Theory and Practice of Differential Entropy Estimation

Yanjun Han (Stanford EE)

Joint work with:

Weihao Gao	UIUC ECE
Jiantao Jiao	Stanford EE
Tsachy Weissman	Stanford EE
Yihong Wu	Yale Stats

July 17, 2018

Outline

Motivation

Information-theoretic measures:

- entropy $H(X)$
- mutual information $I(X; Y)$
- Kullback–Leibler divergence $D(P\|Q)$

Motivation

Information-theoretic measures:

- entropy $H(X)$
- mutual information $I(X;Y)$
- Kullback–Leibler divergence $D(P\|Q)$

Subroutine for many fields and applications:

- machine learning: classification, clustering, feature selection
- causal inference: network flow
- sociology
- computational biology
- ...

Problem Formulation

Problem:

- let f be a continuous density supported on $[0,1]^d$, belonging to some function class $\mathcal{F}$

Problem Formulation

Problem:

- let f be a continuous density supported on $[0,1]^d$, belonging to some function class $\mathcal{F}$
- observe $X^n = (X_1, \cdots, X_n) \overset{i.i.d.}{\sim} f$

Problem Formulation

Problem:

- let f be a continuous density supported on $[0,1]^d$, belonging to some function class $\mathcal{F}$
- observe $X^n = (X_1, \cdots, X_n) \overset{i.i.d.}{\sim} f$
- estimate the differential entropy of f based on X^n:

$$h(f) = \int_{[0,1]^d} -f(x)\log f(x)dx$$

Problem Formulation

Problem:

- let f be a continuous density supported on $[0,1]^d$, belonging to some function class $\mathcal{F}$
- observe $X^n = (X_1, \cdots, X_n) \overset{i.i.d.}{\sim} f$
- estimate the differential entropy of f based on X^n:

$$h(f) = \int_{[0,1]^d} -f(x)\log f(x)dx$$

Target: charactermize the minimax risk

$$|\hat{h}(X^n) - h(f)|$$

Problem Formulation

Problem:

- let f be a continuous density supported on $[0,1]^d$, belonging to some function class $\mathcal{F}$
- observe $X^n = (X_1, \cdots, X_n) \overset{i.i.d.}{\sim} f$
- estimate the differential entropy of f based on X^n:

$$h(f) = \int_{[0,1]^d} -f(x) \log f(x) dx$$

Target: charactermize the minimax risk

$$\mathbb{E}_f |\hat{h}(X^n) - h(f)|$$

Problem Formulation

Problem:

- let f be a continuous density supported on $[0,1]^d$, belonging to some function class $\mathcal{F}$
- observe $X^n = (X_1, \cdots, X_n) \overset{i.i.d.}{\sim} f$
- estimate the differential entropy of f based on X^n:

$$h(f) = \int_{[0,1]^d} -f(x) \log f(x) dx$$

Target: charactermize the minimax risk

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f |\hat{h}(X^n) - h(f)|$$

Problem Formulation

Problem:

- let f be a continuous density supported on $[0,1]^d$, belonging to some function class $\mathcal{F}$
- observe $X^n = (X_1, \cdots, X_n) \overset{i.i.d.}{\sim} f$
- estimate the differential entropy of f based on X^n:

$$h(f) = \int_{[0,1]^d} -f(x) \log f(x) dx$$

Target: charactermize the minimax risk

$$\inf_{\hat{h}} \sup_{f \in \mathcal{F}} \mathbb{E}_f |\hat{h}(X^n) - h(f)|$$

Choice of Function Class

Hölder ball $\mathcal{H}_d^s(L)$

- $0 < s \leq 1$: $|f(x) - f(y)| \leq L\|x - y\|^s$
- $1 < s \leq 2$: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|^{s-1}$
- intuition: $\|f^{(s)}\|_\infty \leq L$

Choice of Function Class

Hölder ball $\mathcal{H}_d^s(L)$

- $0 < s \le 1$: $|f(x) - f(y)| \le L\|x-y\|^s$
- $1 < s \le 2$: $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|^{s-1}$
- intuition: $\|f^{(s)}\|_\infty \le L$

