Minimax Estimation of Kernel Mean Embeddings

Ilya Tolstikhin[1], Bharath Sriperumbudur[2], Krikamol Muandet[1,3]

[1]Department of Empirical Inference, MPI for Intelligent Systems, [2]Department of Statistics, Pennsylvania State University, [3]Department of Mathematics, Mahidol University

Notations and preliminaries

The Fourier transform of a finite Borel measure μ on $\mathbb{R}^d$ is defined by

$$\mathcal{F}[\mu](y) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle y,x\rangle} d\mu(x), \quad y \in \mathbb{R}^d.$$

Theorem 1 (Bochner's theorem)**.** *A continuous function $\psi\colon \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite non-negative Borel measure Λ_ψ on $\mathbb{R}^d$, i.e.,*

$$\psi(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle x,w\rangle} d\Lambda_\psi(w), \quad x \in \mathbb{R}^d.$$

Positive-definite kernels and RKHS

A continuous reproducing kernel $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is translation-invariant if $k(x,y) = \psi(x-y)$ for a symmetric p.d. ψ and all $x, y \in \mathbb{R}^d$. It is radial on every $\mathbb{R}^d$ if and only if there is a finite non-negative Borel measure ν on $[0,\infty)$ such that for all $x, y \in \mathbb{R}^d$

$$k(x,y) = \int_0^\infty e^{-t\|x-y\|_2^2} d\nu(t).$$

Separable reproducing kernel Hilbert space $\mathcal{H}_k$ (RKHS) of real-valued functions on $\mathbb{R}^d$.

For any Borel probability measure P on $\mathcal{X}$ with $\int_{\mathcal{X}} \sqrt{k(x,x)} dP(x)$ its kernel mean is defined by

$$\mu_P := \int k(\cdot, x) dP(x) \in \mathcal{H}_k.$$

A kernel k is characteristic if the kernel mean embedding $\mu_k\colon P \to \mu_P$ (KME) is injective. If k is a bounded continuous translation-invariant p.d. kernel on $\mathbb{R}^d$ then it is characteristic if and only if $\operatorname{supp}(\Lambda_\psi) = \mathbb{R}^d$.

Following kernels are radial and characteristic on $\mathbb{R}^d$:

1. *Gaussian*: $k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\eta^2}\right)$, $\eta > 0$;
2. *Mixture of Gaussians*: $k(x,y) = \sum_{i=1}^M \beta_i \exp\left(-\frac{\|x-y\|_2^2}{2\eta_i^2}\right)$, where $M \geq 2$;
3. *Inverse Multiquadratics*: $k(x,y) = (c^2 + \|x-y\|_2^2)^{-\gamma}$, $c, \gamma > 0$;
4. *Matérn*: $k(x,y) = \frac{c^{2\tau-d}}{\Gamma(\tau-d/2)2^{\tau-1-d/2}} \left(\frac{\|x-y\|_2}{c}\right)^{\tau-d/2} K_{d/2-\tau}(c\|x-y\|_2)$, $\tau > d/2$, $c > 0$, where K_α is the *modified Bessel function of the third kind* of order α.

For further details we refer to Wendland (2005).

Minimax probability in the nonparametric estimation

$\mathcal{P} = \{P_\theta\colon \Theta\}$ is a family of distributions indexed by a set of functions Θ;

d is a distance on Θ.

$S_n = (X_1, \ldots, X_n)$ is an i.i.d. sample distributed according to P_θ for some $\theta \in \Theta$.

$\hat{\theta}_n = \hat{\theta}_n(S_n)$ is an estimator of θ.

A minimax risk associated with the set $\mathcal{P}$ and the distance d:

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}_{S_n\sim P_\theta}\left[d^2(\hat{\theta}_n, \theta)\right]$$

A minimax probability associated with the set $\mathcal{P}$ and the distance d:

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{P}_{S_n\sim P_\theta}\left\{d(\hat{\theta}_n, \theta) \geq \epsilon\right\}.$$

Markov's inequality implies

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}_{S_n\sim P_\theta}\left[d^2(\hat{\theta}_n, \theta)\right] \geq \epsilon^2 \cdot \inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{P}_{S_n\sim P_\theta}\left\{d(\hat{\theta}_n, \theta) \geq \epsilon\right\}.$$

For further details we refer to Tsybakov (2008).

