Exponential Concentration Inequality for a Rényi-α Divergence Estimator

Shashank Singh and Barnabás Póczos
Machine Learning Department, Carnegie-Mellon University, Pittsburgh, PA, USA

Carnegie Mellon University

Introduction

- For a fixed $\alpha \in [0,1) \cup (1,\infty)$, we are interested in estimating the Rényi-α divergence
$$D_\alpha(p\|q) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} p^\alpha(x) q^{1-\alpha}(x))\,dx,$$
between two unknown, continuous, nonparametric probability densities p and q over $\mathcal{X} \subseteq \mathbb{R}^d$, using samples from each density.
- Applications of divergence estimation include
  - extending machine learning algorithms designed to operate on finite-dimensional feature vectors to the setting where inputs are sets or distributions.
  - estimating entropy and mutual information.
- Rényi-α Divergence has KL-Divergence as its $\alpha \to 1$ limit case, and is related to Tsallis-α, Jensen-Shannon, and Hellinger divergences.
- Few divergence estimators have known rates, and, to the best of our knowledge, none have known exponential concentration bounds.
- We propose and analyze a plug-in estimator based on kernel density estimation, for densities on the unit cube $\mathcal{X} = [0,1]^d$. We prove
  - the estimator is exponentially concentrated about its mean.
  - for densities in a β-Hölder smoothness class with certain boundary conditions, the bias of the estimator is bounded by $O\left(n^{-\frac{\beta}{\beta+d}}\right)$, 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 the densities p and q, and the kernel K:

- **(Smoothness)** All (mixed) ℓ-order partial derivatives of p and q exist and are $(\beta - \ell)$-Hölder Continuous (i.e., $\exists L \in \mathbb{R}$ such hat, $\forall x, x+v \in \mathcal{X}$, $|\vec{i}| = \ell$,
$$|D^{\vec{i}}p(x+v) - D^{\vec{i}}p(x)|, |D^{\vec{i}}q(x+v) - D^{\vec{i}}q(x)| \leq L\|v\|_2^{\beta-\ell}).$$
- **(Boundedness)** $\exists \kappa_1, \kappa_2 \in \mathbb{R}$ such that, $\forall x \in \mathcal{X}$,
$$0 < \kappa_1 \leq p(x), q(x) \leq \kappa_2 < +\infty.$$
- **(Boundary)** All derivatives of p and q 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 $\mathbf{dist}(x, \partial\mathcal{X}) \to 0$).
- **(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 Rényi-α divergence estimate is computed in 3 steps:

1. Mirror data over subsets of edges of $\mathcal{X}$.
2. Compute clipped kernel density estimates $\hat{p}$ and $\hat{q}$ from the mirrored data, using product kernel K_d and bandwidth h, and clipping the kernel density estimates pointwise below at κ_1 and above at κ_2.
3. Estimate $D_\alpha(p\|q)$ by the plug-in estimator $D_\alpha(\hat{p}\|\hat{q})$.

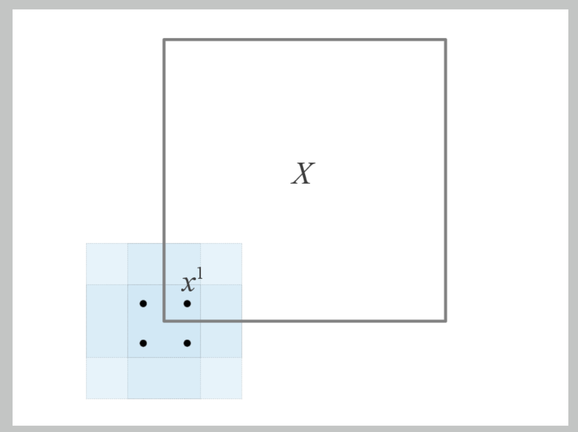

Figure : Four kernels centered on a single data point and its three reflected copies, in the case $d = 2$.

