Minimax Estimation of Kernel Mean Embeddings

Bharath K. Sriperumbudur

Department of Statistics
Pennsylvania State University

Gatsby Computational Neuroscience Unit
May 4, 2016

Collaborators

- Dr. Ilya Tolstikhin : Max Planck Institute for Intelligent Systems, Tübingen.
- Dr. Krikamol Muandet : Max Planck Institute for Intelligent Systems, Tübingen.

Kernel Mean Embedding (KME)

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel.

- Kernel trick:

$$y \mapsto \overbrace{\phi(y)}^{k(\cdot,y)}$$

- Equivalently,

$$\delta_y \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\delta_y(x)$$

- Generalization:

$$\mathbb{P} \mapsto \underbrace{\int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)}_{\textit{kernel mean embedding}} =: \mu_{\mathbb{P}}.$$

Properties

- KME is a generalization of
  - Characteristic function : $k(\cdot, x) = e^{-\sqrt{-1}\langle \cdot, x \rangle}$, $x \in \mathbb{R}^d$
  - Moment generating function : $k(\cdot, x) = e^{\langle \cdot, x \rangle}$, $x \in \mathbb{R}^d$

  to arbitrary $\mathcal{X}$.

- In general, many $\mathbb{P}$ can yield the same KME!!

$$\text{for } k(x, y) = \langle x, y \rangle, \text{ we have } \mathbb{P} \mapsto \mu_{\mathbb{P}}.$$

- Characteristic kernels: They ensure that no two different $\mathbb{P}$ can have the same KME.

$$\mathbb{P} \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) \text{ is one-to-one.}$$

  Examples: Gaussian, Matérn, ... (infinite dimensional RKHS)

Application: Two-Sample Problem

- Given random samples $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Determine: $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \neq \mathbb{Q}$?
- $\gamma(\mathbb{P}, \mathbb{Q})$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

$$\begin{matrix} H_0 : \mathbb{P} = \mathbb{Q} \\ H_1 : \mathbb{P} \neq \mathbb{Q} \end{matrix} \equiv \begin{matrix} H_0 : \gamma(\mathbb{P}, \mathbb{Q}) = 0 \\ H_1 : \gamma(\mathbb{P}, \mathbb{Q}) > 0 \end{matrix}$$

- Test: Say H_0 if $\hat{\gamma}\left(\{X_i\}_{i=1}^m, \{Y_j\}_{j=1}^n\right) < \varepsilon$. Otherwise say H_1.

Idea: Use

$$\gamma(\mathbb{P}, \mathbb{Q}) = \left\| \int k(\cdot, x)\, d\mathbb{P}(x) - \int k(\cdot, x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}_k}$$

with k being characteristic.

More Applications

- Testing for independence (Gretton et al., 2008)
- Conditional independence tests (Fukumizu et al., 2008)
- Feature selection (Song et al., 2012)
- Distribution regression (Szabó et al., 2015)
- Causal inference (Lopez-Paz et al., 2015)
- Mixture density estimation (Sriperumbudur, 2011), ...

Estimators of KME

- In applications, $\mathbb{P}$ is unknown and only samples $\{X_i\}_{i=1}^n$ from it are known.
- A popular estimator of KME that has been employed in all these applications is the empirical estimator:

$$\hat{\mu}_{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)$$

Theorem (Smola et al., 2007; Gretton et al., 2012; Lopez-Paz et al., 2015)

Suppose $\sup_{x\in\mathcal{X}} k(x,x) \le C < \infty$ *where* k *is continuous. Then for any* $\tau > 0$,

$$\mathbb{P}^n\left(\left\{(X_i)_{i=1}^n : \|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}_k} \ge \sqrt{\frac{C}{n}} + \sqrt{\frac{2C\tau}{n}}\right\}\right) \le e^{-\tau}.$$

Alternatively $\mathbb{E}\,\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}_k} \le \frac{C'}{\sqrt{n}}$ for some $C' > 0$.

Shrinkage Estimator

Given $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} N(\mu, \sigma^2 I)$, suppose we are interested in estimating $\mu \in \mathbb{R}^d$.

- Maximum likelihood estimator: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ which is the empirical estimator.

- (James and Stein, 1961): constructed an estimator $\check{\mu}$ such that for $d \geq 3$ for all $\mu \in \mathbb{R}^d$,

$$\mathbb{E}\|\check{\mu} - \mu\|^2 \leq \mathbb{E}\|\hat{\mu} - \mu\|^2$$

  and for at least one μ, the strict inequality holds.

Kernel setting: Based on the above motivation, (Krikamol et al., 2015) proposed a shrinkage estimator, $\check{\mu}_{\mathbb{P}}$ of $\mu_{\mathbb{P}}$ and showed that

$$\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}_k} < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}_k} + O_p(n^{-3/2})$$

as $n \to \infty$ and $\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}_k} \leq C'' n^{-1/2}$ for some $C'' > 0$.

Shrinkage Estimator

Given $(X_i)_{i=1}^n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2 I)$, suppose we are interested in estimating $\mu \in \mathbb{R}^d$.

- Maximum likelihood estimator: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ which is the empirical estimator.

- (James and Stein, 1961): constructed an estimator $\check{\mu}$ such that for $d \geq 3$ for all $\mu \in \mathbb{R}^d$,

$$\mathbb{E}\|\check{\mu} - \mu\|^2 \leq \mathbb{E}\|\hat{\mu} - \mu\|^2$$

  and for at least one μ, the strict inequality holds.

Kernel setting: Based on the above motivation, (Krikamol et al., 2015) proposed a shrinkage estimator, $\check{\mu}_{\mathbb{P}}$ of $\mu_{\mathbb{P}}$ and showed that

$$\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}_k} < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}_k} + O_p(n^{-3/2})$$

as $n \to \infty$ and $\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}_k} \leq C'' n^{-1/2}$ for some $C'' > 0$.

Shrinkage Estimator

Given $(X_i)_{i=1}^n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2 I)$, suppose we are interested in estimating $\mu \in \mathbb{R}^d$.

