Minimax Risk Upper Bounds Based on Shell Analysis of a Quantized Maximum Likelihood Estimator

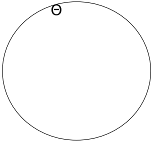
Or Ordentlich
Joint work with Noam Gavish

The Hebrew University of Jerusalem

Algorithmic Structures for Uncoordinated Communications and Statistical Inference in Exceedingly Large Spaces
BIRS, Banff
March 15, 2024

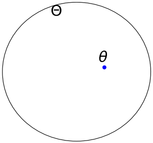
Problem Setting: High Dimensional Parameter Estimation

- Parameter space: $\Theta \subset \mathbb{R}^d$, loss $\ell : \Theta \times \Theta \to \mathbb{R}_+$



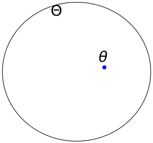
Problem Setting: High Dimensional Parameter Estimation

- Parameter space: $\Theta \subset \mathbb{R}^d$, loss $\ell : \Theta \times \Theta \to \mathbb{R}_+$
- Sample model: $\theta \in \Theta$, $Y^n \sim P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$



Problem Setting: High Dimensional Parameter Estimation

- Parameter space: $\Theta \subset \mathbb{R}^d$, loss $\ell : \Theta \times \Theta \to \mathbb{R}_+$
- Sample model: $\theta \in \Theta$, $Y^n \sim P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$
- Denote possible estimators by $\hat{\theta} : \mathcal{Y}^n \to \Theta$



Problem Setting: High Dimensional Parameter Estimation

- Parameter space: $\Theta \subset \mathbb{R}^d$, loss $\ell : \Theta \times \Theta \to \mathbb{R}_+$
- Sample model: $\theta \in \Theta$, $Y^n \sim P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$
- Denote possible estimators by $\hat{\theta} : \mathcal{Y}^n \to \Theta$

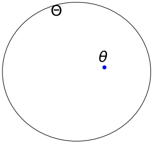
Objective

Upper bound the minimax risk

$$\min_{\hat{\theta}(\cdot)} \max_{\theta} \mathbb{E}_{Y^n \sim P_\theta} \left[\ell(\theta, \hat{\theta}(Y^n)) \right]$$

or its PAC proxy

$$\min_{\hat{\theta}(\cdot)} \max_{\theta} P_\theta \left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta \right]$$



Problem Setting: High Dimensional Parameter Estimation

- Parameter space: $\Theta \subset \mathbb{R}^d$, loss $\ell : \Theta \times \Theta \to \mathbb{R}_+$
- Sample model: $\theta \in \Theta$, $Y^n \sim P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$
- Denote possible estimators by $\hat{\theta} : \mathcal{Y}^n \to \Theta$

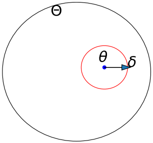
Objective

Upper bound the minimax risk

$$\min_{\hat{\theta}(\cdot)} \max_{\theta} \mathbb{E}_{Y^n \sim P_\theta} \left[\ell(\theta, \hat{\theta}(Y^n)) \right]$$

or its PAC proxy

$$\min_{\hat{\theta}(\cdot)} \max_{\theta} P_\theta \left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta \right]$$



Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include
  - Spiked Wigner ($Y = \lambda\theta\theta^T + \frac{1}{\sqrt{d}}Z$)

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include
  - Spiked Wigner ($Y = \lambda\theta\theta^T + \frac{1}{\sqrt{d}}Z$)
  - Multi Reference Alignment

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include
  - Spiked Wigner ($Y = \lambda\theta\theta^T + \frac{1}{\sqrt{d}}Z$)
  - Multi Reference Alignment
  - Gaussian Mixture Model

    $\vdots$

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include
  - Spiked Wigner ($Y = \lambda\theta\theta^T + \frac{1}{\sqrt{d}}Z$)
  - Multi Reference Alignment
  - Gaussian Mixture Model

    $\vdots$
  - General results for discrete product distributions (under "locality"), e.g., stochastic block model

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include
  - Spiked Wigner ($Y = \lambda\theta\theta^T + \frac{1}{\sqrt{d}}Z$)
  - Multi Reference Alignment
  - Gaussian Mixture Model

    $\vdots$
  - General results for discrete product distributions (under "locality"), e.g., stochastic block model
- Maximum Likelihood Estimator (MLE):

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include
  - Spiked Wigner ($Y = \lambda\theta\theta^T + \frac{1}{\sqrt{d}}Z$)
  - Multi Reference Alignment
  - Gaussian Mixture Model

    $\vdots$
  - General results for discrete product distributions (under "locality"), e.g., stochastic block model
- Maximum Likelihood Estimator (MLE):
  - Consistency, efficiency for fixed d, $n \to \infty$

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include
  - Spiked Wigner ($Y = \lambda\theta\theta^T + \frac{1}{\sqrt{d}}Z$)
  - Multi Reference Alignment
  - Gaussian Mixture Model

    ⋮
  - General results for discrete product distributions (under "locality"), e.g., stochastic block model
- Maximum Likelihood Estimator (MLE):
  - Consistency, efficiency for fixed d, $n \to \infty$
  - Often hard to analyze for involved models

Problem Setting: High Dimensional Parameter Estimation

- $d, n \to \infty$ simultaneously
- Interested in involved non i.i.d. models. Examples include
  - Spiked Wigner ($Y = \lambda\theta\theta^T + \frac{1}{\sqrt{d}}Z$)
  - Multi Reference Alignment
  - Gaussian Mixture Model

    $\vdots$
  - General results for discrete product distributions (under "locality"), e.g., stochastic block model
- Maximum Likelihood Estimator (MLE):
  - Consistency, efficiency for fixed d, $n \to \infty$
  - Often hard to analyze for involved models

Our objective

Develop a **unified** information-theoretic framework for **upper bounding**

$$\min_{\hat{\theta}(\cdot)} \max_{\theta} P_\theta \left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta \right]$$

Reminder: Unified IT framework for lower bounds

Deriving a general **lower bound** on minimax risk is easy

Mutual Information Metohd [Polyanskiy-Wu, Chapter 30]

