Near-Optimal Model Discrimination

https://arxiv.org/abs/2012.02901

Dmitrii M. Ostrovskii **Mohamed Ndaoud**
Adel Javanmard Meisam Razaviyayn

University of Southern California

WIAS, Berlin
January 22, 2021

Problem formulation

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.,

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

where $\ell_z(\theta)$ is the loss function, $L_k(\theta)$ the population risks ($k \in \{0, 1\}$).
The loss function $\ell_z : \mathbb{R}^d \to \mathbb{R}$ is known (and assumed strictly convex).

- **Statistician** has access to $\theta^* \in \{\theta_0, \theta_1\}$ (but **not** to $\bar{\theta} \in \{\theta_0, \theta_1\} \setminus \theta^*$) 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 testing focuses on the **sample**. We focus on the **model**.

- **Classical testing:** both θ_0 and θ_1 are known; one observes $Z \sim \mathbb{P}^{\otimes n}$.

  Which $\theta \in \{\theta_0, \theta_1\}$ *corresponds to the sample?*

  Two simple *hypotheses about the unknown* θ.

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

  Which of the two samples Z^0, Z^1 *corresponds to* θ^**?*

  Two composite *hypotheses about the unknown* (θ_0, θ_1).

- **Statistician** has access to $\theta^* \in \{\theta_0, \theta_1\}$ (but **not** to $\bar{\theta} \in \{\theta_0, \theta_1\} \setminus \theta^*$) 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 $(\theta_0, \theta_1) \in \mathbb{R}^{2d}$:

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

Separation and sample complexity

$$\mathcal{H}_0 : (\theta_0, \theta_1) \in ({\color{red}\theta^*}, \bar{\Theta}_0) \text{ vs. } \mathcal{H}_1 : (\theta_0, \theta_1) \in (\bar{\Theta}_1, {\color{red}\theta^*}) \;\text{ for some } \bar{\Theta}_0, \bar{\Theta}_1.$$

What are $\bar{\Theta}_0, \bar{\Theta}_1$?

- $\mathbb{R}^d$ not an option: then $\mathcal{H}_0$ and $\mathcal{H}_1$ have the common point (θ^*, θ^*).
- Thus we have to *separate* $\bar{\Theta}_0, \bar{\Theta}_1$ from θ^*.
- Assume that θ_0 and θ_1 are **separated** "prediction-wise":

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

(We can explicitly write $\bar{\Theta}_0, \bar{\Theta}_1$ that correspond to this prior information – but we won't.)

Main question

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 $\Delta_0, \Delta_1, \ldots$

Well-specified linear regression

Linear regression setup

Consider the linear regression setup: $z = (x, y)$, and $\mathbb{P}_0, \mathbb{P}_1$ are given by

$$\mathbb{P}_k : \; x \sim \mathcal{N}(0, \boldsymbol{\Sigma}_k), \;\; y|x \sim \mathcal{N}(x^\top \theta_k, \sigma_k^2) \;\text{ for }\; k \in \{0, 1\}.$$

Moreover, let $\sigma_0^2 = \sigma_1^2 = 1$ and denote $r_k = \operatorname{rank}(\boldsymbol{\Sigma}_k)$.

- Write $Z^k = (X^k; Y^k)$, where $X^k \in \mathbb{R}^{n \times d}$ and $Y^k \in \mathbb{R}^n$ for $k \in \{0, 1\}$.
- Note that $\widehat{\boldsymbol{\Sigma}}_k := \frac{1}{n} X^{k\top} X^k$ is an estimate of $\boldsymbol{\Sigma}_k$.
- Separations given by $\Delta_k = \|\theta_1 - \theta_0\|_{\boldsymbol{\Sigma}_k}^2$ and have empirical counterparts

$$\widehat{\Delta}_k = \|\theta_1 - \theta_0\|_{\widehat{\boldsymbol{\Sigma}}_k}^2 = \frac{1}{n}\|X^k(\theta_1 - \theta_0)\|^2.$$

Basic test

Consider 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 $\mathcal{H}_0 : \theta^* = \theta_0$, we have

$$\begin{aligned}
\text{LHS} &= \|\xi^0\|^2 - n, \\
\text{RHS} &= \|\xi^1\|^2 - n - 2\left\langle \xi^1, X_1(\theta_0 - \theta_1)\right\rangle + \|X_1(\theta_1 - \theta_0)\|^2.
\end{aligned}$$

- 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

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

$$\mathbb{1}\left\{\|Y^0 - X^0\theta^*\|^2 - n \geqslant \|Y^1 - X^1\theta^*\|^2 - n\right\}.$$

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

- Recalling the concentration inequalities

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

  (see [LM00]), we bound the (conditional over X_0, X_1) type-I error prob.:

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

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

- 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: decrease χ^2-term fluctuations by projecting residuals on signal spaces.