- Maximum likelihood estimator: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ which is the empirical estimator.

- (James and Stein, 1961): constructed an estimator $\check{\mu}$ such that for $d \geq 3$ for all $\mu \in \mathbb{R}^d$,

$$\mathbb{E}\|\check{\mu} - \mu\|^2 \leq \mathbb{E}\|\hat{\mu} - \mu\|^2$$

  and for at least one μ, the strict inequality holds.

Kernel setting: Based on the above motivation, (Krikamol et al., 2015) proposed a shrinkage estimator, $\check{\mu}_{\mathbb{P}}$ of $\mu_{\mathbb{P}}$ and showed that

$$\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}_k}^2 < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}_k}^2 + O_p(n^{-3/2})$$

as $n \to \infty$ and $\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}_k} \leq C'' n^{-1/2}$ for some $C'' > 0$.

Main Message

Question: Can we do better using some other estimators?

Answer: for a large class of kernels the answer is NO.

We can do better in terms of constant factors (Muandet et al., 2015).

But not in terms of rates w.r.t. sample size n or dimensionality d (if $\mathcal{X} = \mathbb{R}^d$).

Tool: Minimax theory

Estimation Theory: Setup

Given:

- A class of distributions $\mathcal{P}$ on a sample space $\mathcal{X}$;
- A mapping $\theta : \mathcal{P} \to \Theta$, $\mathbb{P} \mapsto \theta(\mathbb{P})$.

Goal:

- Estimate $\theta(\mathbb{P})$ based on i.i.d. observations $(X_i)_{i=1}^n$ drawn from the unknown distribution $\mathbb{P}$.

Examples:

- $\mathcal{P} = \{N(\theta, \sigma^2) \,:\, \theta \in \mathbb{R}\}$ with known variance: $\theta(\mathbb{P}) = \int x \, d\mathbb{P}(x)$.
- $\mathcal{P} = \{\text{set of all distributions}\} : \theta(\mathbb{P}) = \int k(\cdot, x) \, d\mathbb{P}(x)$.

Estimator:

$$\hat{\theta}(X_1, \ldots, X_n)$$

Estimation Theory: Setup

Given:

- A class of distributions $\mathcal{P}$ on a sample space $\mathcal{X}$;
- A mapping $\theta : \mathcal{P} \to \Theta$, $\mathbb{P} \mapsto \theta(\mathbb{P})$.

Goal:

- Estimate $\theta(\mathbb{P})$ based on i.i.d. observations $(X_i)_{i=1}^n$ drawn from the unknown distribution $\mathbb{P}$.

Examples:

- $\mathcal{P} = \{N(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ with known variance: $\theta(\mathbb{P}) = \int x \, d\mathbb{P}(x)$.
- $\mathcal{P} = \{\text{set of all distributions}\}$: $\theta(\mathbb{P}) = \int k(\cdot, x) \, d\mathbb{P}(x)$.

Estimator:

$$\hat{\theta}(X_1, \ldots, X_n)$$

Estimation Theory: Setup

Given:

- A class of distributions $\mathcal{P}$ on a sample space $\mathcal{X}$;
- A mapping $\theta : \mathcal{P} \to \Theta$, $\mathbb{P} \mapsto \theta(\mathbb{P})$.

Goal:

- Estimate $\theta(\mathbb{P})$ based on i.i.d. observations $(X_i)_{i=1}^n$ drawn from the unknown distribution $\mathbb{P}$.

Examples:

- $\mathcal{P} = \left\{N(\theta, \sigma^2) \,:\, \theta \in \mathbb{R}\right\}$ with known variance: $\theta(\mathbb{P}) = \int x \, d\mathbb{P}(x)$.
- $\mathcal{P} = \{\text{set of all distributions}\}$: $\theta(\mathbb{P}) = \int k(\cdot, x) \, d\mathbb{P}(x)$.

Estimator:

$$\hat{\theta}(X_1, \ldots, X_n)$$

Estimation Theory: Setup

Given:

- A class of distributions $\mathcal{P}$ on a sample space $\mathcal{X}$;
- A mapping $\theta : \mathcal{P} \to \Theta$, $\mathbb{P} \mapsto \theta(\mathbb{P})$.

Goal:

- Estimate $\theta(\mathbb{P})$ based on i.i.d. observations $(X_i)_{i=1}^n$ drawn from the unknown distribution $\mathbb{P}$.

Examples:

- $\mathcal{P} = \left\{N(\theta, \sigma^2) \, : \, \theta \in \mathbb{R}\right\}$ with known variance: $\theta(\mathbb{P}) = \int x \, d\mathbb{P}(x)$.
- $\mathcal{P} = \{\text{set of all distributions}\}$: $\theta(\mathbb{P}) = \int k(\cdot, x) \, d\mathbb{P}(x)$.

Estimator:

$$\hat{\theta}(X_1, \ldots, X_n)$$

Minimax Risk

How good is the estimator, $\hat{\theta}$?

- Define a distance $\rho : \Theta \times \Theta \to \mathbb{R}$ to measure the error of $\hat{\theta}$ for the parameter θ.
- The average performance of $\hat{\theta}$ is measured by the risk:

$$R(\hat{\theta}; \mathbb{P}) = \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right].$$

- Obviously, we would want an estimator that has the smallest risk for every $\mathbb{P}$: not achievable!!
- Global view: Minimize the average risk (Bayesian view) or the maximum risk,

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$$

- $\hat{\theta}^*$ is called a minimax estimator if

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}^*, \theta(\mathbb{P}))\right] = \overbrace{\inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]}^{\mathcal{M}_n(\theta(\mathcal{P}))\ :\ \text{minimax risk}}.$$

Minimax Risk

How good is the estimator, $\hat{\theta}$?

