Non-parameteric Estimation of Integral Probability Metrics

Bharath K. Sriperumbudur[⋆], Kenji Fukumizu[†], Arthur Gretton[‡,×], Bernhard Schölkopf[×] and Gert R. G. Lanckriet[⋆]

[⋆]UC San Diego [†]The Institute of Statistical Mathematics
[‡] CMU [×]MPI for Biological Cybernetics

ISIT 2010

Probability Metrics

- X : measurable space.
- $\mathscr{P}$: set of all probability measures defined on X.
- $\gamma : \mathscr{P} \times \mathscr{P} \to \mathbb{R}^+$ is a notion of distance on $\mathscr{P}$, called the *probability metric*.

Popular example: ϕ-divergence

$$D_\phi(\mathbb{P}, \mathbb{Q}) := \begin{cases} \int_X \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q}, & \mathbb{P} \ll \mathbb{Q} \\ +\infty, & \text{otherwise} \end{cases},$$

where $\phi : [0, \infty) \to (-\infty, \infty]$ is a convex function.

Appropriate choice of ϕ: Kullback-Leibler divergence, Jensen-Shannon divergence, Total-variation distance, Hellinger distance, χ^2-distance.

Probability Metrics

- X : measurable space.
- $\mathscr{P}$: set of all probability measures defined on X.
- $\gamma : \mathscr{P} \times \mathscr{P} \to \mathbb{R}^+$ is a notion of distance on $\mathscr{P}$, called the *probability metric*.

Popular example: ϕ-divergence

$$D_\phi(\mathbb{P},\mathbb{Q}) := \begin{cases} \int_X \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q}, & \mathbb{P} \ll \mathbb{Q} \\ +\infty, & \text{otherwise} \end{cases},$$

where $\phi : [0,\infty) \to (-\infty,\infty]$ is a convex function.

Appropriate choice of ϕ: Kullback-Leibler divergence, Jensen-Shannon divergence, Total-variation distance, Hellinger distance, χ^2-distance.

Applications

Two-sample problem:

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, respectively.
- *Determine:* are $\mathbb{P}$ and $\mathbb{Q}$ different?

Applications

Two-sample problem:

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, respectively.
- *Determine:* are $\mathbb{P}$ and $\mathbb{Q}$ different?
- $\gamma(\mathbb{P}, \mathbb{Q})$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

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

- *Test:* Say H_0 if $\widehat{\gamma}(\mathbb{P}, \mathbb{Q}) < \varepsilon$. Otherwise say H_1.

Applications

Two-sample problem:

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, respectively.
- *Determine:* are $\mathbb{P}$ and $\mathbb{Q}$ different?
- $\gamma(\mathbb{P}, \mathbb{Q})$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

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

- *Test:* Say H_0 if $\widehat{\gamma}(\mathbb{P}, \mathbb{Q}) < \varepsilon$. Otherwise say H_1.

Other applications:

- *Hypothesis testing* : Independence test, Goodness of fit test, etc.
- Limit theorems (central limit theorem), density estimation, etc.

Estimation of $D_\phi(\mathbb{P}, \mathbb{Q})$

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, estimate $D_\phi(\mathbb{P}, \mathbb{Q})$.
- Well-studied for $\phi(t) = t \log t$, $t \in [0, \infty)$, i.e., Kullback-Liebler divergence.
- *Approaches:*
  - Histogram estimator based on space partitioning scheme [Wang et al., 2005].
  - M-estimation based on the variational characterization [Nguyen et al., 2008],

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \sup_{f:X\to\mathbb{R}} \left[\int_X f\, d\mathbb{P} - \int_X \phi^*(f)\, d\mathbb{Q} \right],$$

where ϕ^* is the convex conjugate of ϕ.

