The Nearest Neighbor Information Estimator is Adaptively Near Minimax Rate-Optimal

Jiantao Jiao*, Weihao Gao†, Yanjun Han‡

* Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Email: jiantao@berkeley.edu

† Department of ECE, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. Email: wgao9@illinois.edu

‡ Department of Electrical Engineering, Stanford University. Email: yjhan@stanford.edu

Problem Formulation

Differential entropy of a continuous density:

$$h(f) \triangleq \int_{\mathbb{R}^d} -f(x)\log f(x)dx.$$

Applications of differential entropy:

- machine learning tasks, e.g., classification, clustering, feature selection
- other fields: causal inference, sociology, computational biology, etc.

Target: given i.i.d. samples $X_1, \cdots, X_n$ from f, estimate the value of $h(f)$.

Nearest Neighbor Estimator

Notations:

- n: number of samples
- d: dimensionality
- k: number of nearest neighbors
- $R_{i,k}$: Euclidean distance of i-th sample to its k-th nearest neighbor
- $\mathrm{vol}_d(r)$: volumn of the d-dimensional ball with radius r

Insights:

$$h(f) = \mathbb{E}[-\log f(X)] \approx -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i), \qquad f(X_i)\cdot \mathrm{vol}_d(R_{i,k}) \approx \frac{k}{n}$$

Kozachenko–Leonenko (KL) estimator [1]:

$$\hat{h}_{n,k}^{\mathrm{KL}} = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{n}{k}\mathrm{vol}_d(R_{i,k})\right) + \underbrace{\log(k) - \psi(k)}_{\text{bias correction term}}$$

Key contribution: Analyze the performance of $\hat{h}_{n,k}^{\mathrm{KL}}$ without assuming the density is bounded away from zero.

Main Result

Let $\mathcal{H}_d^s$ be the class of probability densities supported on $[0,1]^d$ which are Hölder smooth with parameter $s \ge 0$.

Main Theorem: for fixed k and $s \in (0,2]$, we have

$$\left(\sup_{f\in\mathcal{H}_d^s}\mathbb{E}_f\left(\hat{h}_{n,k}^{\mathrm{KL}} - h(f)\right)^2\right)^{\frac{1}{2}} \le C\left(n^{-\frac{s}{s+d}}\log n + n^{-\frac{1}{2}}\right)$$

where $C>0$ does not depend on n.

Significance

KL estimator is near minimax rate-optimal:

- Minimax lower bound in [2]:

$$\left(\inf_{\hat{h}}\sup_{f\in\mathcal{H}_d^s}\mathbb{E}_f\left(\hat{h} - h(f)\right)^2\right)^{\frac{1}{2}} \ge c\left(n^{-\frac{s}{s+d}}(\log n)^{-\frac{s+2d}{s+d}} + n^{-\frac{1}{2}}\right).$$

- KL estimator is near minimax within logarithmic factors

KL estimator is adaptive in s:

- Construction of $\hat{h}_{n,k}^{\mathrm{KL}}$ does not depend on s
- KL estimator adapts to unknown smoothness and (nearly) achieves the corresponding minimax rate

Different behavior for density bounded away from zero:

- Different rate from $\Theta(n^{-\frac{4s}{4s+d}} + n^{-\frac{1}{2}})$ [3] (the case where $f \ge c > 0$)

Main Tool: Maximal Inequality

Key lemma to deal with small f: define the minimal function of density f as

$$m[f](x) \triangleq \inf_{0<r\le 1}\frac{1}{\mathrm{vol}_d(r)}\int_{\|y-x\|_2\le r} f(y)dy.$$

Then there exists a constant C (depending on d only) such that for any $\varepsilon > 0$,

$$\int_{[0,1]^d} f(x)\cdot 1(f(x)\le\varepsilon)dx \le C\varepsilon.$$

Generalized Hardy–Littlewood Maximal Inequality: let μ_1, μ_2 be two Borel measures on the metric space (Ω, d), then for any $t>0$,

$$\mu_1\left\{x\in\Omega : \sup_{r>0}\frac{\mu_2(B(x;r))}{\mu_1(B(x;r))} \ge t\right\} \le C\cdot\frac{\mu_2(\Omega)}{t}.$$

Proof of the lemma: choose $\mu_2 =$ Lebesgue measure, $\mu_1(dx) = f(x)\mu_2(dx)$.

References

- [1] L. F. Kozachenko and Nikolai N Leonenko. *Sample estimate of the entropy of a random vector.* Problemy Peredachi Informatsii, 23(2):9–16, 1987.
- [2] Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. *Optimal rates of entropy estimation over lipschitz balls.* arXiv preprint arXiv:1711.02141, 2017.
- [3] James Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. *Higher order estimating equations for high-dimensional models.* The Annals of Statistics (To Appear), 2016.