- Define a distance $\rho : \Theta \times \Theta \to \mathbb{R}$ to measure the error of $\hat{\theta}$ for the parameter θ.
- The average performance of $\hat{\theta}$ is measured by the risk:

$$R(\hat{\theta}; \mathbb{P}) = \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right].$$

- Obviously, we would want an estimator that has the smallest risk for every $\mathbb{P}$: not achievable!!
- Global view: Minimize the average risk (Bayesian view) or the maximum risk,

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$$

- $\hat{\theta}^*$ is called a minimax estimator if

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}^*, \theta(\mathbb{P}))\right] = \overbrace{\inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]}^{\mathcal{M}_n(\theta(\mathcal{P}))\ :\ \text{minimax risk}}.$$

Minimax Risk

How good is the estimator, $\hat{\theta}$?

- Define a distance $\rho : \Theta \times \Theta \to \mathbb{R}$ to measure the error of $\hat{\theta}$ for the parameter θ.
- The average performance of $\hat{\theta}$ is measured by the risk:

$$R(\hat{\theta}; \mathbb{P}) = \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right].$$

- Obviously, we would want an estimator that has the smallest risk for every $\mathbb{P}$: not achievable!!
- Global view: Minimize the average risk (Bayesian view) or the maximum risk,

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$$

- $\hat{\theta}^*$ is called a minimax estimator if

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}^*, \theta(\mathbb{P}))\right] = \overbrace{\inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]}^{\mathcal{M}_n(\theta(\mathcal{P}))\ :\ \text{minimax risk}}.$$

Minimax Risk

How good is the estimator, $\hat{\theta}$?

- Define a distance $\rho : \Theta \times \Theta \to \mathbb{R}$ to measure the error of $\hat{\theta}$ for the parameter θ.
- The average performance of $\hat{\theta}$ is measured by the risk:

$$R(\hat{\theta}; \mathbb{P}) = \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right].$$

- Obviously, we would want an estimator that has the smallest risk for every $\mathbb{P}$: not achievable!!
- Global view: Minimize the average risk (Bayesian view) or the maximum risk,

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$$

- $\hat{\theta}^*$ is called a minimax estimator if

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}^*, \theta(\mathbb{P}))\right] = \overbrace{\inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]}^{\mathcal{M}_n(\theta(\mathcal{P}))\ :\ \text{minimax risk}}.$$

Minimax Risk

How good is the estimator, $\hat{\theta}$?

- Define a distance $\rho : \Theta \times \Theta \to \mathbb{R}$ to measure the error of $\hat{\theta}$ for the parameter θ.
- The average performance of $\hat{\theta}$ is measured by the risk:

$$R(\hat{\theta}; \mathbb{P}) = \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right].$$

- Obviously, we would want an estimator that has the smallest risk for every $\mathbb{P}$: not achievable!!
- Global view: Minimize the average risk (Bayesian view) or the maximum risk,

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$$

- $\hat{\theta}^*$ is called a minimax estimator if

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}^*, \theta(\mathbb{P}))\right] = \overbrace{\inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]}^{\mathcal{M}_n(\theta(\mathcal{P}))\ :\ \text{minimax risk}}.$$

Minimax Estimator

- Statistical decision theory has two goals:
  - Find the minimax risk, $\mathcal{M}_n(\theta(\mathcal{P}))$.
  - Find the minimax estimator that achieves this risk.
- Except in simple cases, finding both the minimax risk and the minimax estimator is usually very hard.
- So we settle for an estimator that achieves the minimax rate:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}^a, \theta(\mathbb{P}))\right] \underbrace{\asymp}_{a_n \asymp b_n \,\equiv\, \frac{a_n}{b_n},\, \frac{b_n}{a_n} \text{ are bounded}} \mathcal{M}_n(\theta(\mathcal{P}))$$

Minimax Estimator

- Statistical decision theory has two goals:
  - Find the minimax risk, $\mathcal{M}_n(\theta(\mathcal{P}))$.
  - Find the minimax estimator that achieves this risk.
- Except in simple cases, finding both the minimax risk and the minimax estimator is usually very hard.
- So we settle for an estimator that achieves the minimax rate:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}^a, \theta(\mathbb{P}))\right] \underbrace{\asymp}_{a_n \asymp b_n \equiv \frac{a_n}{b_n}, \frac{b_n}{a_n} \text{ are bounded}} \mathcal{M}_n(\theta(\mathcal{P}))$$

Minimax Estimator

- Suppose we have an estimator $\hat{\theta}_\star$ such that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}_\star, \theta(\mathbb{P}))\right] \le C\psi_n$$

  for some $C > 0$ and $\psi_n \to 0$ as $n \to \infty$.

- If

$$\mathcal{M}_n(\theta(\mathcal{P})) \ge c\psi_n$$

  for some $c > 0$, then $\hat{\theta}_\star$ is minimax ψ_n-rate optimal.

Our Problem:

- $\theta(\mathbb{P}) = \mu_{\mathbb{P}} = \int k(\cdot, x)\, d\mathbb{P}(x)$
- $\rho = \|\cdot\|_{\mathcal{H}}$
- We have that $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}_\star, \theta(\mathbb{P}))\right] \le \frac{C_\star}{\sqrt{n}}$ for $\hat{\theta}_\star$ being an empirical estimator, shrinkage estimator and kernel density based estimator.

What is $\mathcal{M}_n(\mu(\mathcal{P}))$?

Minimax Estimator

- Suppose we have an estimator $\hat{\theta}_\star$ such that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}_\star, \theta(\mathbb{P}))\right] \leq C\psi_n$$

  for some $C > 0$ and $\psi_n \to 0$ as $n \to \infty$.

- If

$$\mathcal{M}_n(\theta(\mathcal{P})) \geq c\psi_n$$

  for some $c > 0$, then $\hat{\theta}_\star$ is minimax ψ_n-rate optimal.

Our Problem:

- $\theta(\mathbb{P}) = \mu_{\mathbb{P}} = \int k(\cdot, x)\, d\mathbb{P}(x)$
- $\rho = \|\cdot\|_{\mathcal{H}}$
- We have that $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}_\star, \theta(\mathbb{P}))\right] \leq \frac{C_\star}{\sqrt{n}}$ for $\hat{\theta}_\star$ being an empirical estimator, shrinkage estimator and kernel density based estimator.