Lipschitz ball (or Besov ball) $\text{Lip}_{p,d}^s(L)$

- definition: for any $t \in \mathbb{R}^d$,

$$\|f(\cdot + t) + f(\cdot - t) - 2f(\cdot)\|_p \le L\|t\|^s$$

- intuition: $\|f^{(s)}\|_p \le L$

Parameters

Reminder of important parameters:

- n: sample size
- d: dimension of support of f
- $s \in (0, 2]$: smoothness parameter of $\mathcal{F}$

Nonparametric Functional Estimation

General Problem

Given $X_1, \cdots, X_n \sim f$, we would like to estimate the functional of the form

$$I(f) = \int w(f(x))dx$$

Nonparametric Functional Estimation

General Problem

Given $X_1, \cdots, X_n \sim f$, we would like to estimate the functional of the form

$$I(f) = \int w(f(x))dx$$

Example

- quadratic functional: $I(f) = \int f(x)^2 dx$
- cubic functional: $I(f) = \int f(x)^3 dx$

Smooth Functional

- quadratic functional: elbow effect

Theorem (Bickel–Ritov'88)

$$\inf_{\hat{I}} \sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f |\hat{I} - I(f)| \asymp n^{-\frac{4s}{4s+d}} + n^{-\frac{1}{2}}$$

Smooth Functional

- quadratic functional: elbow effect

Theorem (Bickel–Ritov'88)

$$\inf_{\hat{I}} \sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f |\hat{I} - I(f)| \asymp n^{-\frac{4s}{4s+d}} + n^{-\frac{1}{2}}$$

- cubic functional: same result with much more involved estimator construction (Kerkyacharian–Picard'96, Tchetgen et al.'08)

Smooth Functional

- quadratic functional: elbow effect

Theorem (Bickel–Ritov'88)

$$\inf_{\hat{I}} \sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f |\hat{I} - I(f)| \asymp n^{-\frac{4s}{4s+d}} + n^{-\frac{1}{2}}$$

- cubic functional: same result with much more involved estimator construction (Kerkyacharian–Picard'96, Tchetgen et al.'08)
- smooth functional: reduce to linear, quadratic and cubic via Taylor expansion (Mukherjee–Newey–Robins'17)

Smooth Functional

- quadratic functional: elbow effect

Theorem (Bickel–Ritov'88)

$$\inf_{\hat{I}} \sup_{f\in\mathcal{H}_d^s} \mathbb{E}_f|\hat{I} - I(f)| \asymp n^{-\frac{4s}{4s+d}} + n^{-\frac{1}{2}}$$

- cubic functional: same result with much more involved estimator construction (Kerkyacharian–Picard'96, Tchetgen et al.'08)
- smooth functional: reduce to linear, quadratic and cubic via Taylor expansion (Mukherjee–Newey–Robins'17)
- almost nothing is known for nonsmooth functionals

Differential Entropy Estimation

Kernel-based methods:

- Joe'89
- Györfi–van der Meulen'91
- Hall–Morton'93
- Paninski–Yajima'08
- Kandasamy et al.'15
- ...

Differential Entropy Estimation

Kernel-based methods:

- Joe'89
- Györfi–van der Meulen'91
- Hall–Morton'93
- Paninski–Yajima'08
- Kandasamy et al.'15
- ...

Nearest neighbor methods:

- Tsybakov–van der Meulen'96
- Sricharan–Raich–Hero'12
- Singh–Póczos'16
- Berrett–Samworth–Yuan'16
- Delattre–Fournier'17
- Gao–Oh–Viswanath'17
- ...

