\top Y - X^\top X\theta^*) \\
&= n^2 (\widehat{\boldsymbol{\Sigma}}(\widehat{\theta} - \theta^*))^\top (X^\top X)^\dagger \widehat{\boldsymbol{\Sigma}}(\widehat{\theta} - \theta^*) \\
&= n(\widehat{\theta} - \theta^*)^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Sigma}}^\dagger \widehat{\boldsymbol{\Sigma}} (\widehat{\theta} - \theta^*) = n(\widehat{\theta} - \theta^*)^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\theta} - \theta^*) \\
&= n \|\widehat{\theta} - \theta^*\|_{\widehat{\boldsymbol{\Sigma}}}^2.
\end{aligned}$$

Thus, our test amounts to $\mathbb{1}\left\{ \|\theta^* - \widehat{\theta}_0\|_{\widehat{\boldsymbol{\Sigma}}_0}^2 - \frac{\widehat{r}_0}{n} \geqslant \|\theta^* - \widehat{\theta}_1\|_{\widehat{\boldsymbol{\Sigma}}_1}^2 - \frac{\widehat{r}_1}{n} \right\}$.

- We compare the empirical prediction distances from $\widehat{\theta}^*$ to $\widehat{\theta}_0$ and $\widehat{\theta}_1$ *after debiasing them under the matching hypothesis*.
- **NB:** *we don't require* $\widehat{\theta}_0, \widehat{\theta}_1$ *to be unique* (i.e. $n \geqslant r_{\max}$).

Model discrimination vs. recovery

Sample complexity for improved test: $O\left(\min\left\{\frac{1}{\Delta^2}, \frac{\sqrt{r_{\max}}}{\Delta}\right\}\right) \ll \frac{r_{\max}}{\Delta}$.

- Sample complexity of **estimating** $\bar{\theta} = \theta_0 + \theta_1 - \theta^*$ up to Δ prediction error (i.e., better than by θ^*) is at least $\frac{r_{\min}}{\Delta}$.
- Thus, when $r_0 \approx r_1$, **recovery is way harder than discrimination!**

Non-disclosure property

We can **discriminate** *between $\mathcal{H}_0$ and $\mathcal{H}_1$ with sample size that does not allow to* **recover** *the complementary model $\bar{\theta}$ (with better quality than θ^*).*

- In fact, our tests access θ^* through "scalar" statistic $\|\Pi_X[Y - X\theta^*]\|^2$ that carries only $O(1)$ Fisher information about θ^*.
- Hence, we also guarantee **non-disclosure of** θ^* (up to accuracy Δ).

We need to prove two bounds:

$$\inf_{\widehat{T}} \sup_{\|\theta_1-\theta_0\|_{I_r}^2 \geqslant \Delta} P_I(\widehat{T})+P_{II}(\widehat{T}) \gtrsim \max\left\{\exp(-cn\Delta), \;\; \exp\left(-c\frac{n^2\Delta^2}{\min\{n,r\}}\right)\right\}.$$

First bound: easier problem with **known** $\bar{\theta}$ and **simple hypotheses:**

$$\widetilde{\mathcal{H}}_0 : (\theta_0, \theta_1) = (\theta^*, \bar{\theta}), \quad \text{vs.} \quad \widetilde{\mathcal{H}}_1 : (\theta_0, \theta_1) = (\bar{\theta}, \theta^*).$$