Fix a prior $\theta \sim \pi$

$$\theta \overset{P_{Y^n|\theta}}{—} Y^n — \hat{\theta}$$

Examine:

$$I\left(\theta; \hat{\theta}(Y^n)\right)$$

Deriving a general **lower bound** on minimax risk is easy

Mutual Information Metohd [Polyanskiy-Wu, Chapter 30]

Fix a prior $\theta \sim \pi$

$$\theta \overset{P_{Y^n|\theta}}{—} Y^n — \hat{\theta}$$

Let $R(D)$ be the RDF for π, ℓ. By definition

$$R\left(\mathbb{E}\left[\ell(\theta, \hat{\theta})\right]\right) \leq I\left(\theta; \hat{\theta}(Y^n)\right)$$

Reminder: Unified IT framework for lower bounds

Deriving a general **lower bound** on minimax risk is easy

Mutual Information Metohd [Polyanskiy-Wu, Chapter 30]

Fix a prior $\theta \sim \pi$

$$\theta \overset{P_{Y^n|\theta}}{—} Y^n — \hat{\theta}$$

Data Processing Inequality gives

$$R\left(\mathbb{E}\left[\ell(\theta, \hat{\theta})\right]\right) \leq I\left(\theta; \hat{\theta}(Y^n)\right) \leq I\left(\theta; Y^n\right)$$

Reminder: Unified IT framework for lower bounds

Deriving a general **lower bound** on minimax risk is easy

Mutual Information Metohd [Polyanskiy-Wu, Chapter 30]

Fix a prior $\theta \sim \pi$

$$\theta \overset{P_{Y^n|\theta}}{—} Y^n — \hat{\theta}$$

By definition of capacity

$$R\left(\mathbb{E}\left[\ell(\theta,\hat{\theta})\right]\right) \leq I\left(\theta;\hat{\theta}(Y^n)\right) \leq I\left(\theta;Y^n\right) \leq C(P_{Y^n|\theta})$$

Reminder: Unified IT framework for lower bounds

Deriving a general **lower bound** on minimax risk is easy

Mutual Information Metohd [Polyanskiy-Wu, Chapter 30]

Fix a prior $\theta \sim \pi$

$$R\left(\mathbb{E}\left[\ell(\theta,\hat{\theta})\right]\right) \leq I\left(\theta;\hat{\theta}(Y^n)\right) \leq I\left(\theta;Y^n\right) \leq C(P_{Y^n|\theta})$$

- Tight in many cases (for "good" choice of π)
- Lower bound decouples analysis of Θ and sample model

Reminder: Unified IT framework for lower bounds

- Lower bound decouples analysis of Θ and sample model

$$R\left(\mathbb{E}\left[\ell(\theta,\hat{\theta})\right]\right) \leq I\left(\theta;\hat{\theta}(Y^n)\right) \leq I\left(\theta;Y^n\right) \leq C(P_{Y^n|\theta})$$

Reminder: Unified IT framework for lower bounds

- Lower bound decouples analysis of Θ and sample model

$$R\left(\mathbb{E}\left[\ell(\theta,\hat{\theta})\right]\right) \leq I\left(\theta;\hat{\theta}(Y^n)\right) \leq I\left(\theta;Y^n\right) \leq C(P_{Y^n|\theta})$$

- Decoupling has a price - the output of $P_{Y^n|\theta}$ can provide much information on θ that is not helpful for estimation under the $\ell(\cdot,\cdot)$ distortion measure
  $\Theta = \mathcal{S}^{d-1}$, $Y^n = \theta + \sigma Z$, $\ell(\theta,\hat{\theta}) = \sum_{i=1}^{d} |[\theta_i - \hat{\theta}_i] \bmod \epsilon|$

Reminder: Unified IT framework for lower bounds

- Lower bound decouples analysis of Θ and sample model

$$R\left(\mathbb{E}\left[\ell(\theta,\hat{\theta})\right]\right) \leq I\left(\theta;\hat{\theta}(Y^n)\right) \leq I\left(\theta;Y^n\right) \leq C(P_{Y^n|\theta})$$

- Decoupling has a price - the output of $P_{Y^n|\theta}$ can provide much information on θ that is not helpful for estimation under the $\ell(\cdot,\cdot)$ distortion measure
  $\Theta = \mathcal{S}^{d-1}$, $Y^n = \theta + \sigma Z$, $\ell(\theta,\hat{\theta}) = \sum_{i=1}^{d} |[\theta_i - \hat{\theta}_i] \bmod \epsilon|$
- Prior work on upper bounds in the spirit of separated analysis (Yang-Barron [1999], Birge [1983], Le Cam [1986], Yatracos [1985] ...) bypassed this by considering $\ell(\theta,\hat{\theta}) = D_f(P_\theta \| P_{\hat{\theta}})$

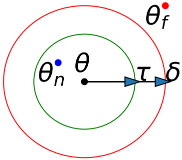
Reminder: Unified IT framework for lower bounds

- Lower bound decouples analysis of Θ and sample model

$$R\left(\mathbb{E}\left[\ell(\theta,\hat{\theta})\right]\right) \leq I\left(\theta;\hat{\theta}(Y^n)\right) \leq I\left(\theta;Y^n\right) \leq C(P_{Y^n|\theta})$$

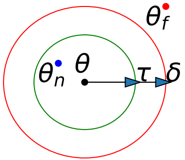
- Decoupling has a price - the output of $P_{Y^n|\theta}$ can provide much information on θ that is not helpful for estimation under the $\ell(\cdot,\cdot)$ distortion measure
  $\Theta = \mathcal{S}^{d-1}$, $Y^n = \theta + \sigma Z$, $\ell(\theta,\hat{\theta}) = \sum_{i=1}^{d} |[\theta_i - \hat{\theta}_i] \bmod \epsilon|$
- Prior work on upper bounds in the spirit of separated analysis (Yang-Barron [1999], Birge [1983], Le Cam [1986], Yatracos [1985] ...) bypassed this by considering $\ell(\theta,\hat{\theta}) = D_f(P_\theta \| P_{\hat{\theta}})$
- We characterize the "sensitivity" of P_θ to large ℓ-variations in θ via mismatched binary hypothesis testing