Upper bounds in the RKHS norm

Theorem 2. *Let $(X_i)_{i=1}^n$ be sampled i.i.d. from P defined on a separable topological space $\mathcal{X}$. Suppose $r\colon \mathcal{X} \to H$ is continuous and $\sup_{x\in\mathcal{X}} \|r(x)\|_H^2 \leq C_k < \infty$, where H is a separable Hilbert space of real-valued functions. Then with probability larger than $1-\delta$*

$$\left\| \int r(x) dP(x) - \frac{1}{n}\sum_{i=1}^n r(X_i) \right\|_H \leq \sqrt{\frac{C_k}{n}} + \sqrt{\frac{2C_k \log(1/\delta)}{n}}.$$

By choosing $H = \mathcal{H}_k$ and $r(x) = k(x,\cdot)$ Theorem 2 proves the convergence of $\mu_k(P_n)$ to $\mu_k(P)$ at the rate of $O(n^{-1/2})$ in the RKHS norm if $\sup_{x\in\mathcal{X}} k(x,x) \leq C_k < \infty$. This result improves the constants of Smola et al. (2007, Theorem 2), Gretton et al. (2012), and Lopez-Paz et al. (2015).

Constants: $C_k = 1$ for Matérn, Gaussian, and mixture of Gaussian kernels, $C_k = c^{-2\gamma}$ for inverse multiquadratics kernels.

Minimax lower bounds in the RKHS norm

A set of all Borel discrete measures

Theorem 3 (Translation invariant kernels)**.** *Let $\mathcal{P}$ be the set of all Borel discrete probability measures on $\mathbb{R}^d$. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Assume there exists $z_0 \in \mathbb{R}^d$ and $\beta > 0$, such that $\psi(0) - \psi(z_0) \geq \beta$. Then the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} P^n\left\{\|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{6}\sqrt{\frac{2\beta}{n}}\right\} \geq \frac{1}{4}.$$

Corollary 4 (Radial kernels)**.** *Let $\mathcal{P}$ be the set of all Borel discrete probability measures on $\mathbb{R}^d$ and k be radial on $\mathbb{R}^d$ with $\operatorname{supp}(\nu) \neq \{0\}$. Assume there exist $0 < t_1 < \infty$ and $\alpha > 0$ satisfying $\nu([t_1,\infty)) \geq \alpha$. Then the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} P^n\left\{\|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{6}\sqrt{\frac{\alpha}{n}}\right\} \geq \frac{1}{4}.$$

Constants: Corollary 4 holds with $\alpha = 1$ for Gaussian and mixture of Gaussian kernels, $\alpha = \frac{c^{-2\gamma}}{2}$ for inverse multiquadrics and $\alpha = \frac{1}{2}$ for Matérn kernels.

A set of measures with smooth densities

Theorem 5 (Translation invariant kernels)**.** *Let $\mathcal{P}$ be the set of distributions over $\mathbb{R}^d$ whose densities are continuously infinitely differentiable. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Define $c_\psi := c_{\psi,1}$ and $\epsilon_\psi := \epsilon_{\psi,1}$ where $c_{\psi,1}$ and $\epsilon_{\psi,1}$ are positive constants that satisfy (1) in Proposition 9. Then for any $n \geq \frac{1}{\epsilon_\psi}$, the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} P^n\left\{\|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{2}\sqrt{\frac{c_\psi}{2n}}\right\} \geq \frac{1}{4}.$$

Theorem 6 (Radial kernels)**.** *Let $\mathcal{P}$ be the set of distributions over $\mathbb{R}^d$ whose densities are continuously infinitely differentiable and k be radial on $\mathbb{R}^d$ with $\operatorname{supp}(\nu) \neq \{0\}$. Assume that there exist $0 < t_0 \leq t_1 < \infty$ and $0 < \beta < \infty$ such that $\nu([t_0,t_1]) \geq \beta$. Then the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} P^n\left\{\|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{50}\sqrt{\frac{1}{n}\cdot\frac{\beta t_0}{t_1 e}\left(1 - \frac{2}{2+d}\right)}\right\} \geq \frac{1}{5}.$$

Constants: Theorem 6 holds with $\frac{\beta t_0}{t_1} = 1$ for Gaussian kernels, $\frac{\beta t_0}{t_1} = \frac{\eta_M^2}{\eta_1^2}$ for mixture of Gaussian kernels, and

$$\frac{\beta t_0}{t_1} = \begin{cases} \frac{c^{-2\gamma}}{2\Gamma(\gamma)}\left(\frac{\gamma}{2e}\right)^\gamma, & \text{for } \gamma \geq 1; \\ \frac{c^{-2\gamma}}{4\Gamma(\gamma)}\left(\frac{\gamma}{e}\right)^\gamma, & \text{for } \gamma \in (0,1) \end{cases} \qquad \frac{\beta t_0}{t_1} = \begin{cases} \frac{1}{2\Gamma(\tau-\frac{d}{2})}\left(\frac{2\tau-d}{4e}\right)^{\tau-\frac{d}{2}}, & \text{for } \tau - \frac{d}{2} \geq 1; \\ \frac{1}{4\Gamma(\tau-\frac{d}{2})}\left(\frac{2\tau-d}{2e}\right)^{\tau-\frac{d}{2}}, & \text{for } \tau - \frac{d}{2} \in (0,1) \end{cases}$$

for inverse multiquadratics and Matérn kernels respectively.

