Mixture Density Estimation Via Hilbert Space Embedding of Measures

Bharath K. Sriperumbudur

Gatsby Unit, University College London

Presenter: Hirakendu Das, UC San Diego

Density Estimation

- Problem: Given $\{X_1, \ldots, X_n\}$ drawn i.i.d. from an unknown probability measure with density f, estimate f.
- Approaches: Parametric estimation using maximum likelihood

$$f_{\theta^*}, \text{ where } \theta^* = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f_\theta(X_i)$$

- Maximum likelihood is not applicable to non-parametric estimation.
- Method of Sieves [Grenander, 1981]
  - Perform maximum likelihood on a restricted class
  - Slowly increase the size of the class with increase in n.
- Examples: Histogram estimators, penalized estimators, etc.

Mixture Sieves

Setup: $(\mathcal{X}, \mathscr{A})$ is a measurable space, μ is a σ-finite measure on $\mathscr{A}$.

- Base class, $\mathcal{C} := \left\{x \mapsto \phi_\theta(x) \,:\, \theta \in \Theta \subset \mathbb{R}^d\right\}$
  - Example: Gaussian family parametrized by mean and variance.
- Convex hull of $\mathcal{C}$: $\mathcal{G} = \left\{g(x) = \int_\Theta \phi_\theta(x)\, d\mathbb{P}(\theta),\, \mathbb{P} \in M_+^1(\Theta)\right\}$.

Suppose $f \in \mathcal{G}$. The maximum likelihood estimator is given as

$$\arg\max_{f \in \mathcal{G}} \prod_{i=1}^n f(X_i) = \arg\max_{\mathbb{P} \in M_+^1(\Theta)} \prod_{i=1}^n \int_\Theta \phi_\theta(X_i)\, d\mathbb{P}(\theta)$$

Mixture Sieves

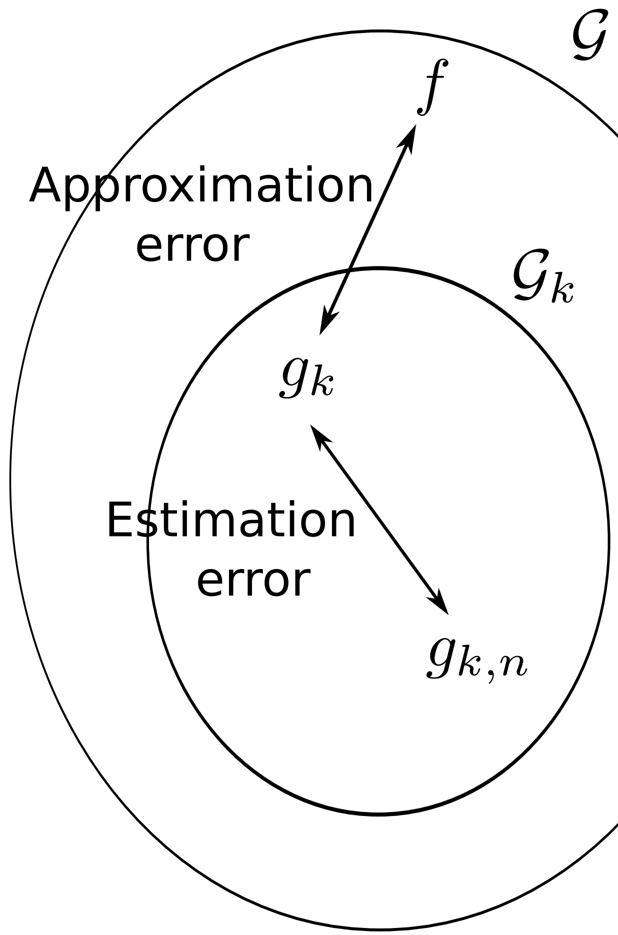
- k-term approximation to $\mathcal{G}$:
  $\mathcal{G}_k = \left\{ g_k(x) = \sum_{j=1}^k \lambda_j \phi_{\theta_j}(x) \,:\, \sum_{j=1}^k \lambda_j = 1,\ \lambda_j \geq 0,\ \forall j \right\}$

$$
\begin{aligned}
g_{k,n} &= \arg\max_{g \in \mathcal{G}_k} \prod_{i=1}^n g(X_i) = \arg \max_{\theta \in \Theta, \|\lambda\|_1 = 1, \lambda \succeq 0} \sum_{i=1}^n \log \left(\sum_{j=1}^k \lambda_j \phi_{\theta_j}(X_i) \right) \\
&= \arg\min_{g \in \mathcal{G}_k} \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i)}{g(X_i)} \\
&= \arg\min_{g \in \mathcal{G}_k} D((X_i)_{i=1}^n \| g)
\end{aligned}
$$

