Near-Optimal Model Discrimination

arxiv.org/abs/2012.02901

Dmitrii M. Ostrovskii Mohamed Ndaoud
Adel Javanmard Meisam Razaviyayn

University of Southern California

USC Probability and Statistics Seminar
February 5, 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.

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

with strictly convex loss $\ell_z : \mathbb{R}^d \to \mathbb{R}$ and 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 ℓ_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 ℓ_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 $(\theta_0, \theta_1) \in \mathbb{R}^{2d}$:

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

Separation and sample complexity

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

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^*.
- "Prediction-wise" separation:

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

Possible to recast this information in terms of $\bar{\Theta}_0, \bar{\Theta}_1$, but hardly useful.

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 models

Linear regression setup

Well-specified linear regression: $z = (x, y)$, and $\mathbb{P}_k$, $k \in \{0, 1\}$, is given by

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

- 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: $r_k = \text{rank}(\boldsymbol{\Sigma}_k)$, and $\widehat{r}_k = \text{rank}(\widehat{\boldsymbol{\Sigma}}_k)$.
- Separations and their empirical counterparts:

$$\begin{aligned}
\Delta_k &= \|\theta_1 - \theta_0\|^2_{\boldsymbol{\Sigma}_k} = \|\boldsymbol{\Sigma}_k^{1/2}(\theta_1 - \theta_0)\|^2 \\
\widehat{\Delta}_k &= \|\theta_1 - \theta_0\|^2_{\widehat{\boldsymbol{\Sigma}}_k} = \tfrac{1}{n}\|X^k(\theta_1 - \theta_0)\|^2.
\end{aligned}$$

Basic test

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

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

- 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

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

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

- Recall tail inequalities: $\mathbb{P}[\mathcal{N}(0,1) \geqslant t] \leqslant \exp(-t^2)$ and χ^2-bound:

  $$\mathbb{P}[|\chi^2_s - s| \geqslant t] \lesssim \exp(-c\min\{t, t^2/s\}),$$

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

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

- 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.

Idea: decrease χ^2-term fluctuations by projecting 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

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

Sample complexity bound

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

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

Comparison

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

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

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

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

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

- **Well-sep. regime:** $\Delta \gtrsim \frac{1}{\sqrt{r_{\max}}}$. Sample complexity $\lesssim r_{\max}$ and rank-independent. No need for projections if $r_0 \asymp r_1$.
- **Ill-sep. regime:** $\Delta \ll \frac{1}{\sqrt{r_{\max}}}$, sample complexity $\gg r_{\max}$, projections.

Testing vs. estimation

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

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

Testing vs. estimation

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

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 complementary model $\bar{\theta}$ (with better quality than θ^*).*

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

Interpretation via least-squares

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

$$\widehat{\mathbf{\Sigma}}_0 \widehat{\theta}_0 = \tfrac{1}{n} X^{0\top} Y^0, \quad \widehat{\mathbf{\Sigma}}_1 \widehat{\theta}_1 = \tfrac{1}{n} X^{1\top} Y^1.$$

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

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

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

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

Lower bound: key ideas

We need to prove two bounds:

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

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

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

  - Likelihood-ratio (LR) 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.

- **Second bound** captures dependence on the rank. Bayesian approach:
  - Put Gaussian prior on $\bar{\theta}$, lower-bound $\max\{\mathbb{P}_I,\mathbb{P}_{II}\}$ for the Bayes test.
  - Lower-bounding is technical and requires that $\widehat{\boldsymbol{\Sigma}}_0, \widehat{\boldsymbol{\Sigma}}_1$ commute.
  - We achieve this by sampling x and $\widetilde{x}$ from $\sqrt{r}\{\pm e_1, ..., \pm e_r\}$.

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:**

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

where ρ_0, ρ_1 are "effective model ranks".

Open questions:

- General nonasymptotic result;
- Full optimality;
- Adaptation.

Thank you!

And check our paper:

arxiv.org/abs/2012.02901

General asymptotics

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

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

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

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

- Replace $\|\mathbf{\Pi}_{X^k}[Y^k - X^k\theta^*]\|^2$ 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

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