Exponential Concentration of a Density Functional Estimator

Shashank Singh and Barnabás Póczos

Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA

Carnegie Mellon University

Introduction

- Many important statistical quantites can be written as
$$F(p_1, \dots, p_k) = \int_{\mathcal{X}_1 \times \cdots \times \mathcal{X}_k} f(p_1(x_1), \dots, p_k(x_k))\, d(x_1, \dots, x_k),$$
where each $p_i : \mathcal{X}_i \subseteq \mathbb{R}^d \to \mathbb{R}^+$ is a probability density, $f : \mathbb{R}^k \to \mathbb{R}$ is smooth.
- Examples of such quantities, which we call *Density Functionals*, are:
  - Shannon/KL, Rényi-α, and Tsallis-α entropies, mutual informations, and divergences
  - f-divergences (e.g., Hellinger, Jensen-Shannon, etc.)
  - L_p norms and distances
  - Conditional versions of the above quantities
- For many of these quantities, few consistent estimators are known, and almost none of these have finite-sample convergence of concentration guarantees.
- We propose and study a nonparametric estimator for such quantities, based on plugging in a boundary-corrected kernel density estimate.
- We prove that, when each $\mathcal{X}_i = [0, 1]^d$ is a unit cube:
  - our estimator is exponentially concentrated about its mean.
  - for densities in a β-Hölder smoothness class with certain boundary conditions, the bias of the estimator decays as $O\left(n^{-\frac{\beta}{\beta+d}}\right)$, where n is the number of samples from each density.

Assumptions

Let $\beta > 0$, and let $\ell := \lfloor \beta \rfloor$ be the greatest integer *strictly* less than β. We make the following four assumptions on f, the densities $p_1, \dots, p_k$, the kernel K:

- **(f-Smoothness)** f is twice continuously differentiable.
- **(Density Smoothness)** All (mixed) ℓ-order partial derivatives of $p_1, \dots, p_k$ exist and are $(\beta - \ell)$-Hölder Continuous (i.e., there exists $L \in \mathbb{R}$ such that, $\forall x, x + v \in \mathcal{X}$, $|\vec{i}| = \ell$, each
$$|D^{\vec{i}} p_i(x + v) - D^{\vec{i}} p_i(x)| \leq L \|v\|_2^{(\beta - \ell)}).$$
- **(Density Boundaries)** All derivatives of $p_1, \dots, p_k$ of order up to ℓ vanish at the boundary
$$\partial \mathcal{X} = \{x \in \mathcal{X} : x_i \in \{0, 1\} \text{ for some } i \in [d]\}$$
(i.e., $\sup_{1 \leq |\vec{i}| \leq \ell} |D^{\vec{i}}(x)| \to 0$ as $\operatorname{dist}(x, \partial \mathcal{X}) \to 0$).
- **(Kernel)** The kernel $K : \mathbb{R} \to \mathbb{R}$ has support in $[-1, 1]$,
$$\int_{-1}^{1} K(u)\, du = 1 \quad \text{and} \quad \int_{-1}^{1} u^j K(u)\, du = 0, \quad \forall j \in \{1, \dots, \ell\}.$$

Mirrored Kernel Density Estimator

Given a bandwidth h, our density functional estimate is computed in 3 steps:

1. Augment data from p_i with reflections over each subset of edges of $\mathcal{X}_i$.
2. Compute kernel density estimates $\hat{p}_1, \dots, \hat{p}_k$ from the augmented data, using a product kernel and bandwidth h.
3. Estimate $F(p_1, \dots, p_k)$ by the plug-in estimator $F(\hat{p}_1, \dots, \hat{p}_k)$.

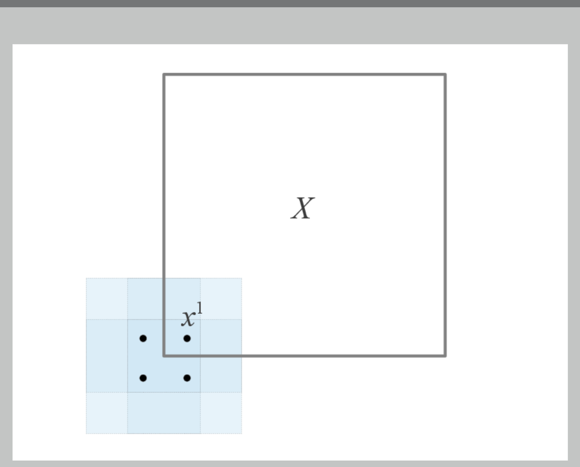

Figure : Four kernels centered on a single data point and its three reflected copies, in the case $d = 2$.