What is $\mathcal{M}_n(\mu(\mathcal{P}))$?

From Estimation to Testing

Key Idea: Reduce the estimation problem to a testing problem and bound $\mathcal{M}_n(\theta(\mathcal{P}))$ in terms of the probability of error in testing problems.

Setup:

- Let $\{\mathbb{P}_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ where $\mathcal{V} = \{1, \ldots, M\}$.
- The family induces a collection of parameters $\{\theta(\mathbb{P}_v)\}_{v \in \mathcal{V}}$.
- Choose $\{\mathbb{P}_v\}_{v \in \mathcal{V}}$ such that

$$\rho(\theta(\mathbb{P}_v), \theta(\mathbb{P}_{v'})) \geq 2\delta, \qquad \text{for all } v \neq v'.$$

- Suppose we observe $(X_i)_{i=1}^n$ is drawn from the n-fold product distribution, $\mathbb{P}_{v^*}^n$ for some $v^* \in \mathcal{V}$.
- Construct $\hat{\theta}(X_1, \ldots, X_n)$.

Testing problem:

- Based on $(X_i)_{i=1}^n$, test which of M hypothesis is true.

From Estimation to Testing

Key Idea: Reduce the estimation problem to a testing problem and bound $\mathcal{M}_n(\theta(\mathcal{P}))$ in terms of the probability of error in testing problems.

Setup:

- Let $\{\mathbb{P}_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ where $\mathcal{V} = \{1, \ldots, M\}$.
- The family induces a collection of parameters $\{\theta(\mathbb{P}_v)\}_{v \in \mathcal{V}}$.
- Choose $\{\mathbb{P}_v\}_{v \in \mathcal{V}}$ such that

$$\rho(\theta(\mathbb{P}_v), \theta(\mathbb{P}_{v'})) \geq 2\delta, \qquad \text{for all } v \neq v'.$$

- Suppose we observe $(X_i)_{i=1}^n$ is drawn from the n-fold product distribution, $\mathbb{P}_{v^*}^n$ for some $v^* \in \mathcal{V}$.
- Construct $\hat{\theta}(X_1, \ldots, X_n)$.

Testing problem:

- Based on $(X_i)_{i=1}^n$, test which of M hypothesis is true.

From Estimation to Testing

Key Idea: Reduce the estimation problem to a testing problem and bound $\mathcal{M}_n(\theta(\mathcal{P}))$ in terms of the probability of error in testing problems.

Setup:

- Let $\{\mathbb{P}_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$ where $\mathcal{V} = \{1, \ldots, M\}$.
- The family induces a collection of parameters $\{\theta(\mathbb{P}_v)\}_{v\in\mathcal{V}}$.
- Choose $\{\mathbb{P}_v\}_{v\in\mathcal{V}}$ such that

$$\rho(\theta(\mathbb{P}_v), \theta(\mathbb{P}_{v'})) \geq 2\delta, \qquad \text{for all } v \neq v'.$$

- Suppose we observe $(X_i)_{i=1}^n$ is drawn from the n-fold product distribution, $\mathbb{P}_{v^*}^n$ for some $v^* \in \mathcal{V}$.
- Construct $\hat{\theta}(X_1, \ldots, X_n)$.

Testing problem:

- Based on $(X_i)_{i=1}^n$, test which of M hypothesis is true.

From Estimation to Testing

- For a measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$, the error probability is defined as
$$\max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi(X_1, \ldots, X_n) \neq v).$$

Minimum distance test:

$$\Psi^* = \arg\min_{v \in \mathcal{V}} \rho(\hat{\theta}, \theta(\mathbb{P}_v))$$

- $\rho(\hat{\theta}, \theta(\mathbb{P}_v)) < \delta \Longrightarrow \Psi^* = v$
- $\Psi^* \neq v \Longrightarrow \rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta$
- $\mathbb{P}_v^n(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta) \geq \mathbb{P}_v^n(\Psi^* \neq v)$

From Estimation to Testing

- For a measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$, the error probability is defined as
$$\max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi(X_1, \ldots, X_n) \neq v).$$

Minimum distance test:

$$\Psi^* = \arg\min_{v \in \mathcal{V}} \rho(\hat{\theta}, \theta(\mathbb{P}_v))$$

- $\rho(\hat{\theta}, \theta(\mathbb{P}_v)) < \delta \Longrightarrow \Psi^* = v$
- $\Psi^* \neq v \Longrightarrow \rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta$
- $\mathbb{P}_v^n(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta) \geq \mathbb{P}_v^n(\Psi^* \neq v)$

From Estimation to Testing

- For a measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$, the error probability is defined as
$$\max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi(X_1, \ldots, X_n) \neq v).$$

Minimum distance test:

$$\Psi^* = \arg\min_{v \in \mathcal{V}} \rho(\hat{\theta}, \theta(\mathbb{P}_v))$$

- $\rho(\hat{\theta}, \theta(\mathbb{P}_v)) < \delta \implies \Psi^* = v$
- $\Psi^* \neq v \implies \rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta$
- $\mathbb{P}_v^n(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta) \geq \mathbb{P}_v^n(\Psi^* \neq v)$

From Estimation to Testing

- For a measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$, the error probability is defined as

$$\max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi(X_1, \ldots, X_n) \neq v).$$

Minimum distance test:

$$\Psi^* = \arg\min_{v \in \mathcal{V}} \rho(\hat{\theta}, \theta(\mathbb{P}_v))$$

- $\rho(\hat{\theta}, \theta(\mathbb{P}_v)) < \delta \Longrightarrow \Psi^* = v$
- $\Psi^* \neq v \Longrightarrow \rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta$
- $\mathbb{P}_v^n(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta) \geq \mathbb{P}_v^n(\Psi^* \neq v)$