Estimation of $D_\phi(\mathbb{P}, \mathbb{Q})$

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, estimate $D_\phi(\mathbb{P}, \mathbb{Q})$.
- Well-studied for $\phi(t) = t \log t$, $t \in [0, \infty)$, i.e., Kullback-Liebler divergence.
- *Approaches:*
  - Histogram estimator based on space partitioning scheme [Wang et al., 2005].
  - M-estimation based on the variational characterization [Nguyen et al., 2008],

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \sup_{f : X \to \mathbb{R}} \left[\int_X f \, d\mathbb{P} - \int_X \phi^*(f) \, d\mathbb{Q} \right],$$

where ϕ^* is the convex conjugate of ϕ.

Estimation of $D_\phi(\mathbb{P}, \mathbb{Q})$

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, estimate $D_\phi(\mathbb{P}, \mathbb{Q})$.
- Well-studied for $\phi(t) = t \log t$, $t \in [0, \infty)$, i.e., Kullback-Liebler divergence.
- *Approaches:*
  - Histogram estimator based on space partitioning scheme [Wang et al., 2005].
  - M-estimation based on the variational characterization [Nguyen et al., 2008],

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \sup_{f : X \to \mathbb{R}} \left[\int_X f \, d\mathbb{P} - \int_X \phi^*(f) \, d\mathbb{Q} \right],$$

where ϕ^* is the convex conjugate of ϕ.

Estimation of $D_\phi(\mathbb{P}, \mathbb{Q})$

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, estimate $D_\phi(\mathbb{P}, \mathbb{Q})$.
- Well-studied for $\phi(t) = t \log t$, $t \in [0, \infty)$, i.e., Kullback-Liebler divergence.
- *Approaches:*
  - Histogram estimator based on space partitioning scheme [Wang et al., 2005].
  - M-estimation based on the variational characterization [Nguyen et al., 2008],

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \sup_{f: X \to \mathbb{R}} \left[\int_X f \, d\mathbb{P} - \int_X \phi^*(f) \, d\mathbb{Q} \right],$$

where ϕ^* is the convex conjugate of ϕ.

Properties of Estimators

- Computability
- Consistency
- Rate of convergence

Issues:

- Though the estimators of $D_\phi(\mathbb{P}, \mathbb{Q})$ are consistent, their *rate of convergence can be arbitrarily slow* depending on $\mathbb{P}$ and $\mathbb{Q}$.
- Let $X \subset \mathbb{R}^d$. For large d, the estimator proposed by [Wang et al., 2005] is *computationally inefficient.*

Properties of Estimators

- Computability
- Consistency
- Rate of convergence

Issues:

- Though the estimators of $D_\phi(\mathbb{P}, \mathbb{Q})$ are consistent, their *rate of convergence can be arbitrarily slow* depending on $\mathbb{P}$ and $\mathbb{Q}$.
- Let $X \subset \mathbb{R}^d$. For large d, the estimator proposed by [Wang et al., 2005] is *computationally inefficient.*

Integral Probability Metrics

- The *integral probability metric* [Müller, 1997] between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left| \int_X f\, d\mathbb{P} - \int_X f\, d\mathbb{Q} \right|.$$

- Many popular probability metrics can be obtained by appropriately choosing $\mathcal{F}$.
  - *Total variation distance* : $\mathcal{F} = \left\{f : \|f\|_\infty := \sup_{x\in X} |f(x)| \le 1\right\}$.
  - *Wasserstein distance* : $\mathcal{F} = \left\{f : \|f\|_L := \sup_{x\neq y\in X} \frac{|f(x)-f(y)|}{\rho(x,y)} \le 1\right\}$.
  - *Dudley metric* : $\mathcal{F} = \{f : \|f\|_L + \|f\|_\infty \le 1\}$.
  - *L^p metric* : $\mathcal{F} = \{f : \|f\|_{L^p(X,\mu)} := (\int_X |f|^p\, d\mu)^{1/p} \le 1,\ 1 \le p < \infty\}$.
- *well-studied* in probability theory, mass transporation problems, etc.