Main Result - Definitions



- Let $\tau > 0$
- $\Theta \subset \mathbb{R}^d$, $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some norm, $\mathrm{diam}(\Theta) = e^{e^{o(d)}}$
- $\theta, \theta_n, \theta_f \in \Theta$
  - $Y^n \sim P_\theta$
  - $\ell(\theta_n, \theta) < \tau < \delta < \ell(\theta_f, \theta)$
- Likelihood test θ_n VS θ_f

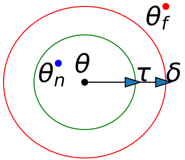
Main Result - Definitions



- Let $\tau > 0$
- $\Theta \subset \mathbb{R}^d$, $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some norm, $\mathrm{diam}(\Theta) = e^{e^{o(d)}}$
- $\theta, \theta_n, \theta_f \in \Theta$
    - $Y^n \sim P_\theta$
    - $\ell(\theta_n, \theta) < \tau < \delta < \ell(\theta_f, \theta)$
- Likelihood test θ_n VS θ_f

$$P_\theta\left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n) \geq 1\right]$$

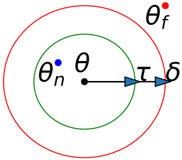
Main Result - Definitions



- Let $\tau > 0$
- $\Theta \subset \mathbb{R}^d$, $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some norm, $\mathrm{diam}(\Theta) = e^{e^{o(d)}}$
- $\theta, \theta_n, \theta_f \in \Theta$
  - $Y^n \sim P_\theta$
  - $\ell(\theta_n, \theta) < \tau < \delta < \ell(\theta_f, \theta)$
- Likelihood test θ_n VS θ_f

$$-\frac{1}{d} \log P_\theta \left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n) \geq 1 \right]$$

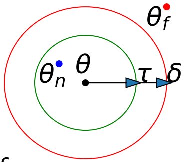
Main Result - Definitions



- Let $\tau > 0$
- $\Theta \subset \mathbb{R}^d$, $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some norm, $\mathrm{diam}(\Theta) = e^{e^{o(d)}}$
- $\theta, \theta_n, \theta_f \in \Theta$
  - $Y^n \sim P_\theta$
  - $\ell(\theta_n, \theta) < \tau < \delta < \ell(\theta_f, \theta)$
- Likelihood test θ_n VS θ_f
- Worst case error exponent w.r.t $\theta, \theta_n, \theta_f$

$$\min_{\substack{\theta,\theta_n,\theta_f \in \Theta \\ \ell(\theta,\theta_n) \leq \tau \\ \ell(\theta,\theta_f) \geq \delta}} -\frac{1}{d} \log P_\theta \left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n) \geq 1 \right]$$

Main Result - Definitions

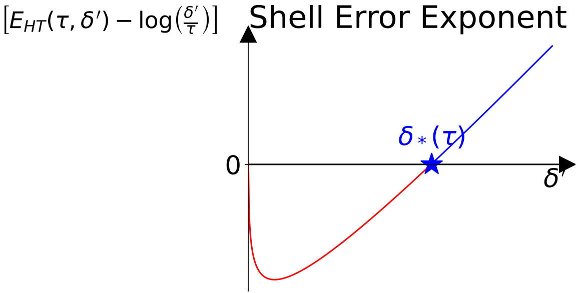


- Let $\tau > 0$
- $\Theta \subset \mathbb{R}^d$, $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some norm, $\mathrm{diam}(\Theta) = e^{e^{o(d)}}$
- $\theta, \theta_n, \theta_f \in \Theta$
  - $Y^n \sim P_\theta$
  - $\ell(\theta_n, \theta) < \tau < \delta < \ell(\theta_f, \theta)$
- Likelihood test θ_n VS θ_f
- Worst case error exponent w.r.t $\theta, \theta_n, \theta_f$

Def: Hypothesis Testing Error Exponent (Mismatched,Worst Case)

$$E_{HT}(\tau, \delta) \triangleq \min_{\substack{\theta,\theta_n,\theta_f \in \Theta \\ \ell(\theta,\theta_n) \leq \tau \\ \ell(\theta,\theta_f) \geq \delta}} -\frac{1}{d} \log P_\theta \left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n) \geq 1 \right]$$

Main Result - Definitions



Def: Critical Loss

$$\delta_*(\tau) \triangleq \sup\left\{\delta' > 0 : \left[E_{HT}(\tau,\delta') - \log\frac{\delta'}{\tau}\right] \leq 0\right\}$$

$$\delta_* \triangleq \min_{\tau>0} \delta_*(\tau)$$

Main Result

Def: Hypothesis Testing Error Exponent (Mismatched,Worst Case)

$$E_{HT}(\tau,\delta) \triangleq \min_{\substack{\theta,\theta_n,\theta_f\in\Theta \\ \ell(\theta,\theta_n)\leq\tau \\ \ell(\theta,\theta_f)\geq\delta}} -\frac{1}{d}\log P_\theta\left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n)\geq 1\right]$$

Def: Critical Loss

$$\delta_*(\tau) \triangleq \sup\left\{\delta' > 0 : \left[E_{HT}(\tau,\delta') - \log\frac{\delta'}{\tau}\right] \leq 0\right\}$$

$$\delta_* \triangleq \min_{\tau>0} \delta_*(\tau)$$

Main Result

Def: Hypothesis Testing Error Exponent (Mismatched, Worst Case)

$$E_{HT}(\tau,\delta) \triangleq \min_{\substack{\theta,\theta_n,\theta_f\in\Theta \\ \ell(\theta,\theta_n)\le\tau \\ \ell(\theta,\theta_f)\ge\delta}} -\frac{1}{d}\log P_\theta\left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n)\ge 1\right]$$

