Exponential Concentration Inequality for a Rényi-α Divergence Estimator
Shashank Singh and Barnabás Póczos
Machine Learning Department, Carnegie-Mellon University, Pittsburgh, PA, USA

Introduction

For a fixed $\alpha \in [0, 1) \cup (1, \infty)$, we are interested in estimating the Rényi-α divergence

$$D_\alpha(p||q) = \frac{1}{\alpha - 1} \log \int_X p^\alpha(x)q^{1-\alpha}(x) \, dx,$$

between two unknown, continuous, nonparametric probability densities $p$ and $q$ over $X \subseteq \mathbb{R}^d$ 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 $\alpha \rightarrow 1$ 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, using the assumed symmetry properties of the densities.

$\alpha > 1$: we make the following four assumptions on the densities $p$ and $q$, and the kernel $K$.

- **Smoothness** All (mixed) $\ell$-order partial derivatives of $p$ and $q$ exist and are $(\beta - \ell)$-Hölder Continuous (i.e., $\exists L \in \mathbb{R}$ such that, $\forall x, x+v \in X$, $\|x\|_r = \ell$, $|D^\ell p(x+v) - D^\ell p(x)|, |D^\ell q(x+v) - D^\ell q(x)| \leq L|x-v|_r^{\beta - \ell}$).

- **Boundedness** $\exists \kappa_1, \kappa_2 \in \mathbb{R}$ such that, $\forall x \in X$, $0 < \kappa_2 \leq p(x), q(x) \leq \kappa_2 < +\infty$.

- **Boundary** All derivatives of $p$ and $q$ vanish at the boundary $\partial X = \{x \in X : x_k \in \{0, 1\} \text{ for some } i \in [d]\}$ (i.e., $\sup_{1 \leq i \leq d} |D^\ell (x_j)| \rightarrow 0$ as $dist(x, \partial X) \rightarrow 0$).

- **Kernel** The kernel $K : \mathbb{R} \rightarrow \mathbb{R}$ has support in $[-1, 1]$, $\int_{-1}^1 K(u) \, du = 1$ and $\int_{-1}^1 uK(u) \, du = 0$, $\forall j \in \{1, \ldots, \ell\}$.

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 $X$.
2. Compute clipped kernel density estimates $\tilde{p}$ and $\tilde{q}$ from the mirrored data, using product kernel $K_h$ and bandwidth $h$, and clipping the kernel density estimates pointwise below at $k_1$ and above at $k_2$.
3. Estimate $D_\alpha(p||q)$ by the plug-in estimator $D_\alpha(\tilde{p}||\tilde{q})$.

Results: Exponential Concentration Bound

We show that, $\forall \epsilon > 0$,

$$P\left(|D_\alpha(\tilde{p}||\tilde{q}) - D_\alpha(p||q)| > \epsilon \right) \leq 2 \exp\left(-C_\alpha \epsilon^2 n\right),$$

where

$$C_\alpha = \frac{|\alpha - 1|}{2C_0 |K|^2 h^d}$$

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 $C_\alpha/n$.

This is achieved by combining a smoothness property of $D_\alpha$ with the observation that the integral of the mirrored kernel density estimate changes by at most $\frac{2}{n} \int_{-1}^1 |K^d(u)| \, du$.

Bias Lemma

Writing the pointwise bias of the clipped and mirrored kernel density as $b_\alpha(x) = \mathbb{E}[\tilde{p}(x) - p(x)]$, we show

$$\int b^2_\alpha(x) \, dx \leq C_\alpha h^{2d}.$$

For somewhat small $h$ and large $\beta$ (in particular, $h \leq \sqrt{\frac{\epsilon}{3C_\alpha}}$ and $\beta \geq 6d + 2$ suffices), one can show $C_\alpha \leq 3L$.

Away from the boundary of $X$ (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 $X$, we combine the Smoothness and Boundary Conditions via a Taylor bound to derive a pointwise bound $b_\alpha(x) \leq C_\alpha h$.

Experimental Results on Synthetic Data

$$\mu_1 = \begin{bmatrix} 0.3 \\ 0.4 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.3 \end{bmatrix}$$

$$p_1 = \mathcal{N}(\mu_1, \Sigma), p_2 = \mathcal{N}(\mu_2, \Sigma)$$

In each trial, $n$ points were drawn from $p_1$ and $p_2$ restricted to $[0, 1]^3$. $D_\alpha(\tilde{p}||\tilde{q})$ was computed from the samples and $D_\alpha(p_1||p_2)$ was computed directly.

Results: Convergence Rate

- We show there exists $C_\delta \in \mathbb{R}$ (constant in $n$ and $h$) such that

$$|\mathbb{E}[D_\alpha(\tilde{p}||\tilde{q})] - D_\alpha(p||q)| \leq C_\delta \left(h^2 + \frac{1}{nh}\right).$$

- 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:

$$\mathbb{V}[F(p_1, \ldots, p_n)] \leq C_\epsilon n^{-d}.$$”

- This does not depend on $h$, so pick $h$ to minimize the bias bound.

- Asymptotically optimal $h$ is $h \propto n^{-\frac{d}{2d}}$, so bias bound is $O\left(n^{-\frac{d}{2d}}\right)$.

- Hence MSE is $O(n^{-\frac{d}{2d}} + n^{-1})$, which is the parametric rate $O(n^{-\frac{d}{2d}})$ if $\beta \geq d$ and $O(n^{-\frac{d}{2d}})$ otherwise.

- Kernel assumptions for the bias bound necessitate $|K|_d \geq 1$ when $\beta \geq 2$ and $C_\alpha \propto |K|_d$, which is exponential in $d$.

- Lower bounds in $d$ are unknown; whether dependence is necessarily exponential is an important open problem.