Outline

- *Relation between $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ and $D_\phi(\mathbb{P},\mathbb{Q})$*
- Estimation of $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$
- Consistency analysis and rate of convergence

$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ *vs.* $D_\phi(\mathbb{P},\mathbb{Q})$

$$D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left[\int_X f\,d\mathbb{P} - \int_X \phi^*(f)\,d\mathbb{Q}\right]$$

- $D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ if $\mathcal{F}$ is the set of all real-valued measurable functions on X.
- $D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ if $\phi(t) = \begin{cases} 0, & t=1 \\ +\infty, & t\neq 1 \end{cases}$.
- $D_\phi(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ if and only if any one of the following hold:
  - *(i)* $\mathcal{F} = \{f : \|f\|_\infty \le \frac{\beta-\alpha}{2}\}$ and $\phi(t) = \begin{cases} \alpha(t-1), & 0\le t\le 1 \\ \beta(t-1), & t\ge 1 \end{cases}$ for some $\alpha<\beta<\infty$.
  - *(ii)* $\mathcal{F} = \{f\ :\ f = c,\ c\in\mathbb{R}\}$, $\phi(t) = \alpha(t-1)$, $t\ge 0$, $\alpha\in\mathbb{R}$
- *Total-variation is the only ϕ-divergence that is also an integral probability metric.*

$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ *vs.* $D_\phi(\mathbb{P},\mathbb{Q})$

$$D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}}\left[\int_X f\,d\mathbb{P} - \int_X \phi^*(f)\,d\mathbb{Q}\right]$$

- $D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ if $\mathcal{F}$ is the set of all real-valued measurable functions on X.
- $D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ if $\phi(t) = \begin{cases} 0, & t=1 \\ +\infty, & t\neq 1 \end{cases}$.
- $D_\phi(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ if and only if any one of the following hold:

  (i) $\mathcal{F} = \{f : \|f\|_\infty \leq \frac{\beta-\alpha}{2}\}$ and $\phi(t) = \begin{cases} \alpha(t-1), & 0\leq t\leq 1 \\ \beta(t-1), & t\geq 1 \end{cases}$ for some $\alpha<\beta<\infty$.

  (ii) $\mathcal{F} = \{f\ :\ f=c,\ c\in\mathbb{R}\}$, $\phi(t)=\alpha(t-1)$, $t\geq 0$, $\alpha\in\mathbb{R}$
- *Total-variation is the only ϕ-divergence that is also an integral probability metric.*

$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ *vs.* $D_\phi(\mathbb{P},\mathbb{Q})$

$$D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left[\int_X f\,d\mathbb{P} - \int_X \phi^*(f)\,d\mathbb{Q}\right]$$

- $D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ if $\mathcal{F}$ is the set of all real-valued measurable functions on X.
- $D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ if $\phi(t) = \begin{cases} 0, & t=1 \\ +\infty, & t\neq 1 \end{cases}$.
- $D_\phi(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ if and only if any one of the following hold:

  (i) $\mathcal{F} = \{f : \|f\|_\infty \leq \frac{\beta-\alpha}{2}\}$ and $\phi(t) = \begin{cases} \alpha(t-1), & 0\leq t\leq 1 \\ \beta(t-1), & t\geq 1 \end{cases}$ for some $\alpha < \beta < \infty$.

  (ii) $\mathcal{F} = \{f\ :\ f = c,\ c\in\mathbb{R}\}$, $\phi(t) = \alpha(t-1)$, $t\geq 0$, $\alpha\in\mathbb{R}$

- *Total-variation is the only ϕ-divergence that is also an integral probability metric.*

$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ *vs.* $D_\phi(\mathbb{P},\mathbb{Q})$

$$D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left[\int_X f\, d\mathbb{P} - \int_X \phi^*(f)\, d\mathbb{Q}\right]$$

- $D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ if $\mathcal{F}$ is the set of all real-valued measurable functions on X.
- $D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ if $\phi(t) = \begin{cases} 0, & t=1 \\ +\infty, & t\neq 1 \end{cases}$.
- $D_\phi(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ if and only if any one of the following hold:

  (i) $\mathcal{F} = \{f : \|f\|_\infty \leq \frac{\beta-\alpha}{2}\}$ and $\phi(t) = \begin{cases} \alpha(t-1), & 0\leq t\leq 1 \\ \beta(t-1), & t\geq 1 \end{cases}$ for some $\alpha < \beta < \infty$.