Def: Critical Loss

$$\delta_*(\tau) \triangleq \sup\left\{\delta' > 0 : \left[E_{HT}(\tau,\delta') - \log\frac{\delta'}{\tau}\right] \le 0\right\}$$

$$\delta_* \triangleq \min_{\tau>0} \delta_*(\tau)$$

Theorem

$$\min_{\hat{\theta}(\cdot)}\max_{\theta} P_\theta\left[\left\|\theta-\hat{\theta}(Y^n)\right\| > \delta_*\right] \xrightarrow{d\to\infty} 0$$

Main Result

Def: Hypothesis Testing Error Exponent (Mismatched,Worst Case)

$$E_{HT}(\tau,\delta) \triangleq \min_{\substack{\theta,\theta_n,\theta_f\in\Theta \\ \ell(\theta,\theta_n)\le\tau \\ \ell(\theta,\theta_f)\ge\delta}} -\frac{1}{d}\log P_\theta\left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n)\ge 1\right]$$

Def: Critical Loss

$$\delta_*(\tau) \triangleq \sup\left\{\delta' > 0 : \left[E_{HT}(\tau,\delta') - \log\frac{\delta'}{\tau}\right] \le 0\right\}$$

$$\delta_* \triangleq \min_{\tau>0} \delta_*(\tau)$$

Theorem

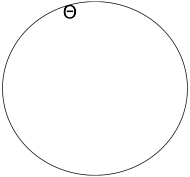
$$\min_{\hat\theta(\cdot)}\max_{\theta} P_\theta\left[\left\|\theta-\hat\theta(Y^n)\right\| > \delta_*\right] \xrightarrow{d\to\infty} 0$$

To use theorem: bound $E_{HT}(\tau,\delta)$, choose τ

$\hat{\theta} \triangleq$ Quantized Maximum Likelihood Estimator

Reminder

Estimation, upper bound: $P_\theta\left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right]$

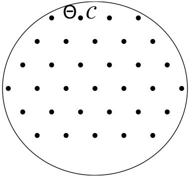


- Θ

$\hat{\theta} \triangleq$ Quantized Maximum Likelihood Estimator

Reminder

Estimation, upper bound: $P_\theta\left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right]$

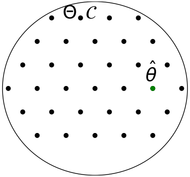


- Θ
- τ-cover Θ by a discrete net $\mathcal{C}$

$\hat{\theta} \triangleq$ Quantized Maximum Likelihood Estimator

Reminder

Estimation, upper bound: $P_\theta \left[\ell \left(\theta, \hat{\theta}(Y^n) \right) > \delta \right]$



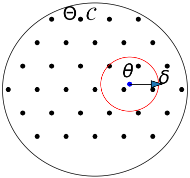
- Θ
- τ-cover Θ by a discrete net $\mathcal{C}$

Estimator

$$\hat{\theta}(y^n) \triangleq argmax_{\theta' \in \mathcal{C}} \left[\frac{dP_{\theta'}}{d\mu}(y^n) \right]$$

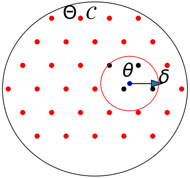
Error Probability Upper Bound - Basic Analysis

$$P_\theta\left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right] = ?$$



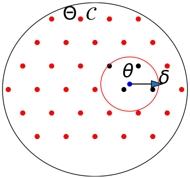
- θ is chosen arbitrarily (for minimax bound)

Error Probability Upper Bound - Basic Analysis



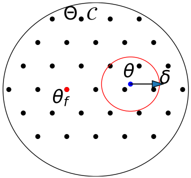
- Denote "bad" candidates $S \triangleq \{\theta' \in \mathcal{C} : \ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\}$

Error Probability Upper Bound - Basic Analysis



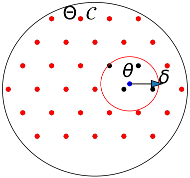
- Denote "bad" candidates $S \triangleq \{\theta' \in \mathcal{C} : \ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\}$
- $P_\theta\left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right] = P_\theta\left[\hat{\theta} \in S\right] = ?$

Error Probability Upper Bound - Basic Analysis



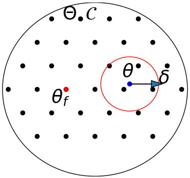
- Denote "bad" candidates $S \triangleq \{\theta' \in \mathcal{C} : \ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\}$
- $P_\theta\left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right] = P_\theta\left[\hat{\theta} \in S\right] = \sum_{\theta_f \in S} P_\theta\left[\hat{\theta} = \theta_f\right]$

Basic Analysis: Candidate Count



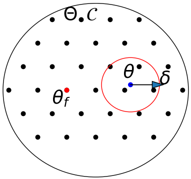
- Bound amount of far candidates by $|\mathcal{C}|$

Basic Analysis: Candidate Error Exponent



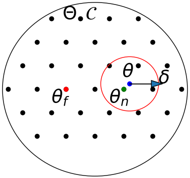
- Let far candidate $\theta_f \in \mathcal{C}$: $\ell(\theta, \theta_f) \geq \delta$.

Basic Analysis: Candidate Error Exponent



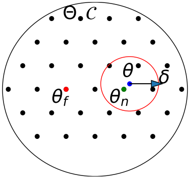
- Let far candidate $\theta_f \in \mathcal{C}$: $\ell(\theta, \theta_f) \geq \delta$.
- $P_\theta \left[\hat{\theta} = \theta_f \right] =??$

Basic Analysis: Candidate Error Exponent



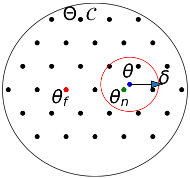
- Let far candidate $\theta_f \in \mathcal{C}$: $\ell(\theta, \theta_f) \geq \delta$.
- $P_\theta\left[\hat{\theta} = \theta_f\right] = ??$
- Exists near neighbour $\theta_n \in \mathcal{C}$: $\ell(\theta, \theta_n) \leq \tau$