Proof techniques

Le Cam's method

Theorem 7 (Lower bound based on two hypotheses)**.** *Assume Θ contains θ_0 and θ_1 such that $d(\theta_0,\theta_1) \geq 2s$ and $\mathrm{KL}(P_{\theta_0}^n \| P_{\theta_1}^n) \leq \alpha$ for some $s > 0$ and $0 < \alpha < \infty$. Then*

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} P_\theta^n\left\{d(\hat{\theta}_n, \theta) \geq s\right\} \geq \max\left(\frac{1}{4}e^{-\alpha}, \frac{1-\sqrt{\alpha/2}}{2}\right).$$

Theorem 8 (Lower bound based on many hypotheses)**.** *Assume $M \geq 2$ and suppose that there exist $\theta_0, \ldots, \theta_M \in \Theta$ such that (i) $d(\theta_i,\theta_j) \geq 2s > 0$, $\forall\, 0 \leq i < j \leq M$; (ii) P_{θ_j} is absolutely continuous w.r.t. P_{θ_0} for all $j = 1, \ldots, M$, and $\frac{1}{M}\sum_{i=1}^M \mathrm{KL}(P_{\theta_j}^n \| P_{\theta_0}^n) \leq \alpha \log M$ with $0 < \alpha < 1/8$. Then*

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} P_\theta^n\left\{d(\hat{\theta}_n, \theta) \geq s\right\} \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right) > 0.$$

In all our lower bounds we apply Le Cam's method with

$$\Theta := \mu_k(\mathcal{P}), \quad d(\theta_1,\theta_2) := \|\theta_1 - \theta_2\|_{\mathcal{H}_k}.$$

Proof of Theorem 3

Apply Le Cam's method (Theorem 7) with $P_0 = \frac{1}{2}(\delta_x + \delta_v)$ and $P_1 = \left(\frac{1}{2} + \frac{1}{3\sqrt{n}}\right)\delta_x + \left(\frac{1}{2} - \frac{1}{3\sqrt{n}}\right)\delta_v$. One can show that $\|\mu_k(P_0) - \mu_k(P_1)\|_{\mathcal{H}_k}^2 = \frac{2}{9n}\big(\psi(0) - \psi(x-v)\big)$.

Proof of Theorem 5

Apply Le Cam's method (Theorem 7) and the following result:

Proposition 9. *Let $\mu_0, \mu_1 \in \mathbb{R}^d$, and $\sigma^2 > 0$. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Let Λ_ψ be a finite non-negative Borel measure corresponding to ψ from Theorem 1. Then there exist constants $\epsilon_{\psi,\sigma^2}, c_{\psi,\sigma^2} > 0$ depending only on ψ and σ^2, such that the following condition holds for any $a \in \mathbb{R}^d$ with $\|a\|_2^2 \leq \epsilon_{\psi,\sigma^2}$:*

$$c_{\psi,\sigma^2} \leq \min_{z\in\mathbb{R}^d\setminus\{0\}} \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \langle e_z, w\rangle^2 \cos\left(\langle a, w\rangle\right) d\Lambda_\psi(w) < \infty, \tag{1}$$

where $e_z := z/\|z\|_2$. Moreover, if Condition (1) is satisfied for each vector a in some symmetric set $A \subseteq \mathbb{R}^d$ (in particular, for a ball $\|a\|_2 \leq R$ of some radius $R > 0$), then the following relation between the RKHS and Euclidean norms holds for all vectors $\mu_0, \mu_1 \in \mathbb{R}^d$, satisfying $\mu_0 - \mu_1 \in A$:

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k} \geq \sqrt{\frac{c_{\psi,\sigma^2}}{2}}\|\mu_0 - \mu_1\|_2,$$

where θ_0 and θ_1 are KMEs of Gaussian measures $G(\mu_0, \sigma^2 I_{d\times d})$ and $G(\mu_1, \sigma^2 I_{d\times d})$ respectively.

Proof of Theorem 6

Apply Le Cam's method (Theorem 8) with $\log M \approx d$. Parameters θ_i, $i = 0, \ldots, M$, correspond to KMEs of Gaussian distributions $G(\mu_i, \sigma^2 I_{d\times d})$ with $\sigma^2 \approx d^{-1}$ and $\|\mu_i - \mu_j\| \approx n^{-1/2}$.

References

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31. Springer-Verlag, 2007.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, NY, 2008.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.