  (ii) $\mathcal{F} = \{f \,:\, f = c,\, c\in\mathbb{R}\}$, $\phi(t) = \alpha(t-1)$, $t\geq 0$, $\alpha\in\mathbb{R}$

- *Total-variation is the only ϕ-divergence that is also an integral probability metric.*

Outline

- Relation between $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ and $D_\phi(\mathbb{P}, \mathbb{Q})$
- *Estimation of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$*
- Consistency analysis and rate of convergence

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, estimate $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$.

- *Estimator:*

$$\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sup_{f \in \mathcal{F}} \left[\frac{1}{m} \sum_{i=1}^{m} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right],$$

  where $\mathbb{P}_m := \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i}$ and $\mathbb{Q}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$.

- *Computability:* Possible for certain choices of $\mathcal{F}$.
  - $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_L \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_L + \|f\|_\infty \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ where $\mathcal{H}$ is a *reproducing kernel Hilbert space.*

- *Consistency and rate of convergence:* determined by the "size" of $\mathcal{F}$.

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, estimate $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$.

- *Estimator:*

$$\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sup_{f \in \mathcal{F}} \left[\frac{1}{m} \sum_{i=1}^{m} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right],$$

where $\mathbb{P}_m := \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i}$ and $\mathbb{Q}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$.

- *Computability:* Possible for certain choices of $\mathcal{F}$.
  - $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_L \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_L + \|f\|_\infty \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ where $\mathcal{H}$ is a *reproducing kernel Hilbert space.*

- *Consistency and rate of convergence:* determined by the "size" of $\mathcal{F}$.

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, estimate $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$.
- *Estimator:*

$$\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sup_{f \in \mathcal{F}} \left[\frac{1}{m} \sum_{i=1}^{m} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right],$$

  where $\mathbb{P}_m := \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i}$ and $\mathbb{Q}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$.
- *Computability:* Possible for certain choices of $\mathcal{F}$.
  - $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_L \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_L + \|f\|_\infty \leq 1\}$
  - $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ where $\mathcal{H}$ is a *reproducing kernel Hilbert space.*
- *Consistency and rate of convergence:* determined by the “size” of $\mathcal{F}$.

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$

$V := \{X_1,\ldots,X_m,Y_1,\ldots,Y_n\}$, $S := \{\frac{1}{m},\ldots,\frac{1}{m},-\frac{1}{n},\ldots,-\frac{1}{n}\}$,
$N := m+n$.

Theorem

- $\mathcal{F}=\{f:\|f\|_L\le 1\}$: $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)=\sum_{i=1}^N S_i a_i^*$, where

$$\{a_i^*\}_{i=1}^N = \arg\max\left\{\sum_{i=1}^N S_i a_i \,:\, -\rho(V_i,V_j)\le a_i - a_j \le \rho(V_i,V_j), \forall\, i,j\right\}.$$

- $\mathcal{F}=\{f:\|f\|_L+\|f\|_\infty\le 1\}$: $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)=\sum_{i=1}^N S_i b_i^*$, where

$$\begin{aligned}\{b_i^*\}_{i=1}^N = \arg\max_{b_1,\ldots,b_N,e,c} \quad & \sum_{i=1}^N S_i b_i \\ s.t. \quad & -e\,\rho(V_i,V_j)\le b_i-b_j\le e\,\rho(V_i,V_j), \forall\, i,j \\ & -c\le b_i\le c, \forall\, i,\ e+c\le 1.\end{aligned}$$

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$

$V := \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$, $S := \{\frac{1}{m}, \ldots, \frac{1}{m}, -\frac{1}{n}, \ldots, -\frac{1}{n}\}$,
$N := m + n$.