Basic Analysis: Candidate Error Exponent



- Let far candidate $\theta_f \in \mathcal{C}$: $\ell(\theta, \theta_f) \geq \delta$.
- Exists near neighbour $\theta_n \in \mathcal{C}$: $\ell(\theta, \theta_n) \leq \tau$
- $\log P_\theta \left[\hat{\theta} = \theta_f \right] \leq \log P_\theta \left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n) \geq 1 \right]$

Basic Analysis: Candidate Error Exponent

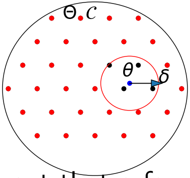


- Let far candidate $\theta_f \in \mathcal{C}$: $\ell(\theta, \theta_f) \geq \delta$.
- Exists near neighbour $\theta_n \in \mathcal{C}$: $\ell(\theta, \theta_n) \leq \tau$
- $\log P_\theta \left[\hat{\theta} = \theta_f\right] \leq \log P_\theta \left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n) \geq 1\right]$

Def: Hypothesis Testing Error Exponent (Mismatched,Worst Case)

$$E_{HT}(\tau, \delta) = \min_{\substack{\theta, \theta_n, \theta_f \in \Theta \\ \ell(\theta, \theta_n) \leq \tau \\ \ell(\theta, \theta_f) \geq \delta}} -\frac{1}{d} \log P_\theta \left[\frac{dP_{\theta_f}}{dP_{\theta_n}}(Y^n) \geq 1\right]$$

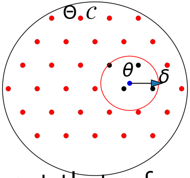
Basic Analysis: Global Error Exponent



- Upper bound on the event that a far candidate is the winner

$$\log P_\theta\left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right] \lessapprox -d\left[E_{HT}(\tau, \delta) - \frac{1}{d}\log|\mathcal{C}|\right]$$

Basic Analysis: Global Error Exponent

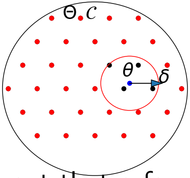


- Upper bound on the event that a far candidate is the winner

$$\log P_\theta \left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta \right] \lessapprox -d \left[E_{HT}(\tau, \delta) - \frac{1}{d} \log |\mathcal{C}| \right]$$

- "Competition": E_{HT} VS amount of candidates

Basic Analysis: Global Error Exponent

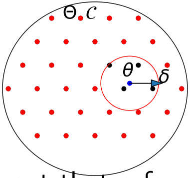


- Upper bound on the event that a far candidate is the winner

$$\log P_\theta\left[\ell\left(\theta,\hat{\theta}(Y^n)\right)>\delta\right]\lessapprox -d\left[E_{HT}(\tau,\delta)-\frac{1}{d}\log|\mathcal{C}|\right]$$

- "Competition": E_{HT} VS amount of candidates
- Large $\mathrm{Vol}(\Theta) \Rightarrow$ useless bound

Basic Analysis: Global Error Exponent



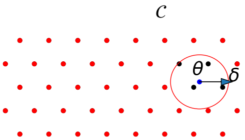
- Upper bound on the event that a far candidate is the winner

$$\log P_\theta \left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right] \lessapprox -d\left[E_{HT}(\tau, \delta) - \frac{1}{d}\log|\mathcal{C}|\right]$$

- "Competition": E_{HT} VS amount of candidates
- Large $\mathrm{Vol}(\Theta) \Rightarrow$ useless bound

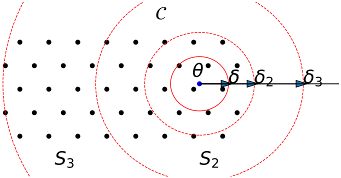
If P_θ and ℓ "matched": far candidates $\iff$ high error exponent
$\Rightarrow$ Can exploit this using a finer bounding method

Error Probability Upper Bound - Shell Analysis



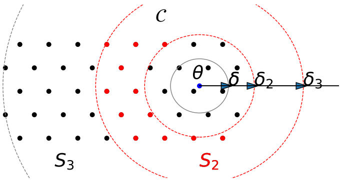
- $P_\theta\left[\hat{\theta} \in S\right] = ?$

Error Probability Upper Bound - Shell Analysis



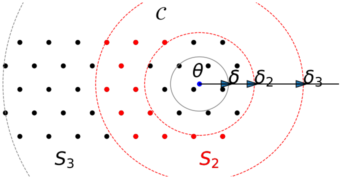
- $P_\theta\left[\hat{\theta} \in S\right] =?$
- Radii: $\delta = \delta_1 < \delta_2 < \cdots < \delta_k = \mathrm{diam}(\Theta)$

Error Probability Upper Bound - Shell Analysis



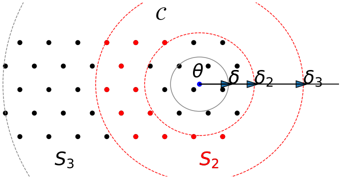
- $P_\theta\left[\hat{\theta} \in S\right] = ?$
- Radii: $\delta = \delta_1 < \delta_2 < \cdots < \delta_k = \mathrm{diam}(\Theta)$
- Shells: $S_i = \{\theta' \in \mathcal{C} : \delta_i < \ell(\theta, \theta') \leq \delta_{i+1}\}$

Error Probability Upper Bound - Shell Analysis



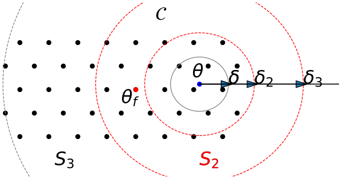
- Radii: $\delta = \delta_1 < \delta_2 < \cdots < \delta_k = \mathrm{diam}(\Theta)$
- Shells: $S_i = \{\theta' \in \mathcal{C} : \delta_i < \ell(\theta, \theta') \leq \delta_{i+1}\}$
- $P_\theta \left[\hat{\theta} \in S\right] = \sum_{i=1,\ldots,k-1} P_\theta \left[\hat{\theta} \in S_i\right]$

Error Probability Upper Bound - Shell Analysis



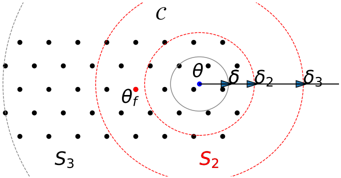
- $P_\theta\left[\hat{\theta} \in S_i\right] = ?$

