Exponential Concentration Inequality for a Rényi-α Divergence Estimator

Shashank Singh [1]   Barnabás Póczos [1]

June 22, 2014

[1] Carnegie Mellon University, Pittsburgh, PA, USA

Problem

Given $\alpha \in [0,1) \cup (1,\infty)$, estimate the Rényi-$\alpha$ divergence

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

between two unknown, continuous, nonparametric probability densities p and q over $\mathcal{X} = [0,1]^d$, using n samples from each density.

Contribution

- plug-in estimator of Rényi-α divergence based on kernel density estimation
- bound bias of the estimator
- prove a concentration inequality
- simple proof-of-concept experiment

Motivation

- 'distributional' machine learning algorithms
  - finite-dimensional feature vectors $\rightarrow$ distribution features

Motivation

- 'distributional' machine learning algorithms
  - finite-dimensional feature vectors $\rightarrow$ distribution features
- KL-divergence, entropy, and mutual information special cases
  - applications to feature selection, clustering, ICA, etc.

Motivation

- ‘distributional’ machine learning algorithms
  - finite-dimensional feature vectors $\rightarrow$ distribution features
- KL-divergence, entropy, and mutual information special cases
  - applications to feature selection, clustering, ICA, etc.
- with concentration inequality:
  - can simultaneously bound error of multiple estimates (e.g., forest density estimation)
  - can derive hypothesis test for independence

Related Work

- Few known rates
- No estimators have concentration inequalities or other results describing their distribution

Smoothness (Hölder) Condition

Same assumptions on p and q.

Smoothness (Hölder) Condition

Same assumptions on p and q.

β-Hölder condition on p:

- $\beta, L > 0$, $\ell := \lfloor \beta \rfloor$ (so $\beta - 1 \leq \ell < \beta$)

All ℓ-order (mixed) partial derivatives of p and q exist and

$$\sup_{\substack{x \neq y \in \mathcal{X} \\ |\vec{i}| = \ell}} \frac{|D^{\vec{i}} p(x) - D^{\vec{i}} p(y)|}{\|x - y\|_r^{\beta - \ell}} \leq L.$$

Boundedness

There exist known $\kappa_1, \kappa_2 \in \mathbb{R}$ such that, $\forall x \in \mathcal{X}$,

$$0 < \kappa_1 \leq p(x), q(x) \leq \kappa_2 < +\infty.$$

- *Existence* of κ_2 is trivial, but our estimator requires it to be *known* beforehand.
- Assuming κ_1 for q is natural (to ensure $D_\alpha(p\|q) < +\infty$).
- κ_1 for p is technical, and can be weakened/eliminated in certain cases.
- Reason for working on finite measure domain $\mathcal{X} = [0,1]^d$.

Boundary Condition

All derivatives of p vanish at the boundary; i.e.,

$$\sup_{1 \leq |\vec{i}| \leq \ell} |D^{\vec{i}} p(x)| \to 0$$

as

$$\text{dist}(x, \partial\mathcal{X}) \to 0.$$

Boundary Condition

All derivatives of p vanish at the boundary; i.e.,

$$\sup_{1 \leq |\vec{i}| \leq \ell} |D^{\vec{i}} p(x)| \to 0$$

as

$$\text{dist}(x, \partial\mathcal{X}) \to 0.$$

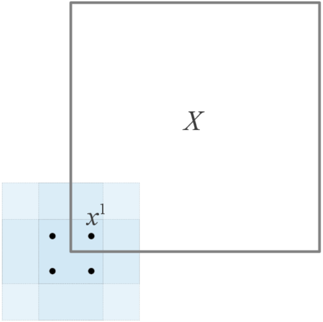
Strong assumption, but needed to eliminate boundary bias.