Theorem

- $\mathcal{F} = \{f : \|f\|_L \leq 1\}$: $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^{N} S_i a_i^*$, where

$$\{a_i^*\}_{i=1}^N = \arg\max\left\{\sum_{i=1}^{N} S_i a_i \;:\; -\rho(V_i, V_j) \leq a_i - a_j \leq \rho(V_i, V_j), \forall\, i, j\right\}.$$

- $\mathcal{F} = \{f : \|f\|_L + \|f\|_\infty \leq 1\}$: $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^{N} S_i b_i^*$, where

$$\begin{aligned} \{b_i^*\}_{i=1}^N = \arg \max_{b_1, \ldots, b_N, e, c} \quad & \sum_{i=1}^{N} S_i b_i \\ s.t. \quad & -e\,\rho(V_i, V_j) \leq b_i - b_j \leq e\,\rho(V_i, V_j),\ \forall\, i, j \\ & -c \leq b_i \leq c,\ \forall\, i,\ e + c \leq 1. \end{aligned}$$

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$

$\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

Definition

A Hilbert space $\mathcal{H}$ is said to be an RKHS if the evaluation functionals ($\delta_x(f) = f(x)$, $x \in X$, $f \in \mathcal{H}$) are bounded and continuous.

- There exists a unique kernel, $k : X \times X \to \mathbb{R}$ such that $\forall x \in X, \forall f \in \mathcal{H}, \langle f, k(\cdot,x)\rangle_{\mathcal{H}} = f(x)$.
- k is the *reproducing kernel* (r.k.) of $\mathcal{H}$ as $k(x,y) = \langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}}$, $x, y \in X$.
- Every r.k. is a *positive definite function.*
- For every positive definite function, k on $X \times X$, there exists a unique RKHS, $\mathcal{H}$ as k as its r.k.
- *Example:* $k(x,y) = e^{-|x-y|}$, $x, y \in \mathbb{R}$ induces a Sobolev space.

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$

$\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

Definition

A Hilbert space $\mathcal{H}$ is said to be an RKHS if the evaluation functionals $(\delta_x(f) = f(x)$, $x \in X$, $f \in \mathcal{H})$ are bounded and continuous.

- There exists a unique kernel, $k : X \times X \to \mathbb{R}$ such that $\forall x \in X, \forall f \in \mathcal{H}, \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- k is the *reproducing kernel* (r.k.) of $\mathcal{H}$ as $k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}}$, $x, y \in X$.
- Every r.k. is a *positive definite function.*
- For every positive definite function, k on $X \times X$, there exists a unique RKHS, $\mathcal{H}$ as k as its r.k.
- *Example:* $k(x, y) = e^{-|x-y|}$, $x, y \in \mathbb{R}$ induces a Sobolev space.

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$

$\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

Definition

A Hilbert space $\mathcal{H}$ is said to be an RKHS if the evaluation functionals ($\delta_x(f) = f(x)$, $x \in X$, $f \in \mathcal{H}$) are bounded and continuous.

- There exists a unique kernel, $k : X \times X \to \mathbb{R}$ such that $\forall x \in X, \forall f \in \mathcal{H}, \langle f, k(\cdot,x)\rangle_{\mathcal{H}} = f(x)$.
- k is the *reproducing kernel* (r.k.) of $\mathcal{H}$ as $k(x,y) = \langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}}$, $x, y \in X$.
- Every r.k. is a *positive definite function.*
- For every positive definite function, k on $X \times X$, there exists a unique RKHS, $\mathcal{H}$ as k as its r.k.
- *Example:* $k(x,y) = e^{-|x-y|}$, $x, y \in \mathbb{R}$ induces a Sobolev space.

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$

$\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

Definition

A Hilbert space $\mathcal{H}$ is said to be an RKHS if the evaluation functionals $(\delta_x(f) = f(x),\ x \in X,\ f \in \mathcal{H})$ are bounded and continuous.

- There exists a unique kernel, $k : X \times X \to \mathbb{R}$ such that $\forall x \in X, \forall f \in \mathcal{H}, \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- k is the *reproducing kernel* (r.k.) of $\mathcal{H}$ as $k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}},\ x, y \in X$.
- Every r.k. is a *positive definite function.*
- For every positive definite function, k on $X \times X$, there exists a unique RKHS, $\mathcal{H}$ as k as its r.k.
- *Example:* $k(x, y) = e^{-|x-y|},\ x, y \in \mathbb{R}$ induces a Sobolev space.