Error Probability Upper Bound - Shell Analysis



- $P_\theta\left[\hat{\theta} \in S_i\right] = \sum_{\theta_f \in S_i} P_\theta\left[\hat{\theta} = \theta_f\right]$

Error Probability Upper Bound - Shell Analysis



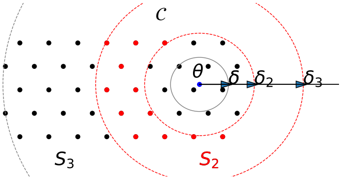
- $P_\theta\left[\hat{\theta} \in S_i\right] = \sum_{\theta_f \in S_i} P_\theta\left[\hat{\theta} = \theta_f\right]$
- $-\frac{1}{d}\log P_\theta\left[\hat{\theta} = \theta_f\right] \geq E_{HT}(\tau, \delta_i)$

Shell Analysis - Density Control

Should control candidate count

Def: Ball around θ with radius r w.r.t loss ℓ: $\mathcal{B}(\theta, r, \ell)$

$|S_i|$

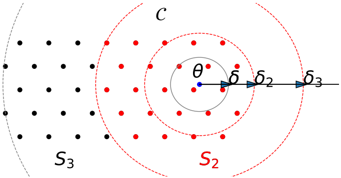


Shell Analysis - Density Control

Should control candidate count in balls

Def: Ball around θ with radius r w.r.t loss ℓ: $\mathcal{B}(\theta, r, \ell)$

$$|S_i| \leq |\mathcal{C} \cap \mathcal{B}(\theta, \delta_{i+1}, \ell)|$$

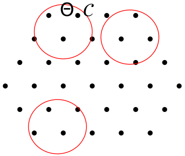


Shell Analysis - Density Control

Should control candidate count in balls for every center

Def: Ball around θ with radius r w.r.t loss ℓ: $\mathcal{B}(\theta, r, \ell)$

$$|S_i| \leq |\mathcal{C} \cap \mathcal{B}(\theta, \delta_{i+1}, \ell)| \leq max_{\theta' \in \Theta} |\mathcal{C} \cap \mathcal{B}(\theta', \delta_{i+1}, \ell)|$$

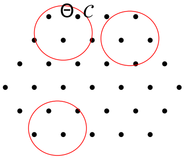


Shell Analysis - Density Control

Should control candidate count in balls for every center

Def: Ball around θ with radius r w.r.t loss ℓ: $\mathcal{B}(\theta, r, \ell)$

$$\frac{1}{d}\log|S_i| \leq \frac{1}{d}\log\left(max_{\theta' \in \Theta}|\mathcal{C} \cap \mathcal{B}(\theta', \delta_{i+1}, \ell)|\right)$$

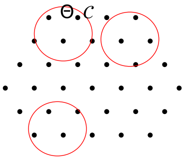


Shell Analysis - Density Control

Should control candidate count in balls for every center

Def: Ball around θ with radius r w.r.t loss ℓ: $\mathcal{B}(\theta, r, \ell)$

$$\frac{1}{d}\log|S_i| \leq \frac{1}{d}\log\left(max_{\theta'\in\Theta}|\mathcal{C}\cap\mathcal{B}(\theta', \delta_{i+1}, \ell)|\right) \triangleq M_{\mathcal{C}}(\delta_{i+1})$$



Def: Net Population Function

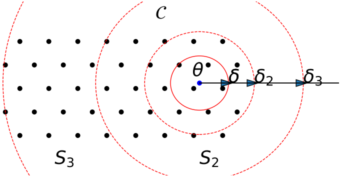
$$M_{\mathcal{C}}(r) = \frac{1}{d}\log\max_{\theta'\in\mathcal{C}}|\mathcal{C}\cap\mathcal{B}(\theta', r, \ell)|$$

Shell Analysis - Density Control

Should control candidate count in balls for every center, radius

Def: Ball around θ with radius r w.r.t loss ℓ: $\mathcal{B}(\theta, r, \ell)$

$$\frac{1}{d}\log|S_i| \leq \frac{1}{d}\log\left(max_{\theta'\in\Theta}|\mathcal{C}\cap\mathcal{B}(\theta', \delta_{i+1}, \ell)|\right)$$



Def: Net Population Function

$M_{\mathcal{C}}(r) = \frac{1}{d}\log\max_{\theta'\in\mathcal{C}}|\mathcal{C}\cap\mathcal{B}(\theta', r, \ell)|$

- Should bound $M_{\mathcal{C}}(r)$ for $r = \delta_1, \delta_2, \ldots, \delta_k$

Existence of τ-net with nearly optimal density

Def: Net Population Function

$M_{\mathcal{C}}(r) = \frac{1}{d} \log \max_{\theta' \in \mathcal{C}} |\mathcal{C} \cap \mathcal{B}(\theta', r, \ell)|$

Should bound $M_{\mathcal{C}}(r)$ for $r = \delta_1, \delta_2, \ldots, \delta_k$

Theorem

Let $d > 25$, $\Theta \subset \mathbb{R}^d$, and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some arbitrary norm on $\mathbb{R}^d$. There exists a (lattice) τ-cover $\mathcal{C}$ of Θ satisfying

$$M_{\mathcal{C}}(\delta) \leq \log\left(\frac{\delta}{\tau}\right) + 3\frac{\log d}{d} + \frac{133 \log 2}{d}, \quad \forall \delta > \tau.$$

Existence of τ-net with nearly optimal density

Def: Net Population Function

$M_{\mathcal{C}}(r) = \frac{1}{d} \log \max_{\theta' \in \mathcal{C}} |\mathcal{C} \cap \mathcal{B}(\theta', r, \ell)|$