Kernel Assumptions

$K : \mathbb{R} \to \mathbb{R}$ with support in $[-1, 1]$ and satisfies

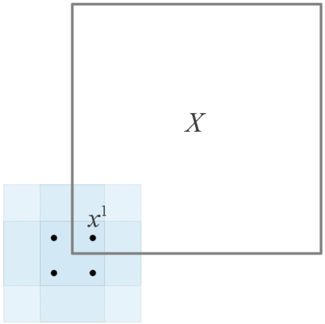
$$\int_{-1}^{1} K(u)\, du = 1 \quad \text{and} \quad \int_{-1}^{1} u^j K(u)\, du = 0, \quad \forall j \in \{1, \cdots, \ell\}.$$

Mirrored Kernel Density Estimate



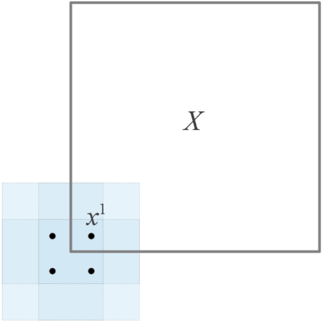
1. Mirror data $x^1, \cdots, x^n$ across all subsets of edges of $\mathcal{X}$

Mirrored Kernel Density Estimate



1. Mirror data $x^1, \cdots, x^n$ across all subsets of edges of $\mathcal{X}$
2. Using a bandwidth h and product kernel K^d, compute kernel density estimate (KDE) $\widetilde{p}$ from resulting $3^d n$ data points

Mirrored Kernel Density Estimate



1. Mirror data $x^1, \cdots, x^n$ across all subsets of edges of $\mathcal{X}$
2. Using a bandwidth h and product kernel K^d, compute kernel density estimate (KDE) $\tilde{p}$ from resulting $3^d n$ data points
   - Removes boundary bias because we assume derivatives of p vanish near $\partial\mathcal{X}$.

Rényi-α Divergence Estimator

1. Clip mirrored KDE below by κ_1 and above by κ_2

$$\text{i.e.,} \quad \hat{p}(x) = \min\{\kappa_2, \max\{\kappa_1, \tilde{p}(x)\}\}.$$

2. Compute $\hat{q}$ by the same process
3. Plug $\hat{p}, \hat{q}$ into D_α:

$$D_\alpha(\hat{p}\|\hat{q}) = \frac{1}{\alpha - 1} \log \int_{\mathcal{X}} \hat{p}^\alpha(x)\hat{q}^{1-\alpha}(x)\, dx.$$

Bounds

- **Bias Bound:** $\exists C_B \in \mathbb{R}$ such that

$$\left|\mathbb{E}D_\alpha(\hat{p}\|\hat{q}) - D_\alpha(p\|q)\right| \leq C_B\left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right).$$

- **Concentration Inequality ('Variance' Bound):** $\exists C_V \in \mathbb{R}$ such that, $\forall \varepsilon > 0$,

$$\mathbb{P}\left(\left|D_\alpha(\hat{p}\|\hat{q}) - \mathbb{E}D_\alpha(\hat{p}\|\hat{q})\right| > \varepsilon\right) \leq 2\exp\left(-C_V^2\varepsilon^2 n\right).$$

Bias Bound

$$\left|\mathbb{E}D_\alpha(\hat{p}\|\hat{q}) - D_\alpha(p\|q)\right| \leq C_B\left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right).$$

Proof Sketch:

1. Main step is to analyze boundary bias of mirrored KDE:

$$\int_{\mathcal{X}} \left(\mathbb{E}\hat{p}(x) - p(x)\right)^2 dx \leq C_b h^{2\beta}.$$

2. Rest is a technical blend of standard proof techniques

Concentration Inequality

$$\mathbb{P}\left(|D_\alpha(\hat{p}\|\hat{q}) - \mathbb{E}D_\alpha(\hat{p}\|\hat{q})| > \varepsilon\right) \leq 2\exp\left(-C_V^2\varepsilon^2 n\right)$$

Proof Sketch:

Concentration Inequality

$$\mathbb{P}\left(\left|D_\alpha(\hat{p}\|\hat{q}) - \mathbb{E}D_\alpha(\hat{p}\|\hat{q})\right| > \varepsilon\right) \leq 2\exp\left(-C_V^2\varepsilon^2 n\right)$$

Proof Sketch:

- By McDiarmid's Inequality, suffices to bound change in estimator by C_V/n when resampling one data point.

Concentration Inequality

$$\mathbb{P}\left(|D_\alpha(\hat{p}\|\hat{q}) - \mathbb{E}D_\alpha(\hat{p}\|\hat{q})| > \varepsilon\right) \leq 2\exp\left(-C_V^2\varepsilon^2 n\right)$$

Proof Sketch:

- By McDiarmid's Inequality, suffices to bound change in estimator by C_V/n when resampling one data point.
- By Mean Value Theorem, change is proportional to integrated change in mirrored KDE.