Estimation of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$

$V := \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$, $S := \{\frac{1}{m}, \ldots, \frac{1}{m}, -\frac{1}{n}, \ldots, -\frac{1}{n}\}$,
$N := m + n$.

Theorem

Let $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ *with* k *being bounded and measurable. Then*

$$\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sqrt{\sum_{i,j=1}^{N} S_i S_j k(V_i, V_j)}.$$

Outline

- Relation between $\gamma_{\mathfrak{F}}(\mathbb{P}, \mathbb{Q})$ and $D_\phi(\mathbb{P}, \mathbb{Q})$
- Estimation of $\gamma_{\mathfrak{F}}(\mathbb{P}, \mathbb{Q})$
- *Consistency analysis and rate of convergence*

Consistency and Rate of Convergence

Theorem

Suppose $\mathcal{F}$ *be such that* $\nu := \sup_{f\in\mathcal{F},x\in X}|f(x)| < \infty$. *Fix* $\delta \in (0,1)$. *Then with probability* $1-\delta$ *over the choice of samples,* $\{X_i\}_{i=1}^m$ *and* $\{Y_i\}_{i=1}^n$, *the following holds:*

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| \le \sqrt{18\nu^2\log\frac{4}{\delta}}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right) + 2R_m(\mathcal{F};\{X_i\}) + 2R_n(\mathcal{F};\{Y_i\}),$$

where

$$R_m(\mathcal{F};\{x_i\}_{i=1}^m) := \mathbb{E}_\sigma \sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right|,$$

is called the Rademacher complexity of $\mathcal{F}$ *and* $\{\sigma_i\}$ *are independent Rademacher random variables defined as* $\sigma_i = 2B_i - 1$, *with* $\{B_i\}$ *being Bernoulli random variables.*

Consistency and Rate of Convergence

Note that if $R_m(\mathfrak{F}; \{X_i\}_{i=1}^m) = O_{\mathbb{P}}(r_m)$ and $R_n(\mathfrak{F}; \{Y_i\}_{i=1}^n) = O_{\mathbb{Q}}(r_n)$, then

$$|\gamma_{\mathfrak{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathfrak{F}}(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_m \vee \ m^{-1/2} + r_n \vee \ n^{-1/2}),$$

where $a \vee \ b := \max(a, b)$.

Theorem ([von Luxburg and Bousquet, 2004])

For every $\epsilon > 0$, the following holds:

$$R_m(\mathfrak{F}; \{x_i\}_{i=1}^m) \leq 2\epsilon + \frac{4\sqrt{2}}{m} \int_{\epsilon/4}^{\infty} \sqrt{\log \mathcal{N}(\tau, \mathfrak{F}, L^2(\mathbb{P}_m))}\, d\tau.$$

Consistency and Rate of Convergence

Corollary

- *Let X be a bounded subset of $(\mathbb{R}^d, \|\cdot\|_s)$ for some $1 \le s \le \infty$. Then, for $\mathcal{F} = \{f : \|f\|_L \le 1\}$ and $\mathcal{F} = \{f : \|f\|_\infty + \|f\|_L \le 1\}$, we have*

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_m + r_n)$$

where

$$r_m = \begin{cases} m^{-1/2}\log m, & d = 1 \\ m^{-1/(d+1)}, & d \ge 2 \end{cases}.$$

In addition if X is a bounded, convex subset of $(\mathbb{R}^d, \|\cdot\|_s)$ with non-empty interior, then

$$r_m = \begin{cases} m^{-1/2}, & d = 1 \\ m^{-1/2}\log m, & d = 2 \\ m^{-1/d}, & d > 2 \end{cases}.$$

Consistency and Rate of Convergence

Corollary

- *Let X be a measurable space. Suppose k is measurable and* $\sup_{x\in M} k(x,x) \leq C < \infty$. *Then, for $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, we have*