From Estimation to Testing

$$\begin{aligned}
\mathcal{M}_n(\theta(\mathcal{P})) &= \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right] \\
&\geq \delta \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n \left(\rho(\hat{\theta}, \theta(\mathbb{P})) \geq \delta\right) \\
&\geq \delta \inf_{\hat{\theta}} \max_{v\in\mathcal{V}} \mathbb{P}_v^n \left(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta\right) \\
&\geq \delta \underbrace{\inf_{\Psi} \max_{v\in\mathcal{V}} \mathbb{P}_v^n(\Psi \neq v)}_{\text{minimax probability of error}}
\end{aligned}$$

From Estimation to Testing

$$\mathcal{M}_n(\theta(\mathcal{P})) = \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$$

$$\geq \delta \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n\left(\rho(\hat{\theta}, \theta(\mathbb{P})) \geq \delta\right)$$

$$\geq \delta \inf_{\hat{\theta}} \max_{v\in\mathcal{V}} \mathbb{P}_v^n\left(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta\right)$$

$$\geq \delta \underbrace{\inf_{\Psi} \max_{v\in\mathcal{V}} \mathbb{P}_v^n(\Psi \neq v)}_{\text{minimax probability of error}}$$

From Estimation to Testing

$$
\begin{aligned}
\mathcal{M}_n(\theta(\mathcal{P})) &= \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right] \\
&\geq \delta \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n\left(\rho(\hat{\theta}, \theta(\mathbb{P})) \geq \delta\right) \\
&\geq \delta \inf_{\hat{\theta}} \max_{v\in\mathcal{V}} \mathbb{P}_v^n\left(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta\right) \\
&\geq \delta \underbrace{\inf_{\Psi} \max_{v\in\mathcal{V}} \mathbb{P}_v^n(\Psi \neq v)}_{\text{minimax probability of error}}
\end{aligned}
$$

From Estimation to Testing

$$\mathcal{M}_n(\theta(\mathcal{P})) = \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$$

$$\geq \delta \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n\left(\rho(\hat{\theta}, \theta(\mathbb{P})) \geq \delta\right)$$

$$\geq \delta \inf_{\hat{\theta}} \max_{v\in\mathcal{V}} \mathbb{P}^n_v\left(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta\right)$$

$$\geq \delta \underbrace{\inf_{\Psi} \max_{v\in\mathcal{V}} \mathbb{P}^n_v(\Psi \neq v)}_{\text{minimax probability of error}}$$

From Estimation to Testing

$$\mathcal{M}_n(\theta(\mathcal{P})) = \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$$
$$\geq \delta \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n\left(\rho(\hat{\theta}, \theta(\mathbb{P})) \geq \delta\right)$$
$$\geq \delta \inf_{\hat{\theta}} \max_{v\in\mathcal{V}} \mathbb{P}_v^n\left(\rho(\hat{\theta}, \theta(\mathbb{P}_v)) \geq \delta\right)$$
$$\geq \delta \underbrace{\inf_{\Psi} \max_{v\in\mathcal{V}} \mathbb{P}_v^n(\Psi \neq v)}_{\text{minimax probability of error}}$$

Minimax Probability of Error

Suppose $M = 2$, i.e., $\mathcal{V} = \{1, 2\}$. Then

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \inf_{\Psi} \left[\mathbb{P}_1^n(\Psi \neq 1) + \mathbb{P}_2^n(\Psi \neq 2)\right]$$

The minimizer is the likelihood ratio test and so

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \int \min(d\mathbb{P}_1^n, d\mathbb{P}_2^n) = \frac{1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}}{2}.$$

$$\mathcal{M}_n(\theta(\mathcal{P})) \geq \frac{\delta}{2}\left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}\right)$$

Recipe: Pick $\mathbb{P}_1$ and $\mathbb{P}_2$ in $\mathcal{P}$ such that $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV} \leq \frac{1}{2}$ and $\rho(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) \geq 2\delta$. (Le Cam, 1973)

General theme: The minimax risk is related to the distance between distributions.

Minimax Probability of Error

Suppose $M = 2$, i.e., $\mathcal{V} = \{1, 2\}$. Then

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \inf_{\Psi} \left[\mathbb{P}_1^n(\Psi \neq 1) + \mathbb{P}_2^n(\Psi \neq 2)\right]$$

The minimizer is the likelihood ratio test and so

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \int \min(d\mathbb{P}_1^n, d\mathbb{P}_2^n) = \frac{1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}}{2}.$$

$$\mathcal{M}_n(\theta(\mathcal{P})) \geq \frac{\delta}{2} \left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}\right)$$

Recipe: Pick $\mathbb{P}_1$ and $\mathbb{P}_2$ in $\mathcal{P}$ such that $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV} \leq \frac{1}{2}$ and $\rho(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) \geq 2\delta$. (Le Cam, 1973)

General theme: The minimax risk is related to the distance between distributions.

Minimax Probability of Error

Suppose $M = 2$, i.e., $\mathcal{V} = \{1, 2\}$. Then

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \inf_{\Psi} \left[\mathbb{P}_1^n(\Psi \neq 1) + \mathbb{P}_2^n(\Psi \neq 2)\right]$$

The minimizer is the likelihood ratio test and so

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \int \min(d\mathbb{P}_1^n, d\mathbb{P}_2^n) = \frac{1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}}{2}.$$

$$\mathcal{M}_n(\theta(\mathcal{P})) \geq \frac{\delta}{2}\left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}\right)$$

Recipe: Pick $\mathbb{P}_1$ and $\mathbb{P}_2$ in $\mathcal{P}$ such that $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV} \leq \frac{1}{2}$ and $\rho(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) \geq 2\delta$. (Le Cam, 1973)

General theme: The minimax risk is related to the distance between distributions.