Results: Exponential Concentration Bound

- We show that, $\forall \varepsilon > 0$,
$$\mathbb{P}\left(|D_\alpha(\hat{p}\|\hat{q}) - \mathbb{E}D_\alpha(\hat{p}\|\hat{q})| > \varepsilon\right) \leq 2\exp\left(-C_V^2 \varepsilon^2 n\right),$$
where
$$C_V = \frac{|\alpha - 1|}{2C_f C_L \|K\|_1^d}$$
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 a smoothness property of D_α with the observation that the integral of the mirrored kernel density estimate changes by at most $\frac{2}{n}\int_{[-1,1]^d} |K^d(u)|\,du$.

Bias Lemma

- **Bias Lemma:** Writing the pointwise bias of the clipped and mirrored kernel density as $b_p(x) = \mathbb{E}\hat{p}(x) - p(x)$, we show
$$\int_{\mathcal{X}} b_p^2(x)\,dx \leq C_b h^{2\beta}.$$
  - For somewhat small h and large β (in particular, $h \leq \sqrt{\frac{1}{3d^{1/r}}}$ and $\beta \geq 6d + 2$ suffices), one can show $C_b \leq 3L$.
  - Away from the boundary of $\mathcal{X}$ (i.e., in $[h, 1-h]^d$), there is no boundary bias, and so we simply cite well-known results in kernel density estimation, using the assumed symmetry properties of the kernel.
  - For x near (within h of) the boundary of $\mathcal{X}$, we combine the Smoothness and Boundary Conditions via a Taylor bound to derive a pointwise bound $b_p(x) \leq C_b h$.

Results: Convergence Rate

- We show there exists $C_B \in \mathbb{R}$ (constant in n and h) such that
$$|\mathbb{E}D_\alpha(\hat{p}\|\hat{q}) - D_\alpha(p\|q)| \leq C_B\left(h^\beta + \frac{1}{nh^d}\right).$$
- Proven by making a second-order Taylor estimate and then using Hölder's Inequality to reduce terms to the Bias Lemma and the integrated mean squared error of a standard kernel density estimator.

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 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 $O(n^{-\frac{\beta}{\beta+d}} + n^{-1})$, 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.

Experimental Results on Synthetic Data

- $\vec{\mu}_1 = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \end{bmatrix}, \vec{\mu}_2 = \begin{bmatrix} 0.7 \\ 0.7 \\ 0.7 \end{bmatrix}$
- $\Sigma = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.3 \end{bmatrix}$
- $p_1 = \mathcal{N}(\vec{\mu}_1, \Sigma), p_2 = \mathcal{N}(\vec{\mu}_2, \Sigma)$
- In each trial, n points were drawn from p_1 and p_2 restricted to $[0,1]^3$. $D_\alpha(\hat{p}\|\hat{q})$ was computed from the samples and $D_\alpha(p_1\|p_2)$ was computed directly.

Mean squared error and standard deviation of our estimator were computed from 100 trials and plotted below.

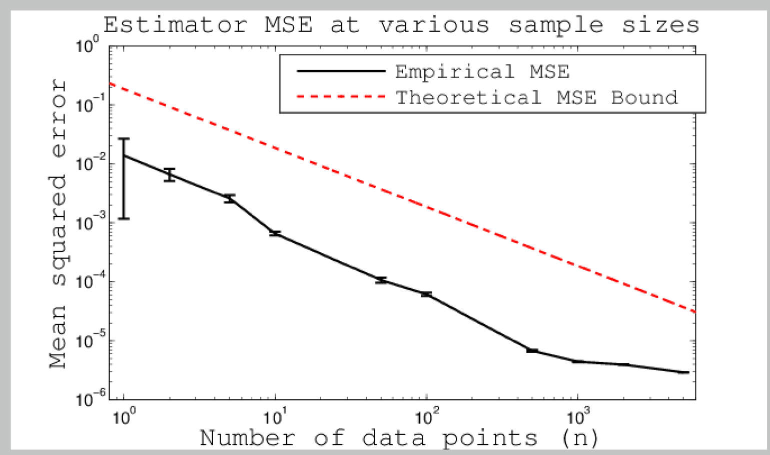

Figure : Log-log plot of mean squared (computed over 100 trials) for various sample sizes n, alongside our theoretical bound. Error bars indicate standard deviation of estimator over 100 trials.

http://www.contrib.andrew.cmu.edu/~sss1
sss1@andrew.cmu.edu