Likelihood-ratio (LR) test

$$T_{\text{LR}} = \mathbb{1}\{\|Y_0 - X_0\theta^*\|^2 + \|Y_1 - X_1\bar{\theta}\|^2 \geqslant \|Y_0 - X_0\bar{\theta}\|^2 + \|Y_1 - X_1\theta^*\|^2\}$$

is optimal (w.r.t. sum of errors) by the Neyman-Pearson lemma, and for it

$$\begin{aligned}
&\mathbb{P}_{\widetilde{\mathcal{H}}_0}[T_{\text{LR}} = 1 | X_0, X_1] \\
&= \mathbb{P}\left[\|Y_0 - X_0\theta_0\|^2 + \|Y_1 - X_1\theta_1\|^2 \geqslant \|Y_0 - X_0\theta_1\|^2 + \|Y_1 - X_1\theta_0\|^2 \,\middle|\, X_0, X_1\right] \\
&= \mathbb{P}\left[2\langle\xi_0, X_0(\theta_0 - \theta_1)\rangle + 2\langle\xi_1, X_1(\theta_0 - \theta_1)\rangle \geqslant \|X_0(\theta_0 - \theta_1)\|^2 + \|X_1(\theta_0 - \theta_1)\|^2\right] \\
&\geqslant \mathbb{P}\left[2\mathcal{N}(0, n\widehat{\Delta}_0) + 2\mathcal{N}(0, n\widehat{\Delta}_1) \geqslant n\widehat{\Delta}_0 + n\widehat{\Delta}_1\right] \\
&\geqslant \mathbb{P}\left[\mathcal{N}(0, n\widehat{\Delta}_0) \geqslant n\widehat{\Delta}_0/2\right] \cdot \mathbb{P}[\mathcal{N}(0, n\widehat{\Delta}_1) \geqslant n\widehat{\Delta}_1/2] \gtrsim \exp\left(-cn\max\{\widehat{\Delta}_0, \widehat{\Delta}_1\}\right).
\end{aligned}$$

Then $\max\{\widehat{\Delta}_0, \widehat{\Delta}_1\} \lesssim \Delta$ with fixed probability by Markov's inequality.

Lower bound: key ideas

We need to prove two bounds:

$$\inf_{\widehat{T}} \sup_{\|\theta_1-\theta_0\|^2_{\boldsymbol{I}_r} \geqslant \Delta} P_I(\widehat{T}) + P_{II}(\widehat{T}) \gtrsim \max\left\{\exp(-cn\Delta),\ \ \exp\left(-c\frac{n^2\Delta^2}{\min\{n,r\}}\right)\right\}.$$

Second bound captures dependence on the rank. Bayesian approach:

- Put a Gaussian prior π on $\bar{\theta}$ such that $\pi\{\|\bar{\theta}-\theta^*\|^2_{\boldsymbol{I}_r} \leqslant \Delta\}$ is very small.
- This allows to lower-bound the maximal risk by the Bayes risk.
- The Bayes risk can be lower-bounded by the Neyman-Pearson lemma —through surprisingly tedious calculations. $\square$

Beyond linear models

In our paper:

- General result for parametric models in asymptotic regime $n \to \infty$ with fixed r_0, r_1 and $n\Delta \to \lambda$.
- Technical result for generalized linear models (GLMs) allowing for heavy tails and misspecification.
- **Same general picture:**

$$\max\{P_I, P_{II}\} \asymp \exp\left(-c \min\left\{n\Delta, \frac{n^2\Delta^2}{\max\{\rho_0, \rho_1\}}\right\}\right)$$

where ρ_0, ρ_1 are "effective model ranks".

Open questions:

- Closing the gap for linear models
- General nonasymptotic result
- Mixtures
- New insights on two-sample testing?

Thank you!

And check our paper:

arxiv.org/abs/2012.02901

General asymptotics

Linear model: $\mathbb{1}\left\{\|\mathbf{\Pi}_{X_0}[Y_0 - X_0\theta^*]\|^2 - \widehat{r}_0 \geqslant \|\mathbf{\Pi}_{X_1}[Y_1 - X_1\theta^*]\|^2 - \widehat{r}_1\right\}.$

General setup:

- Empirical risk $\widehat{L}_k(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, z_i^{(k)})$ has gradient $\nabla\widehat{L}_k(\theta)$ and Hessian $\widehat{\boldsymbol{H}}_k(\theta)$:

$$\widehat{\boldsymbol{H}}_k(\theta) := \nabla^2\widehat{L}_k(\theta), \quad \boldsymbol{H}_k(\theta) := \nabla^2 L_k(\theta).$$