Minimax Probability of Error

Suppose $M = 2$, i.e., $\mathcal{V} = \{1, 2\}$. Then

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \inf_{\Psi} \left[\mathbb{P}_1^n(\Psi \neq 1) + \mathbb{P}_2^n(\Psi \neq 2)\right]$$

The minimizer is the likelihood ratio test and so

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \int \min(d\mathbb{P}_1^n, d\mathbb{P}_2^n)$$
$$= \frac{1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}}{2}.$$

$$\mathcal{M}_n(\theta(\mathcal{P})) \geq \frac{\delta}{2} \left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}\right)$$

Recipe: Pick $\mathbb{P}_1$ and $\mathbb{P}_2$ in $\mathcal{P}$ such that $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV} \leq \frac{1}{2}$ and $\rho(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) \geq 2\delta$. (Le Cam, 1973)

General theme: The minimax risk is related to the distance between distributions.

Minimax Probability of Error

Suppose $M = 2$, i.e., $\mathcal{V} = \{1, 2\}$. Then

$$\inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) \geq \frac{1}{2} \inf_{\Psi} \left[\mathbb{P}_1^n(\Psi \neq 1) + \mathbb{P}_2^n(\Psi \neq 2) \right]$$

The minimizer is the likelihood ratio test and so

$$\begin{aligned} \inf_{\Psi} \max_{v \in \mathcal{V}} \mathbb{P}_v^n(\Psi \neq v) &\geq \frac{1}{2} \int \min(d\mathbb{P}_1^n, d\mathbb{P}_2^n) \\ &= \frac{1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}}{2}. \end{aligned}$$

$$\mathcal{M}_n(\theta(\mathcal{P})) \geq \frac{\delta}{2} \left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV}\right)$$

Recipe: Pick $\mathbb{P}_1$ and $\mathbb{P}_2$ in $\mathcal{P}$ such that $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{TV} \leq \frac{1}{2}$ and $\rho(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) \geq 2\delta$. (Le Cam, 1973)

General theme: The minimax risk is related to the distance between distributions.

Le Cam's Method

Theorem

Suppose there exists $\mathbb{P}_1$, $\mathbb{P}_2 \in \mathcal{P}$ *such that:*

- $\rho(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) \geq 2\delta > 0$;
- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) \leq \alpha < \infty$.

Then

$$\mathcal{M}_n(\theta(\mathcal{P})) \geq \delta \max\left(\frac{e^{-\alpha}}{4}, \frac{1 - \sqrt{\alpha/2}}{2}\right).$$

Strategy: Choose δ and guess two elements $\mathbb{P}_1$ and $\mathbb{P}_2$ so that the conditions are satisfied with α independent of n.

Main Results

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = p_1\delta_y + (1-p_1)\delta_z \qquad \text{and} \qquad \mathbb{P}_2 = p_2\delta_y + (1-p_2)\delta_z$$

where $y, z \in \mathbb{R}^d$, $p_1 > 0$ and $p_2 > 0$.

-

$$\begin{aligned}\rho^2(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) &= \|\mu_{\mathbb{P}_1} - \mu_{\mathbb{P}_2}\|^2_{\mathcal{H}_k} \\ &= 2(p_1-p_2)^2\left(1 - \exp\left(-\frac{\|y-z\|^2}{2\eta^2}\right)\right) \\ &\geq 2(p_1-p_2)^2\frac{\|y-z\|^2}{2\eta^2} \text{ if } \|y-z\|^2 \leq 2\eta^2.\end{aligned}$$

- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) \leq \frac{n(p_1-p_2)^2}{p_2(1-p_2)}$.
- Choose $p_2 = \frac{1}{2}$ and p_1 such that $(p_1-p_2)^2 = \frac{1}{9n}$; y, z such that $\frac{\|y-z\|^2}{2\eta^2} \geq \beta > 0$.

$$\delta = \sqrt{\frac{\beta}{9n}}$$

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = p_1\delta_y + (1-p_1)\delta_z \quad \text{and} \quad \mathbb{P}_2 = p_2\delta_y + (1-p_2)\delta_z$$

where $y, z \in \mathbb{R}^d$, $p_1 > 0$ and $p_2 > 0$.

- $$\begin{aligned}\rho^2(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) &= \|\mu_{\mathbb{P}_1} - \mu_{\mathbb{P}_2}\|^2_{\mathcal{H}_k} \\ &= 2(p_1-p_2)^2\left(1 - \exp\left(-\frac{\|y-z\|^2}{2\eta^2}\right)\right) \\ &\geq 2(p_1-p_2)^2 \frac{\|y-z\|^2}{2\eta^2} \text{ if } \|y-z\|^2 \leq 2\eta^2.\end{aligned}$$
- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) \leq \frac{n(p_1-p_2)^2}{p_2(1-p_2)}$.
- Choose $p_2 = \frac{1}{2}$ and p_1 such that $(p_1-p_2)^2 = \frac{1}{9n}$; y, z such that $\frac{\|y-z\|^2}{2\eta^2} \geq \beta > 0$.

$$\delta = \sqrt{\frac{\beta}{9n}}$$

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = p_1\delta_y + (1-p_1)\delta_z \qquad \text{and} \qquad \mathbb{P}_2 = p_2\delta_y + (1-p_2)\delta_z$$

where $y, z \in \mathbb{R}^d$, $p_1 > 0$ and $p_2 > 0$.

- $$\begin{aligned}\rho^2(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) &= \|\mu_{\mathbb{P}_1} - \mu_{\mathbb{P}_2}\|^2_{\mathcal{H}_k} \\ &= 2(p_1 - p_2)^2\left(1 - \exp\left(-\frac{\|y-z\|^2}{2\eta^2}\right)\right) \\ &\geq 2(p_1-p_2)^2 \frac{\|y-z\|^2}{2\eta^2} \text{ if } \|y-z\|^2 \leq 2\eta^2.\end{aligned}$$
- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) \leq \frac{n(p_1-p_2)^2}{p_2(1-p_2)}$.
- Choose $p_2 = \frac{1}{2}$ and p_1 such that $(p_1 - p_2)^2 = \frac{1}{9n}$; y, z such that $\frac{\|y-z\|^2}{2\eta^2} \geq \beta > 0$.

$$\delta = \sqrt{\frac{\beta}{9n}}$$

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = p_1\delta_y + (1-p_1)\delta_z \qquad \text{and} \qquad \mathbb{P}_2 = p_2\delta_y + (1-p_2)\delta_z$$

where $y, z \in \mathbb{R}^d$, $p_1 > 0$ and $p_2 > 0$.