Test for linear model

$$\widehat{T} = \mathbb{1}\left\{\|\boldsymbol{\Pi}_{X^0}[Y^0 - X^0\theta^*]\|^2 - \widehat{r}_0 \geqslant \|\boldsymbol{\Pi}_{X^1}[Y^1 - X^1\theta^*]\|^2 - \widehat{r}_1\right\},$$

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

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

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

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 $\max\{P_I, P_{II}\} \leqslant \bar{P}$ with

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

Improved test: sample complexity

Theorem. Denoting $r_{\max} := \max\{r_0, r_1\}$, we have $\max\{P_I, P_{II}\} \leqslant \bar{P}$ with

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

Lemma Assume $\Delta \lesssim 1$. Then $-\log(\bar{P}) \gtrsim 1$ is equivalent to

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

Proof. The above bound on n is equivalent to

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

On the other hand, $\bar{P} \lesssim 1$ reads $n\Delta \min\left\{1, \frac{n\Delta}{\min\{n, r_{\max}\}}\right\} \gtrsim 1$. Equivalently,

$$n\Delta \gtrsim \max\left\{1, \min\left\{\frac{1}{\Delta}, \frac{r_{\max}}{n\Delta}\right\}\right\} \iff n\Delta \gtrsim \min\left\{\frac{1}{\Delta}, \max\left\{1, \frac{r_{\max}}{n\Delta}\right\}\right\},$$

where the last step uses $\Delta \lesssim 1$. Now, the first cases under minimum are identical, and the second cases are equivalent: $n\Delta \gtrsim \sqrt{r_{\max}} \iff n\Delta \geqslant \max\left\{1, \frac{r_{\max}}{n\Delta}\right\}$. □

Comparison

Basic test: $\mathbb{1}\left\{\|Y^0 - X^0\theta^*\|^2 - n \geqslant \|Y^1 - X^1\theta^*\|^2 - n\right\}.$

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

Improved test: $\mathbb{1}\left\{\|\mathbf{\Pi}_{X^0}[Y^0 - X^0\theta^*]\|^2 - \widehat{r}_0 \geqslant \|\mathbf{\Pi}_{X^1}[Y^1 - X^1\theta^*]\|^2 - \widehat{r}_1\right\}.$

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

Note: $\widehat{r}_k \overset{a.s.}{=} \min\{r_k, n\}$ and $\mathbf{\Pi}_{X^k}$ projects on $\mathrm{col}(X^k) \subset \mathbb{R}^n$ with dim. $\widehat{r}_k$. Thus, when $n \leqslant \min\{r_0, r_1\}$, the two tests coincide.

- **Well-sep. regime:**

$$\Delta \gtrsim \frac{1}{\sqrt{r_{\max}}}.$$

  Samp. comp. $\lesssim r_{\max}$ and rank-indep. No need for projections if $r_0 \asymp r_1$.
- **Ill-sep. regime:** $\Delta \ll \frac{1}{\sqrt{r_{\max}}}$, samp. comp. $\gg r_{\max}$, need 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:

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

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

- 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. $n \geqslant r_{\max}$).

Testing vs. estimation

Improved test: $\mathbb{1}\left\{n\|\theta^* - \widehat{\theta}_0\|^2_{\widehat{\boldsymbol{\Sigma}}_0} - \widehat{r}_0 \geqslant n\|\theta^* - \widehat{\theta}_1\|^2_{\widehat{\boldsymbol{\Sigma}}_1} - \widehat{r}_1\right\}.$

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

Testing vs. estimation

Improved test: $\mathbb{1}\left\{n\|\theta^* - \widehat{\theta}_0\|^2_{\widehat{\boldsymbol{\Sigma}}_0} - \widehat{r}_0 \geqslant n\|\theta^* - \widehat{\theta}_1\|^2_{\widehat{\boldsymbol{\Sigma}}_1} - \widehat{r}_1\right\}.$

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

- 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 $\frac{r_{\min}}{\Delta}$ $\left[\approx \frac{r_{\max}}{\Delta} \text{ when } r_0 \asymp r_1\right]$.

Non-disclosure

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

- Rich potential for applications in “privacy-aware ML” (see our paper).

Improved test has sample complexity (whenever $\min\{\Delta_0, \Delta_1\} \geqslant \Delta$):

$$O\left(\min\left\{\frac{1}{\Delta^2}, \frac{\sqrt{r_{\max}}}{\Delta}\right\}\right).$$

Near-optimal – up to replacing $r_{\max}$ with $r_{\min}$ and min. sep. with max. sep.