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(m^{-1/2} + n^{-1/2}).$$

Examples:

- Gaussian kernel: $k(x,y) = e^{-\sigma\|x-y\|_2^2}$, $\sigma > 0$, $x, y \in \mathbb{R}^d$
- Laplacian kernel: $k(x,y) = e^{-\sigma\|x-y\|_1}$, $\sigma > 0$, $x, y \in \mathbb{R}^d$
- Inverse multi-quadratic kernel: $k(x,y) = (c^2 + \|x-y\|_2^2)^{-t}$, $c > 0$, $t > d/2$, $x, y \in \mathbb{R}^d$.

Estimation of Total Variation Distance

Total variation distance is both a ϕ-divergence and integral probability metric given by

$$TV(\mathbb{P}, \mathbb{Q}) = \sup \left\{ \int_X f \, d(\mathbb{P} - \mathbb{Q}) \, : \, \|f\|_\infty \leq 1 \right\}.$$

- *Estimator:* $TV(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^{N} S_i a_i^*$ where $\{a_i^*\}_{i=1}^{N}$ solve the linear program:

  $$\max \left\{ \sum_{i=1}^{N} S_i a_i \, : \, -1 \leq a_i \leq 1, \, \forall \, i \right\}.$$

  Easy to see that $a_i^* = \text{sign}(S_i)$ and therefore $TV(\mathbb{P}_m, \mathbb{Q}_n) = 2$ for any m, n. *Not consistent.*

- Can be estimated consistently using *kernel density estimators.*

Lower Bounds on Total Variation Distance

- $W(\mathbb{P}, \mathbb{Q}) = \sup\{\int_X f \, d(\mathbb{P} - \mathbb{Q}) \, : \, \|f\|_L \leq 1\}$
- $\beta(\mathbb{P}, \mathbb{Q}) = \sup\{\int_X f \, d(\mathbb{P} - \mathbb{Q}) \, : \, \|f\|_L + \|f\|_\infty \leq 1\}$
- $\gamma_k(\mathbb{P}, \mathbb{Q}) = \sup\{\int_X f \, d(\mathbb{P} - \mathbb{Q}) \, : \, \|f\|_{\mathcal{H}} \leq 1\}$

Theorem

(i) For all $\mathbb{P} \neq \mathbb{Q}$, we have

$$TV(\mathbb{P}, \mathbb{Q}) \geq \frac{W(\mathbb{P}, \mathbb{Q})\beta(\mathbb{P}, \mathbb{Q})}{W(\mathbb{P}, \mathbb{Q}) - \beta(\mathbb{P}, \mathbb{Q})}.$$

(ii) Suppose $C := \sup_{x \in X} k(x, x) < \infty$. Then

$$TV(\mathbb{P}, \mathbb{Q}) \geq \frac{\gamma_k(\mathbb{P}, \mathbb{Q})}{\sqrt{C}}.$$

- *Lower bounds on Kullback-Leibler divergence through Pinsker's inequality.*

Summary

- Integral probability metrics vs. ϕ-divergences.
- Estimation of integral probability metrics from finite samples: *easily computable* compared to ϕ-divergences.
- *Fast rates of convergence* compared to ϕ-divergences.
- *Open question:* Minimax rates for estimating integral probability metrics.

Thank You

References

- Müller, A. (1997).
  Integral probability metrics and their generating classes of functions.
  Advances in Applied Probability, 29:429–443.
- Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2008).
  Estimating divergence functionals and the likelihood ratio by convex risk minimization.
  Technical Report 764, Department of Statistics, University of California, Berkeley.
- von Luxburg, U. and Bousquet, O. (2004).
  Distance-based classification with Lipschitz functions.
  Journal for Machine Learning Research, 5:669–695.
- Wang, Q., Kulkarni, S. R., and Verdú, S. (2005).
  Divergence estimation of continuous distributions based on data-dependent partitions.
  IEEE Trans. Information Theory, 51(9):3064–3074.