- $$\begin{aligned}\rho^2(\theta(\mathbb{P}_1),\theta(\mathbb{P}_2)) &= \|\mu_{\mathbb{P}_1} - \mu_{\mathbb{P}_2}\|^2_{\mathcal{H}_k} \\ &= 2(p_1-p_2)^2\left(1-\exp\left(-\frac{\|y-z\|^2}{2\eta^2}\right)\right) \\ &\geq 2(p_1-p_2)^2\frac{\|y-z\|^2}{2\eta^2} \text{ if } \|y-z\|^2 \leq 2\eta^2.\end{aligned}$$
- $KL(\mathbb{P}_1^n\|\mathbb{P}_2^n) \leq \frac{n(p_1-p_2)^2}{p_2(1-p_2)}$.
- Choose $p_2 = \frac{1}{2}$ and p_1 such that $(p_1-p_2)^2 = \frac{1}{9n}$; y, z such that $\frac{\|y-z\|^2}{2\eta^2} \geq \beta > 0$.

$$\delta = \sqrt{\frac{\beta}{9n}}$$

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = p_1\delta_y + (1-p_1)\delta_z \qquad \text{and} \qquad \mathbb{P}_2 = p_2\delta_y + (1-p_2)\delta_z$$

where $y, z \in \mathbb{R}^d$, $p_1 > 0$ and $p_2 > 0$.

- $$\begin{aligned}\rho^2(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) &= \|\mu_{\mathbb{P}_1} - \mu_{\mathbb{P}_2}\|^2_{\mathcal{H}_k} \\ &= 2(p_1-p_2)^2\left(1 - \exp\left(-\frac{\|y-z\|^2}{2\eta^2}\right)\right) \\ &\geq 2(p_1-p_2)^2\frac{\|y-z\|^2}{2\eta^2} \text{ if } \|y-z\|^2 \leq 2\eta^2.\end{aligned}$$
- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) \leq \frac{n(p_1-p_2)^2}{p_2(1-p_2)}$.
- Choose $p_2 = \frac{1}{2}$ and p_1 such that $(p_1-p_2)^2 = \frac{1}{9n}$; y, z such that $\frac{\|y-z\|^2}{2\eta^2} \geq \beta > 0$.

$$\delta = \sqrt{\frac{\beta}{9n}}$$

Gaussian Kernel

In words:

- If $\mathcal{P}$ is the set of all discrete distributions, then

$$\mathcal{M}_n(\mu(\mathcal{P})) \geq \frac{1}{12}\sqrt{\frac{\beta}{n}}.$$

- For any estimator $\hat{\theta}$, there always exists a discrete distribution, $\mathbb{P}$ such that $\mu_{\mathbb{P}}$ cannot be estimated at a rate faster than $n^{-1/2}$.

Is such a result true if $\mathcal{P}$ is a class of distributions with smooth density?

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = N(\mu_1, \sigma^2 I) \qquad \text{and} \qquad \mathbb{P}_2 = N(\mu_2, \sigma^2 I)$$

where $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\sigma > 0$.

- $$\begin{aligned}\rho^2(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) &= 2\left(\frac{2\eta^2}{2\eta^2 + 4\sigma^2}\right)^{\frac{d}{2}}\left(1 - \exp\left(-\frac{\|\mu_1 - \mu_2\|^2}{2\eta^2 + 4\sigma^2}\right)\right) \\ &\geq \left(\frac{2\eta^2}{2\eta^2 + 4\sigma^2}\right)^{\frac{d}{2}} \frac{\|\mu_1 - \mu_2\|^2}{2\eta^2 + 4\sigma^2} \text{ if } \|\mu_1 - \mu_2\|^2 \leq 2\eta^2 + 4\sigma^2.\end{aligned}$$
- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) = \frac{n\|\mu_1 - \mu_2\|^2}{2\sigma^2}$.
- Choose μ_1 and μ_2 such that $\|\mu_1 - \mu_2\|^2 \leq \frac{2\sigma^2\alpha}{n}$ and $\sigma^2 = \frac{\eta^2}{2d}$.

$$\delta = \sqrt{\frac{C'}{n}}$$

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = N(\mu_1, \sigma^2 I) \qquad \text{and} \qquad \mathbb{P}_2 = N(\mu_2, \sigma^2 I)$$

where $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\sigma > 0$.

- $$\rho^2(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) = 2\left(\frac{2\eta^2}{2\eta^2 + 4\sigma^2}\right)^{\frac{d}{2}} \left(1 - \exp\left(-\frac{\|\mu_1 - \mu_2\|^2}{2\eta^2 + 4\sigma^2}\right)\right)$$
  $$\geq \left(\frac{2\eta^2}{2\eta^2 + 4\sigma^2}\right)^{\frac{d}{2}} \frac{\|\mu_1 - \mu_2\|^2}{2\eta^2 + 4\sigma^2} \text{ if } \|\mu_1 - \mu_2\|^2 \leq 2\eta^2 + 4\sigma^2.$$
- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) = \frac{n\|\mu_1 - \mu_2\|^2}{2\sigma^2}$.
- Choose μ_1 and μ_2 such that $\|\mu_1 - \mu_2\|^2 \leq \frac{2\sigma^2\alpha}{n}$ and $\sigma^2 = \frac{\eta^2}{2d}$.

$$\delta = \sqrt{\frac{C'}{n}}$$

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = N(\mu_1, \sigma^2 I) \qquad \text{and} \qquad \mathbb{P}_2 = N(\mu_2, \sigma^2 I)$$

where $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\sigma > 0$.