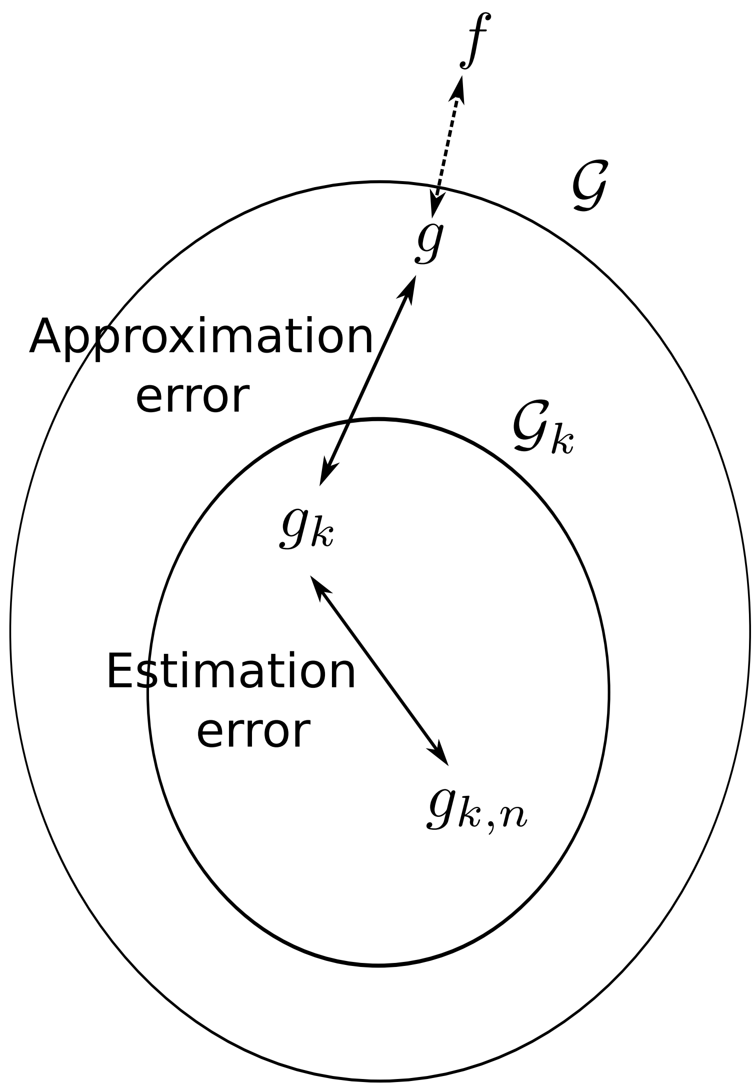
where

$$
D(f\|g) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \, d\mu(x)
$$

Approximation and Estimation Errors



Approximation and Estimation Errors



Approximation error

Theorem ([Li and Barron, 1999])

For any f, there exists $g_k \in \mathcal{G}_k$ such that

$$D(f\|g_k) \leq inf_{g\in\mathcal{G}} D(f\|g) + \frac{4(a + \log(3\sqrt{e}))c_{f,\mathbb{P}}}{k},$$

where

$$a = \sup_{\theta_1,\theta_2,x} \log \frac{\phi_{\theta_1}(x)}{\phi_{\theta_2}(x)}$$

and

$$c_{f,\mathbb{P}} = \int \frac{\int \phi_\theta^2(x)\, d\mathbb{P}(\theta)}{(\int \phi_\theta(x)\, d\mathbb{P}(\theta))^2}\, d\mu$$

In fact such a g_k can be obtained iteratively as

$$D(f\|g_k) \leq \min_{\lambda,\theta} D(f\|(1-\lambda)g_{k-1} + \lambda\phi_\theta).$$

Greedy Estimation

Choose $g_{k,n} \in \mathcal{G}_k$ such that

$$\sum_{i=1}^{n} \log g_{k,n}(X_i) \geq \max_{\lambda,\theta} \sum_{i=1}^{n} \log\left((1-\lambda)g_{k-1,n}(X_i) + \lambda\phi_\theta(x)\right)$$

Clearly the maximum likelihood estimator satisfies the above inequality, i.e., choose $g_{k,n} = \arg\max_{g_k \in \mathcal{G}_k} \sum_{i=1}^{n} \log g_k(X_i)$.