- Let $\boldsymbol{G}_k(\theta) := \mathrm{Cov}_{\mathbb{P}_k}[\nabla\ell_z(\theta)]$. **For well-specified models:**

$$\boldsymbol{G}_k(\theta_k) = \boldsymbol{H}_k(\theta_k).$$

- Standardized Fisher matrix: $\boldsymbol{J}_k(\theta) := \boldsymbol{H}_k(\theta)^{-\dagger/2}\boldsymbol{G}_k(\theta)\boldsymbol{H}_k(\theta)^{-\dagger/2}$.
- Effective rank $\rho_k := \mathrm{Tr}[\boldsymbol{J}_k(\theta_k)]$. **For well-specified models:** $\rho_k = r_k$.

In linear regression $\nabla\widehat{L}(\theta) = \frac{1}{n}X^\top(Y - X\theta)$ and $\widehat{\boldsymbol{H}}(\theta) \equiv \frac{1}{n}X^\top X$, hence

$$\|\mathbf{\Pi}_X[Y - X\theta^*]\|^2 = \|(X^\top X)^{\dagger/2}X^\top(Y - X\theta^*)\|^2 = n\|\widehat{\boldsymbol{H}}(\theta^*)^{\dagger/2}\nabla\widehat{L}(\theta^*)\|^2.$$

$$\mathbb{1}\left\{\|\boldsymbol{\Pi}_{X_0}[Y_0 - X_0\theta^*]\|^2 - \widehat{r}_0 \geqslant \|\boldsymbol{\Pi}_{X_1}[Y_1 - X_1\theta^*]\|^2 - \widehat{r}_1\right\}.$$

- Replace $\|\boldsymbol{\Pi}_{X_k}[Y_k - X_k\theta^*]\|^2$ with $n\|\widehat{\boldsymbol{H}}_k(\theta_k)^{\dagger/2}\nabla\widehat{L}_k(\theta^*)\|^2$.
- When $n \to \infty$,

$$\mathbb{E}_k[n\|\widehat{\boldsymbol{H}}_k(\theta_k)^{\dagger/2}\nabla\widehat{L}_k(\theta_k)\|^2] \to \rho_k = \mathrm{Tr}[\boldsymbol{J}_k(\theta_k)].$$

- Cannot use ρ_k's as one of them uses $\bar{\theta}$ which is unknown. Instead use

$$\mathrm{Tr}[\boldsymbol{J}_k(\theta^*)] = n_k\mathbb{E}_k\left[\left\|\boldsymbol{H}_k(\theta^*)^{\dagger/2}\left(\nabla\widehat{L}_k(\theta^*) - \nabla L_k(\theta^*)\right)\right\|^2\right],$$

or, more precisely, its asymptotically (as $n \to \infty$) unbiased estimate:

$$\widehat{\mathrm{T}}_k = \tfrac{1}{2}n_k\left\|\boldsymbol{H}_k(\theta^*)^{\dagger/2}\left(\nabla\widehat{L}_k(\theta^*) - \widehat{\nabla}L'_k(\theta^*)\right)\right\|^2.$$

$$\widehat{T} = \mathbb{1}\left\{n_0\|\widehat{\boldsymbol{H}}_0(\theta^*)^{\dagger/2}\nabla\widehat{L}_0(\theta^*)\|^2 - \widehat{\mathrm{T}}_0 \geqslant n_1\|\widehat{\boldsymbol{H}}_1(\theta^*)^{\dagger/2}\nabla\widehat{L}_1(\theta^*)\|^2 - \widehat{\mathrm{T}}_1\right\}.$$