Concentration Inequality

$$\mathbb{P}\left(|D_\alpha(\hat{p}\|\hat{q}) - \mathbb{E}D_\alpha(\hat{p}\|\hat{q})| > \varepsilon\right) \leq 2\exp\left(-C_V^2\varepsilon^2 n\right)$$

Proof Sketch:

- By McDiarmid's Inequality, suffices to bound change in estimator by C_V/n when resampling one data point.
- By Mean Value Theorem, change is proportional to integrated change in mirrored KDE.
- By construction of KDE, this is proportional to $2\|K\|_1^d/n$.

Consequences

- Can bound variance by integrating concentration inequality:

$$\mathbb{V}[D_\alpha(\hat{p}\|\hat{q})] \leq C_V^2 n^{-1}.$$

Consequences

- Can bound variance by integrating concentration inequality:

$$\mathbb{V}[D_\alpha(\hat{p}\|\hat{q})] \leq C_V^2 n^{-1}.$$

- Choose bandwidth h to minimize bias bound asymptotically: $h \asymp n^{-\frac{1}{\beta+d}}$. Then,
  - Bias is $O\left(n^{-\frac{\beta}{\beta+d}}\right)$
  - MSE is $O\left(n^{-\frac{2\beta}{\beta+d}} + n^{-1}\right)$
  - parametric rate $O(n^{-1})$ if $\beta \geq d$ and slower $O\left(n^{-\frac{2\beta}{\beta+d}}\right)$ else

Experiment Results

Estimated divergence between two Gaussians in $\mathbb{R}^3$.

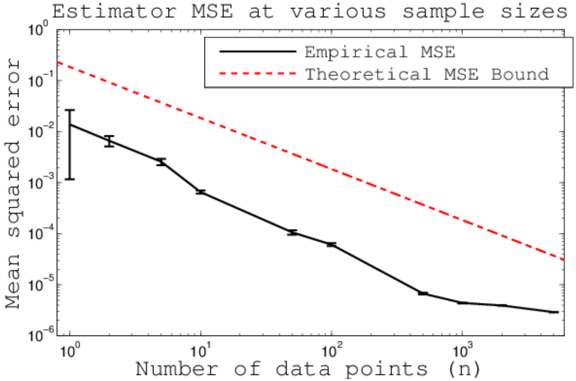


Figure : Log-log plot of empirical MSE alongside theoretical bound. Error bars indicate standard deviation of estimator from 100 trials.

Summary

- Present an estimator of Rényi-α Divergence

Summary

- Present an estimator of Rényi-α Divergence
- Prove $O\left(n^{-\frac{\beta}{\beta+d}}\right)$ bias bound

Summary

- Present an estimator of Rényi-α Divergence
- Prove $O\left(n^{-\frac{\beta}{\beta+d}}\right)$ bias bound
- Prove exponential concentration of estimator

Summary

- Present an estimator of Rényi-α Divergence
- Prove $O\left(n^{-\frac{\beta}{\beta+d}}\right)$ bias bound
- Prove exponential concentration of estimator
- Experimentally verify results

Future Work

1. Study role of dimension d
2. Prove concentration inequality for estimator of *conditional* quantities
    - e.g., Conditional Mutual Information:

$$I_\alpha(X;Y|Z) = \int_{\mathcal{Z}} D_\alpha\left(P(X,Y|Z) \| P(X|Z)P(Y|Z)\right)\, dP(Z)$$

    - hypothesis test for conditional independence

Thanks!

References

- f-Divergence estimation:
  - Nguyen, X., Wainwright, M.J., and Jordan., M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.
- k-NN estimation:
  - Poczos, B. and Schneider, J. On the estimation of alpha-divergences. In *International Conference on AI and Statistics (AISTATS)*, volume 15 of JMLR Workshop and Conference Proceedings, pp. 609-617, 2011.
- Lower bounds for single-density functional estimation:
  - Birge, L. and Massart, P. Estimation of integral functions of a density. *The Annals of Statistics*, 23:11-29, 1995.
- Distributional Machine Learning:
  - Oliva, J., Poczos, B., and Schneider, J. Distribution to distribution regression. In *International Conference on Machine Learning (ICML)*, 2013.