Theorem. Let $r_0, r_1 \in \mathbb{N}$ and $d \geqslant r_{\max}$ be arbitrary. Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two distributions (depending on θ_0, θ_1) in the form

$$\mathbb{P}_k : x \sim \mathbb{D}_k, \; y|x \sim \mathcal{N}(x^\top \theta_k, 1),$$

with $\mathbb{D}_0$, $\mathbb{D}_1$ supported on $\mathbb{R}^d$ and having zero mean and covariances $\boldsymbol{I}_{r_0}, \boldsymbol{I}_{r_1}$. Then $\mathbb{D}_0$ and $\mathbb{D}_1$ can be chosen (depending only on r_0 and r_1) such that:

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

where inf is over all measurable maps $\widehat{T} : (\theta^*, X^0, Y^0, X^1, Y^1) \to \{0, 1\}$.

Lower bound: key ideas

We need to prove two bounds:

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

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

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

Likelihood-ratio test

$$T_{\mathrm{LR}} = \mathbb{1}\{\|Y^0 - X^0\theta^*\|^2 + \|Y^1 - X^1\bar{\theta}\|^2 \geqslant \|Y^0 - X^0\bar{\theta}\|^2 + \|Y^1 - X^1\theta^*\|^2\}$$

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

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_0^o}[T_{\mathrm{LR}} = 1 | X^0, X^1] \\
&= \mathbb{P}\left[\|Y^0 - X^0\theta_0\|^2 + \|Y^1 - X^1\theta_1\|^2 \geqslant \|Y^0 - X^0\theta_1\|^2 + \|Y^1 - X^1\theta_0\|^2 \middle| X^0, X^1\right] \\
&= \mathbb{P}\left[2\langle \xi^0, X^0(\theta_0 - \theta_1)\rangle + 2\langle \xi^1, X^1(\theta_0 - \theta_1)\rangle \geqslant \|X^0(\theta_0 - \theta_1)\|^2 + \|X^1(\theta_0 - \theta_1)\|^2 \middle| X^0, X^1\right] \\
&\geqslant \mathbb{P}\left[2\mathcal{N}(0, n\widehat{\Delta}_0) + 2\mathcal{N}(0, n\widehat{\Delta}_1) \geqslant n\widehat{\Delta}_0 + n\widehat{\Delta}_1\right] \\
&\geqslant \mathbb{P}\left[\mathcal{N}(0, n\widehat{\Delta}_0) \geqslant n\widehat{\Delta}_0/2\right] \cdot \mathbb{P}[\mathcal{N}(0, n\widehat{\Delta}_1) \geqslant n\widehat{\Delta}_1/2] \\
&\gtrsim \exp\left(-cn\max\{\widehat{\Delta}_0, \widehat{\Delta}_1\}\right).
\end{aligned}$$

Then $\widehat{\Delta}_k \lesssim \Delta_k$ with probability $O(1)$ by Markov's inequality.

We need to prove two bounds:

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

Second bound captures dependence on the ranks. Proof is technical.

- Fixing θ^*, put a (conditional) Gaussian prior on $\bar{\theta}$ with covariance

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 $\widehat{L}_k(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_{z_i^k}(\theta)$ 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 $\nabla^2\widehat{L}(\theta) \equiv \frac{1}{n}X^\top X$, hence

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

- Replace $\|\mathbf{\Pi}_{X^k}[Y^k - X^k\theta^*]\|^2$ with the Newton decrement for $\widehat{L}_k(\theta^*)$.

$$\mathbb{1}\left\{\|\boldsymbol{\Pi}_{X^0}[Y^0 - X^0\theta^*]\|^2 - \widehat{r}_0 \geqslant \|\boldsymbol{\Pi}_{X^1}[Y^1 - X^1\theta^*]\|^2 - \widehat{r}_1\right\}.$$

- Replace $\|\boldsymbol{\Pi}_{X^k}[Y^k - X^k\theta^*]\|^2$ with the Newton decrement for $\widehat{L}_k(\theta^*)$.
- When $n \geqslant r_k$, $\widehat{r}_k \overset{a.s.}{=} r_k$. We could replace $r_k = \rho_k = \mathrm{Tr}[\boldsymbol{J}_k(\theta_k)]$, but we only have access to θ^*. So we use

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

instead. more precisely, its asymptotically ($n \to \infty$) unbiased estimate:

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

This leads to the test

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

Theorem. Denoting $\rho_{max} := \max\{\rho_0, \rho_1\}$, $\lim_{n\to\infty}[\max\{P_I, P_{II}\}] \leqslant \bar{P}$ with

$$\bar{P} = \exp\left(-c\min\left\{n\Delta, \frac{n^2\Delta^2}{\rho_{\max}}\right\}\right).$$

References

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.