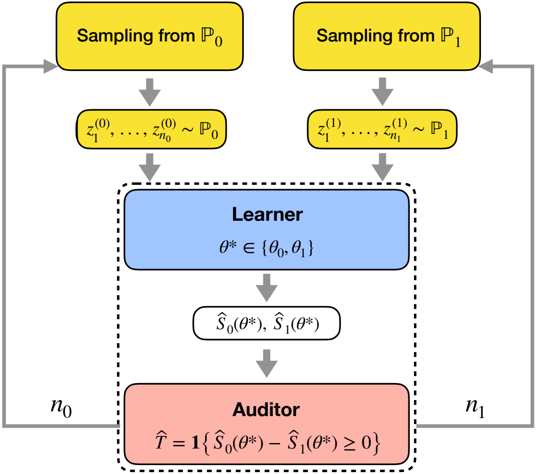
Theorem. Denoting $\rho_{max} := \max\{\rho_0, \rho_1\}$, we have that

$$\lim_{n\to\infty}[\max\{P_I, P_{II}\}] \lesssim \exp\left(-c\min\left\{n\Delta, \frac{n^2\Delta^2}{\rho_{\max}}\right\}\right).$$

Applications

Applications: generic testing protocol

Key observation: θ^* *does not have to be known to run the test, and it cannot be inferred from* Z_0, Z_1 *when they are small.*



- We want to protect θ^* and $\overline{\theta} = \theta_0 + \theta_1 - \theta^*$ from inference via Z_0, Z_1.

Application #1: testing for data deletion

Testing for data deletion

- Company FAANG[1] trained a prediction model θ^* on a large dataset $\mathbb{P}^*$ pertaining to many users.
- Some users ask their data to be removed—and θ^* retrained accordingly.
- FAANG should comply—and would like to demonstrate the compliance.
- Model θ^* is proprietary, hence FAANG would like to avoid disclosing it.

Given a subsample $Z^* \sim \mathbb{P}^*$ of FAANG's dataset and the pool Q of deletion queries, we can check that FAANG indeed retrained the model excluding Q.

- Let $\mathbb{P}_0, \mathbb{P}_1$ correspond to hypotheses $\mathcal{H}_0 : \mathbb{P}^* = \mathbb{P}_0$ ("clean data") and

$$\mathcal{H}_1 : \mathbb{P}^* = \mathbb{P}_1 := (1-\delta)\mathbb{P}_0 + \delta Q,$$

  where $\mathbb{P}_0$ is "clean" data, and $\delta \in (0,1)$ is the share of deletion queries.
- FAANG ("Learner") gives to the tester ("Auditor") access to $Z_0 \sim \mathbb{P}_0$.
- Having Z_0, Q, and δ, Auditor can generate $Z_1 = (1-\delta)Z_0 + \delta Q \sim \mathbb{P}_1$.

Application #2: testing fair representation of subpopulations

- Let $\mathbb{P}_{\text{dem}}$ and $\mathbb{P}_{\text{rep}}$ be two populations: Democrats and Republicans.
- We want them to be equally represented in the dataset $\mathbb{P}^*$.

- Define the hypotheses:

$$\begin{aligned}
&\mathcal{H}_0 : \mathbb{P}^* = \mathbb{P}_0 := \tfrac{1}{2}\mathbb{P}_{\text{dem}} + \tfrac{1}{2}\mathbb{P}_{\text{rep}},\\
&\mathcal{H}_1 : \mathbb{P}^* = \mathbb{P}_1 := (\tfrac{1}{2} + \delta)\mathbb{P}_{\text{dem}} + (\tfrac{1}{2} - \delta)\mathbb{P}_{\text{rep}} \ \text{ for some } \ \delta \in (-\tfrac{1}{2}, \tfrac{1}{2}).
\end{aligned}$$

- Knowing δ, we can easily implement the sampling oracles for $\mathbb{P}_0$ and $\mathbb{P}_1$ can be implemented using those for $\mathbb{P}_{\text{dem}}$ and $\mathbb{P}_{\text{rep}}$.