Error Bound

Theorem ([Li and Barron, 1999])

Suppose Θ is a d-dimensional cube with side length A and that

$$\sup_{x \in \mathcal{X}} |\log \phi_\theta(x) - \log \phi_{\theta'}(x)| \leq B \|\theta - \theta'\|_1$$

for any $\theta, \theta' \in \Theta$. Let $g_{k,n}$ satisfy the inequality in red. Then

$$\mathbb{E}\left[D(f \| g_{k,n})\right] \leq inf_{g \in \mathcal{G}} D(f \| g) + \frac{c_1}{k} + \frac{c_2 k \log(n c_3)}{n},$$

where c_1, c_2 and c_3 are constants (dependent on A, B and d) independent of k and n.

Optimal rate: $O_f\left(\sqrt{\frac{\log n}{n}}\right)$ with $k \sim \sqrt{\frac{n}{\log n}}$

Improved Error Bound

[Rakhlin et al., 2005] showed that for any $g_k \in \mathcal{G}_k$ and any f,

$$D(f||g_{k,n}) - D(f||g_k) \leq \frac{c_1}{k} + \frac{c_2}{\sqrt{n}} + c_3 \int_0^b \sqrt{\frac{\log \mathcal{N}(\mathcal{C}, \epsilon, d_n)}{n}}\, d\epsilon,$$

where $0 < a \leq \phi_\theta(x) \leq b < \infty$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$.

- $d_n^2(\phi_1, \phi_2) := \frac{1}{n}\sum_{j=1}^n (\phi_1(X_j) - \phi_2(X_j))^2$
- $\mathcal{N}(\mathcal{C}, \epsilon, d_n)$ represents the ϵ-covering number of $\mathcal{C}$
- If $\mathcal{C}$ is a VC-class, then the optimal rate is $O_f\left(\frac{1}{\sqrt{n}}\right)$ with $k \sim \sqrt{n}$.

Issues: Boundedness of f and ϕ_θ; finite entropy integral of $\mathcal{C}$.

Outline

$$\gamma_K(\mathbb{P},\mathbb{Q}) = \left\| \int_{\mathcal{X}} K(\cdot,x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot,x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}},$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space and K is a reproducing kernel.

- Interpretation
- M-estimator
- Rates of convergence

Reproducing Kernel Hilbert Space

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, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, K(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- 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 $\mathcal{X} \times \mathcal{X}$, 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.

Interpretation [Sriperumbudur et al., 2010]

$$\mathbb{P} \mapsto \int_{\mathcal{X}} K(\cdot, x)\, d\mathbb{P}(x) := \Phi(\mathbb{P}) \in \mathcal{H}$$

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|_{\mathcal{H}}$$

- Suppose $K(x, y) = e^{-i\langle x,y\rangle_2}$, $x, y \in \mathbb{R}^d$. Then $\gamma_K(\mathbb{P}, \mathbb{Q})$ is the L_2 distance between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$.
- If $K(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, then $\gamma_K(\mathbb{P}, \mathbb{Q})$ is the weighted L_2 distance (weighted by the Fourier transform of ψ) between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$.
- Φ is a generalization of the characteristic function of $\mathbb{P}$.

Choice of k

- Not all K are interesting: $K(x,y) = \langle x, y\rangle_2,\ \ x, y \in \mathbb{R}^d$.

$$\gamma_K(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_2 \,.$$

  Therefore, $\gamma_K(\mathbb{P},\mathbb{Q}) = 0 \nRightarrow \mathbb{P} = \mathbb{Q}$.

- Interesting kernels: Let $K(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$. If the support of the Fourier transform of ψ is $\mathbb{R}^d$, then

$$\gamma_K(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

- Examples: $e^{-\sigma\|x-y\|_2^2}$, $e^{-\sigma\|x-y\|_1}$, $\sigma > 0$, etc.

M-Estimator

$$\gamma_K(f,g) := \left\| \int_{\mathcal{X}} K(\cdot,x)(f(x)-g(x))\,d\mu(x) \right\|_{\mathcal{H}},$$

$$\gamma_K(S,g) := \left\| \frac{1}{n}\sum_{i=1}^{n} K(\cdot,X_i) - \int K(\cdot,x)g(x)\,d\mu(x) \right\|_{\mathcal{H}},$$

and

$$g_{\text{emp}} := \arg\min_{g\in\mathcal{G}_k} \gamma_K(S,g),$$

where $S := \{X_1,\ldots,X_n\}$. g_{emp} is called an M-estimator.