- $$\begin{aligned}\rho^2(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) &= 2\left(\frac{2\eta^2}{2\eta^2 + 4\sigma^2}\right)^{\frac{d}{2}} \left(1 - \exp\left(-\frac{\|\mu_1 - \mu_2\|^2}{2\eta^2 + 4\sigma^2}\right)\right) \\ &\geq \left(\frac{2\eta^2}{2\eta^2 + 4\sigma^2}\right)^{\frac{d}{2}} \frac{\|\mu_1 - \mu_2\|^2}{2\eta^2 + 4\sigma^2} \text{ if } \|\mu_1 - \mu_2\|^2 \leq 2\eta^2 + 4\sigma^2.\end{aligned}$$
- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) = \frac{n\|\mu_1 - \mu_2\|^2}{2\sigma^2}$.
- Choose μ_1 and μ_2 such that $\|\mu_1 - \mu_2\|^2 \leq \frac{2\sigma^2 \alpha}{n}$ and $\sigma^2 = \frac{\eta^2}{2d}$.

$$\delta = \sqrt{\frac{C'}{n}}$$

Gaussian Kernel

Let $k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta>0$. Choose

$$\mathbb{P}_1=N(\mu_1,\sigma^2 I) \qquad \text{and} \qquad \mathbb{P}_2=N(\mu_2,\sigma^2 I)$$

where $\mu_1,\mu_2\in\mathbb{R}^d$ and $\sigma>0$.

-

$$\rho^2(\theta(\mathbb{P}_1),\theta(\mathbb{P}_2)) = 2\left(\frac{2\eta^2}{2\eta^2+4\sigma^2}\right)^{\frac{d}{2}}\left(1-\exp\left(-\frac{\|\mu_1-\mu_2\|^2}{2\eta^2+4\sigma^2}\right)\right)$$
$$\geq \left(\frac{2\eta^2}{2\eta^2+4\sigma^2}\right)^{\frac{d}{2}}\frac{\|\mu_1-\mu_2\|^2}{2\eta^2+4\sigma^2} \text{ if } \|\mu_1-\mu_2\|^2\leq 2\eta^2+4\sigma^2.$$

- $KL(\mathbb{P}_1^n\|\mathbb{P}_2^n)=\frac{n\|\mu_1-\mu_2\|^2}{2\sigma^2}$.
- Choose μ_1 and μ_2 such that $\|\mu_1-\mu_2\|^2\leq\frac{2\sigma^2\alpha}{n}$ and $\sigma^2=\frac{\eta^2}{2d}$.

$$\delta=\sqrt{\frac{C'}{n}}$$

Gaussian Kernel

Let $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\eta^2}\right)$, $\eta > 0$. Choose

$$\mathbb{P}_1 = N(\mu_1, \sigma^2 I) \qquad \text{and} \qquad \mathbb{P}_2 = N(\mu_2, \sigma^2 I)$$

where $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\sigma > 0$.

- $$\rho^2(\theta(\mathbb{P}_1), \theta(\mathbb{P}_2)) = 2\left(\frac{2\eta^2}{2\eta^2 + 4\sigma^2}\right)^{\frac{d}{2}} \left(1 - \exp\left(-\frac{\|\mu_1 - \mu_2\|^2}{2\eta^2 + 4\sigma^2}\right)\right)$$
$$\geq \left(\frac{2\eta^2}{2\eta^2 + 4\sigma^2}\right)^{\frac{d}{2}} \frac{\|\mu_1 - \mu_2\|^2}{2\eta^2 + 4\sigma^2} \text{ if } \|\mu_1 - \mu_2\|^2 \leq 2\eta^2 + 4\sigma^2.$$
- $KL(\mathbb{P}_1^n \| \mathbb{P}_2^n) = \frac{n\|\mu_1 - \mu_2\|^2}{2\sigma^2}$.
- Choose μ_1 and μ_2 such that $\|\mu_1 - \mu_2\|^2 \leq \frac{2\sigma^2\alpha}{n}$ and $\sigma^2 = \frac{\eta^2}{2d}$.

$$\delta = \sqrt{\frac{C'}{n}}$$

General Result

Theorem

Suppose $\mathcal{P}$ is the set of all discrete distributions on $\mathbb{R}^d$. Let k be shift-invariant, i.e., $k(x,y) = \psi(x-y)$ with $\psi \in C_b(\mathbb{R}^d)$ and characteristic. Assume there exists $x_0 \in \mathbb{R}^d$ and $\beta > 0$ such that

$$\psi(0) - \psi(x_0) \geq \beta.$$

Then

$$\mathcal{M}_n(\mu(\mathcal{P})) \geq \frac{1}{24}\sqrt{\frac{2\beta}{n}}.$$

General Result

Theorem

Suppose $\mathcal{P}$ is the set of all distributions with infinitely differentiable densities on $\mathbb{R}^d$. Let k be shift-invariant, i.e., $k(x, y) = \psi(x - y)$ with $\psi \in C_b(\mathbb{R}^d)$ and characteristic. Then there exists constants c_ψ, $\epsilon_\psi > 0$ depending only on ψ such that for any $n \geq \frac{1}{\epsilon_\psi}$:

$$\mathcal{M}_n(\mu(\mathcal{P})) \geq \frac{1}{8}\sqrt{\frac{c_\psi}{2n}}.$$

Idea: Exactly same as that of the Gaussian kernel. But the crucial work is in showing that there exists constants ϵ_{ψ,σ^2} and c_{ψ,σ^2} such that if

$$\|\mu_1 - \mu_2\|^2 \leq \epsilon_{\psi,\sigma^2}$$

then

$$\|\mu(N(\mu_1, \sigma^2 I)) - \mu(N(\mu_2, \sigma^2 I))\|_{\mathcal{H}_k} \geq c_{\psi,\sigma^2}\|\mu_1 - \mu_2\|.$$

Summary

- Mean embedding of distributions is popular in various applications.
- Various estimators of kernel mean are available.
- We provide a theoretical justification for using these estimators, particularly the empirical estimator.
- The empirical estimator of the mean embedding is minimax rate optimal with rate $n^{-1/2}$.

Thank You