Should bound $M_{\mathcal{C}}(r)$ for $r = \delta_1, \delta_2, \ldots, \delta_k$

Theorem

Let $d > 25$, $\Theta \subset \mathbb{R}^d$, and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some arbitrary norm on $\mathbb{R}^d$. There exists a (lattice) τ-cover $\mathcal{C}$ of Θ satisfying

$$M_{\mathcal{C}}(\delta) \leq \log\left(\frac{\delta}{\tau}\right) + 3\frac{\log d}{d} + \frac{133 \log 2}{d}, \quad \forall \delta > \tau.$$

Bound is tight: $\forall$ τ-cover $M_{\mathcal{C}}(\delta) \geq \log\left(\frac{\delta}{\tau}\right)$

Existence of τ-net with nearly optimal density

Def: Net Population Function

$M_{\mathcal{C}}(r) = \frac{1}{d} \log \max_{\theta' \in \mathcal{C}} |\mathcal{C} \cap \mathcal{B}(\theta', r, \ell)|$

Should bound $M_{\mathcal{C}}(r)$ for $r = \delta_1, \delta_2, \ldots, \delta_k$

Theorem

Let $d > 25$, $\Theta \subset \mathbb{R}^d$, and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some arbitrary norm on $\mathbb{R}^d$. There exists a (lattice) τ-cover $\mathcal{C}$ of Θ satisfying

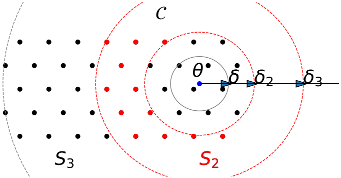
$$M_{\mathcal{C}}(\delta) \leq \log\left(\frac{\delta}{\tau}\right) + 3\frac{\log d}{d} + \frac{133 \log 2}{d}, \quad \forall \delta > \tau.$$

Bound is tight: $\forall$ τ-cover $M_{\mathcal{C}}(\delta) \geq \log\left(\frac{\delta}{\tau}\right)$

Proof is based on a uniform lattice covering result of O.-Regev-Weiss. Result of similar spirit can be derived using Erdős and Rogers '62

Shell Analysis - "Shell Error Exponent"

$$P_\theta\left[\hat{\theta} \in S_i\right] = \sum_{\theta_f \in S_i} P_\theta\left[\hat{\theta} = \theta_f\right]$$

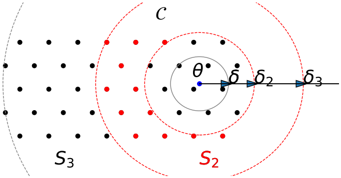


"Shell error exponent":

$$\log P_\theta\left[\hat{\theta} \in S_i\right] \lessapprox -d\left[E_{HT}(\tau, \delta_i) - M_{\mathcal{C}}(\delta_{i+1})\right]$$

Shell Analysis - "Shell Error Exponent"

$$P_\theta\left[\hat{\theta} \in S_i\right] = \sum_{\theta_f \in S_i} P_\theta\left[\hat{\theta} = \theta_f\right]$$

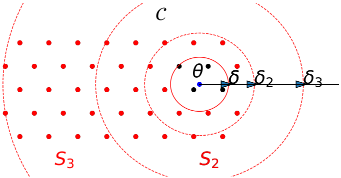


"Shell error exponent":

$$\log P_\theta\left[\hat{\theta} \in S_i\right] \lessapprox -d\left[E_{HT}(\tau, \delta_i) - \log\left(\frac{\delta_{i+1}}{\tau}\right) + o(1)\right]$$

Shell Analysis - Global Error Exponent

$$P_\theta\left[\hat{\theta} \in S\right] = \sum_{i=1}^{k} P_\theta\left[\hat{\theta} \in S_i\right]$$



If number of shells k is sub-exponential, P_e dictated by "Dominant shell error exponent":

$$\min_{i=1,\ldots,k-1}\left[E_{HT}(\tau,\delta_i) - \log\left(\frac{\delta_{i+1}}{\tau}\right) + o(1)\right] \overset{?}{>} 0$$

Shell Analysis: General Result

Theorem

Let

- $\mathcal{C}$ be a τ-cover with "good" density
- $\delta_i = \delta \cdot e^{\frac{i-1}{d}}$, $i = 1, \ldots, k = e^{o(d)}$
  note that $\log\left(\frac{\delta_{i+1}}{\tau}\right) = \log\left(\frac{\delta_i}{\tau}\right) + o(1)$

$$-\frac{1}{d} \log P_\theta \left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right] \geq \min_{i=1,\ldots,k-1} \left[E_{HT}(\tau, \delta_i) - \log\left(\frac{\delta_i}{\tau}\right) + o(1)\right]$$

Shell Analysis: General Result

Theorem

Let

- $\mathcal{C}$ be a τ-cover with "good" density
- $\delta_i = \delta \cdot e^{\frac{i-1}{d}}$, $i = 1, \ldots, k = e^{o(d)}$
  note that $\log\left(\frac{\delta_{i+1}}{\tau}\right) = \log\left(\frac{\delta_i}{\tau}\right) + o(1)$

$$-\frac{1}{d}\log P_\theta\left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right]$$
$$\geq \min_{\delta' > \delta}\left[E_{HT}(\tau, \delta') - \log\left(\frac{\delta'}{\tau}\right) + o(1)\right]$$

Shell Analysis: General Result

Theorem

Let

- $\mathcal{C}$ be a τ-cover with "good" density
- $\delta_i = \delta \cdot e^{\frac{i-1}{d}}$, $i = 1, \ldots, k = e^{o(d)}$
  note that $\log\left(\frac{\delta_{i+1}}{\tau}\right) = \log\left(\frac{\delta_i}{\tau}\right) + o(1)$

$$-\frac{1}{d}\log P_\theta\left[\ell\left(\theta, \hat{\theta}(Y^n)\right) > \delta\right] \geq \min_{\delta' > \delta}\left[E_{HT}(\tau, \delta') - \log\left(\frac{\delta'}{\tau}\right) + o(1)\right]$$

Upper bound $\iff$ Lower bound

- $\log\left(\frac{\delta'}{\tau}\right) \iff$ Rate distortion function of $\Theta, \ell = \|\theta - \hat{\theta}\|$
- $E_{HT}(\tau, \delta) \iff C(P_{Y^n|\theta})$