Main Result

Theorem

Let $C := \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}$*, where* K *is a continuous kernel on a separable topological space,* $\mathcal{X}$*. Then with probability at least* $1-\delta$ *over the choice of samples* $\{X_j\}_{j=1}^n$ *drawn i.i.d. from* f*, the following hold:*

$$\gamma_K(f, g_{emp}) - \inf_{g \in \mathcal{G}} \gamma_K(f,g) \leq \frac{4C}{\sqrt{n}} + \sqrt{\frac{8C^2}{n} \log \frac{2}{\delta}} + \frac{2C}{\sqrt{k}}.$$

In addition,

$$-\frac{2C}{\sqrt{n}} - \sqrt{\frac{2C^2}{n} \log \frac{1}{\delta}} \leq \gamma_K(S, g_{emp}) - \inf_{g \in \mathcal{G}} \gamma_K(f,g) \leq \frac{2C}{\sqrt{n}} + \sqrt{\frac{2C^2}{n} \log \frac{1}{\delta}} + \frac{2C}{\sqrt{k}}.$$

Remarks

- No assumptions on f, ϕ_θ and $\mathcal{C}$
- Approximation error: $O\left(\frac{1}{\sqrt{k}}\right)$
- Estimation error: $O_f\left(\frac{1}{\sqrt{n}}\right)$
- Optimal rate: $O_f\left(\frac{1}{\sqrt{n}}\right)$ with $k \sim n$

$$\gamma_K(\mathbb{P}, \mathbb{Q}) \leq C\sqrt{2\,D(\mathbb{P}\|\mathbb{Q})}$$

- Fast rates

Proof Idea

Let us fix an $\varepsilon > 0$ and a function $g_\varepsilon \in \mathcal{G}$ such that

$$\gamma_K(f, g_\varepsilon) \leq \inf_{g \in \mathcal{G}} \gamma_K(f, g) + \varepsilon.$$

$$\begin{aligned}
\gamma_K(f, g_{\mathrm{emp}}) - \inf_{g \in \mathcal{G}} \gamma_K(f, g) \quad &= \quad \overbrace{\gamma_K(f, g_{\mathrm{emp}}) - \gamma_K(S, g_{\mathrm{emp}})}^{A_1} \\
&\quad + \overbrace{\gamma_K(S, g_{\mathrm{emp}}) - \gamma_K(S, \tilde{g}_k)}^{A_2} \\
&\quad + \overbrace{\gamma_K(S, \tilde{g}_k) - \gamma_K(f, \tilde{g}_k)}^{A_3} \\
&\quad + \gamma_K(f, \tilde{g}_k) - \inf_{g \in \mathcal{G}} \gamma_K(f, g) \\
&\leq \quad A_1 + A_2 + A_3 \\
&\quad + \overbrace{\gamma_K(f, \tilde{g}_k) - \gamma_K(f, g_\varepsilon)}^{A_4} + \varepsilon.
\end{aligned}$$

Proof Idea

$$A_1 \leq \gamma_K(S, f), \quad A_2 \leq 0, \quad A_3 \leq \gamma_K(S, f), \quad A_4 \leq \gamma_K(\tilde{g}_k, g_\varepsilon)$$

Using concentration in Hilbert spaces (e.g., Hoeffding's inequality), it can be shown that

$$\gamma_K(\tilde{g}_k, g_\varepsilon) \leq \frac{2C}{\sqrt{k}}$$

and with probability at least $1 - \frac{\delta}{2}$ over the choice of $\{X_j\}_{j=1}^n$,

$$\gamma_K(S, f) \leq \frac{2C}{\sqrt{n}} + \sqrt{\frac{2C^2}{n} \log \frac{2}{\delta}},$$

Letting $\varepsilon \to 0$ yields the result.

Summary

- Mixture sieve density estimation via RKHS embedding of measures
- No assumptions of f and $\mathcal{C}$
- Fast error rates
- Disadvantage: Weaker distance than the KL divergence (convergence in γ_K does not imply the convergence in KL)

Thank You

References

- Grenander, U. (1981).
  Abstract Inference.
  Wiley, New York.
- Li, J. and Barron, A. (1999).
  Mixture density estimation.
  In Solla, S. A., Leon, T. K., and Muller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 279–285, San Mateo, CA. Morgan Kaufmann Publishers.
- Rakhlin, A., Panchenko, D., and Mukherjee, S. (2005).
  Risk bounds for mixture density estimation.
  ESAIM: Probability and Statistics, 9:220–229.
- Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010).
  Hilbert space embeddings and metrics on probability measures.
  Journal of Machine Learning Research, 11:1517–1561.