Results: Exponential Concentration Bound

- We show that, $\forall \varepsilon > 0$,
$$\mathbb{P}\left(|F(\hat{p}_1, \dots, \hat{p}_k) - \mathbb{E}F(\hat{p}_1, \dots, \hat{p}_k)| > \varepsilon\right) \leq 2 \exp\left(-\frac{\varepsilon^2 n}{C_V^2}\right),$$
where $C_V = 2C_f \|K\|_1^d$ is constant in n and h.
- Main tool in proof is McDiarmid's Inequality, by which it suffices to bound the change in the estimate when resampling a single data point by C_V/n.
- This is achieved by combining the smoothness of f with the observation that the integral of the mirrored kernel density estimate changes by at most $\frac{2}{n}\|K\|_1^d$.

Results: Convergence Rate

- We show there exists $C_B \in \mathbb{R}$ (constant in n and h) such that
$$|\mathbb{E}F(\hat{p}_1, \dots, \hat{p}_k) - F(p_1, \dots, p_k)| \leq C_B\left(h^\beta + \frac{1}{nh^d}\right).$$
- Previous work [2] bounded the integral of each mirrored kernel density estimator's pointwise squared bias:
$$\int_{\mathcal{X}_i} (\mathbb{E}\hat{p}_i(x) - p(x))^2\, dx \leq C_b h^{2\beta} \qquad (1)$$
- To derive our convergence rate, we make a second-order Taylor estimate of f and then use Hölder's Inequality to reduce the resulting terms to (1) and the integrated mean squared error of a standard kernel density estimator.

Condition Density Functionals

- It is often useful to condition density functionals on one or more additional variables; e.g., to estimate
$$F(P) = \int_{\mathcal{Z}} P_Z(z) f\left(\int_{\mathcal{X}} g\left(\frac{P_{X,Z}(x, z)}{P_Z(z)}\right) dx\right) dz.$$
- For example, conditional entropy estimation has applications to clustering [3] and conditional mutual information is useful for learning graphical models [1].
- As long as the density of the conditioned variable (e.g., P_Z) has a positive lower bound, our results extend to the mirrored kernel density plug-in estimator for such *conditional density functionals*.

Discussion

- The exponential concentration bound gives a bound on the variance of the estimator:
$$\mathbb{V}[F(\hat{p}_1, \dots, \hat{p}_k)] \leq C_V^2 n^{-1}.$$
- This does not depend on h, so pick h to minimize the bias bound.
  - Asymptotically, the optimal h is $h \asymp n^{-\frac{1}{\beta+d}}$, so bias bound is $O\left(n^{-\frac{\beta}{\beta+d}}\right)$.
- Hence MSE is $O(n^{-\frac{2\beta}{\beta+d}} + n^{-1})$, which is the parametric rate $O(n^{-1})$ if $\beta \geq d$ and $O(n^{-\frac{\beta}{\beta+d}})$ otherwise.
- Kernel assumptions for the bias bound necessitate $\|K\|_1 > 1$ when $\beta \geq 2$ and C_V includes $\|K\|_1^d$, which is exponential in d.
  - Lower bounds in d are unknown; whether dependence is necessarily exponential is an important open problem.
- For divergences and information theoretic density functionals, the integral in the plug-in estimator can be well estimated by a sample mean. In this case, our estimator can be computed in $O(2^d n \log n)$, making it effective for large, low-dimensional datasets.

References

- D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, Cambridge, MA, 2009.
- S. Singh and B. Poczos. Generalized exponential concentration inequality for Rényi divergence estimation. In *International Conference on Machine Learning (ICML)*, 2014.
- G. ver Steeg, A. Galstyan, F. Sha, and S. DeDeo. Demystifying information-theoretic clustering. In *International Conference on Machine Learning (ICML)*, 2014.

LaTeX beamerposter http://www-i6.informatik.rwth-aachen.de/~dreuw/latexbeamerposter.php

http://www.contrib.andrew.cmu.edu/∼sss1

sss1@andrew.cmu.edu