Main Result

Def: Critical Loss

$$\delta_*(\tau) \triangleq \sup \left\{ \delta' > 0 : \left[E_{HT}(\tau, \delta') - \log \frac{\delta'}{\tau} \right] \leq 0 \right\}$$

$$\delta_* \triangleq \min_{\tau > 0} \delta_*(\tau)$$

Main Result

Def: Critical Loss

$$\delta_*(\tau) \triangleq \sup \left\{ \delta' > 0 : \left[E_{HT}(\tau, \delta') - \log \frac{\delta'}{\tau} \right] \leq 0 \right\}$$

$$\delta_* \triangleq \min_{\tau > 0} \delta_*(\tau)$$

Main Result

Def: Critical Loss

$$\delta_*(\tau) \triangleq \sup \left\{ \delta' > 0 : \left[E_{HT}(\tau, \delta') - \log \frac{\delta'}{\tau} \right] \leq 0 \right\}$$

$$\delta_* \triangleq \min_{\tau > 0} \delta_*(\tau)$$

Theorem

$$\min_{\hat{\theta}(\cdot)} \max_{\theta} P_\theta \left[\left\| \theta - \hat{\theta}(Y^n) \right\| > \delta_* \right] \xrightarrow{d \to \infty} 0$$

Main Result

Def: Critical Loss

$$\delta_*(\tau) \triangleq \sup\left\{\delta' > 0 : \left[E_{HT}(\tau,\delta') - \log\frac{\delta'}{\tau}\right] \leq 0\right\}$$

$$\delta_* \triangleq \min_{\tau>0} \delta_*(\tau)$$

Theorem

$$\min_{\hat{\theta}(\cdot)} \max_{\theta} P_\theta\left[\left\|\theta - \hat{\theta}(Y^n)\right\| > \delta_*\right] \xrightarrow{d\to\infty} 0$$

- To use theorem: bound $E_{HT}(\tau,\delta)$, choose τ

Error Exponent in Gaussian Cases + Examples

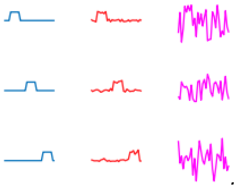
- $\ell(\theta,\theta') = \|\theta-\theta'\|_2^2$
- Transformation $g:\Theta \to \mathbb{R}^n$
- $P_\theta = \mathcal{N}(g(\theta), \sigma^2 I_n)$

Bound $E_{HT}(\tau,\delta) \geq \frac{1}{4}\psi(\tau,\delta)^2$

$$\psi(\tau,\delta) \triangleq \frac{1}{\sqrt{d}} \min_{\substack{\theta,\theta_n,\theta_f\in\Theta \\ \ell(\theta,\theta_n)\leq\tau \\ \ell(\theta,\theta_f)\geq\delta}} \left[\sqrt{\frac{\|g(\theta_f)-g(\theta)\|_2^2}{2\sigma^2}} - \sqrt{\frac{\|g(\theta_n)-g(\theta)\|_2^2}{2\sigma^2}}\right]$$

- Interpretation: Euclidean geometry optimization
- For GLM: $g(\theta)=\theta \to \delta^* < 32\log(2)\sigma^2 d$
- For spiked Wigner: $g(\theta) = \lambda \mathrm{vec}(\theta\theta^T)$, $\sigma^2 = \frac{1}{d}$ (we assume $\|\theta\| > 1$, $\lambda > \sqrt{58}$) $\to \delta^* < \frac{58}{\lambda^2}$

Example - Multi Reference Alignment



- $Y_j = R_{k_j}\theta + \sigma Z_j$, $j = 1, \ldots, m$, and $k_j \overset{i.i.d.}{\sim} \mathrm{Unif}([d])$
- Define extended parameter space $\tilde{\Theta} = \mathbb{R}^d \times [d]^m$, such that $\tilde{\theta} = (\theta, k_1, \ldots, k_m)$ also includes the (nuisance) shifts, and $\tilde{\ell}(\tilde{\theta}, \hat{\tilde{\theta}}) = \frac{1}{m}\|g(\tilde{\theta}) - g(\hat{\tilde{\theta}})\|_2^2$, where $g(\tilde{\theta}) = \mathrm{vec}(R_{k_1}\theta, \ldots, R_{k_m}\theta)$

$$\min_{\hat{\theta}(\cdot)} \max_{\theta < e^{e^{o(d)}}} P_\theta \left[\min_k \|R_k\theta - \hat{\theta}(Y^m)\|_2^2 \geq \frac{32\sigma^2 d}{m} \left(\log 2 + m \frac{\log d}{d} \right) \right] \to 0$$

- Upper bound equivalent to GLM for $m = O\left(\frac{d}{\log d}\right)$

Conclusions

- We presented a general framework to upper bound minimax risk (PAC setup), which is applicable to essentially unbounded parameter spaces
- The bound is based on a delicate shell analysis and mismatched BHT
- For $\Theta \subset \mathbb{R}^d, \ell(\theta, \theta') = \|\theta - \theta'\|$ our bound takes a relatively simple form
- For discrete product distributions on $\mathcal{Y}^n$, if $\frac{P_{\theta,j}(y)}{P_{\theta',j}(y)} \in 1 \pm \kappa_{n,d}$ for all $j \in [n], y \in \mathcal{Y}, \theta, \theta' \in \Theta$ we can prove that

$$E_{HT}(\tau, \delta) \geq \frac{1}{4d}\left[\sqrt{D_\delta} - \sqrt{D_\tau}\right]^2 + \frac{n}{d} \cdot O(\kappa_{n,d}^3),$$

where

$D_\tau \triangleq \sup_{\substack{\theta,\theta_n \in \Theta \\ \ell(\theta,\theta_n) \leq \tau}} D\left(P_\theta \| P_{\theta_n}\right), D_\delta \triangleq \inf_{\substack{\theta,\theta_n \in \Theta \\ \ell(\theta,\theta_n) \geq \delta}} D\left(P_\theta \| P_{\theta_f}\right)$