Differential Entropy Estimation (Cont'd)

Drawbacks of previous works:

- extra assumption: the density f is lower bounded by a positive universal constant, e.g., $f(x) \geq 0.01$ everywhere

Differential Entropy Estimation (Cont'd)

Drawbacks of previous works:

- extra assumption: the density f is lower bounded by a positive universal constant, e.g., $f(x) \geq 0.01$ everywhere
- only prove consistency

Differential Entropy Estimation (Cont'd)

Drawbacks of previous works:

- extra assumption: the density f is lower bounded by a positive universal constant, e.g., $f(x) \geq 0.01$ everywhere
- only prove consistency
- no new lower bound beyond quadratic case

Main Result

Theorem

For any d and $p \in [2, \infty), s \in (0, 2]$, we have

$$\inf_{\hat{h}} \sup_{f \in \mathrm{Lip}^s_{p,d}} \mathbb{E}_f |\hat{h} - h(f)| \asymp (n \log n)^{-\frac{s}{s+d}} + n^{-\frac{1}{2}}$$

Main Result

Theorem

For any d and $p \in [2, \infty), s \in (0, 2]$, we have

$$\inf_{\hat{h}} \sup_{f \in \mathrm{Lip}^s_{p,d}} \mathbb{E}_f |\hat{h} - h(f)| \asymp (n \log n)^{-\frac{s}{s+d}} + n^{-\frac{1}{2}}$$

Significance

- first exact expression for the minimax rate, including sharp exponent and exact logarithmic factor

Main Result

Theorem

For any d and $p \in [2, \infty), s \in (0, 2]$, we have

$$\inf_{\hat{h}} \sup_{f \in \mathrm{Lip}_{p,d}^{s}} \mathbb{E}_f |\hat{h} - h(f)| \asymp (n \log n)^{-\frac{s}{s+d}} + n^{-\frac{1}{2}}$$

Significance

- first exact expression for the minimax rate, including sharp exponent and exact logarithmic factor
- parametric rate $n^{-\frac{1}{2}}$ requires $s \geq d$

Main Result

Theorem

For any d and $p \in [2, \infty), s \in (0, 2]$, we have

$$\inf_{\hat{h}} \sup_{f \in \mathrm{Lip}^s_{p,d}} \mathbb{E}_f |\hat{h} - h(f)| \asymp (n \log n)^{-\frac{s}{s+d}} + n^{-\frac{1}{2}}$$

Significance

- first exact expression for the minimax rate, including sharp exponent and exact logarithmic factor
- parametric rate $n^{-\frac{1}{2}}$ requires $s \geq d$
- does not use any extra assumption (e.g., boundedness of f)

Main Result

Theorem

For any d and $p \in [2, \infty), s \in (0, 2]$, we have

$$\inf_{\hat{h}} \sup_{f \in \mathrm{Lip}_{p,d}^{s}} \mathbb{E}_f |\hat{h} - h(f)| \asymp (n \log n)^{-\frac{s}{s+d}} + n^{-\frac{1}{2}}$$

Significance

- first exact expression for the minimax rate, including sharp exponent and exact logarithmic factor
- parametric rate $n^{-\frac{1}{2}}$ requires $s \geq d$
- does not use any extra assumption (e.g., boundedness of f)
- improves the best known lower bound

Idea: Two-stage Approximation

Recall

$$h(f) = \int_{[0,1]^d} -f(x) \log f(x) dx$$

Idea: Two-stage Approximation

Recall

$$h(f) = \int_{[0,1]^d} -f(x) \log f(x) dx$$

- can estimate $-f(x) \log f(x)$ for every x and then integrate

Idea: Two-stage Approximation

Recall

$$h(f) = \int_{[0,1]^d} -f(x) \log f(x) dx$$

- can estimate $-f(x) \log f(x)$ for every x and then integrate
- involves both function $f(x)$ and functional $y \mapsto -y \log y$

Idea: Two-stage Approximation

Recall

$$h(f) = \int_{[0,1]^d} -f(x) \log f(x) dx$$

- can estimate $-f(x) \log f(x)$ for every x and then integrate
- involves both function $f(x)$ and functional $y \mapsto -y \log y$
- two-stage approximation: first approximate the function and then approximate the functional

First Stage

How to estimate $f(x)$ at a given x?

First Stage

How to estimate $f(x)$ at a given x?

- no unbiased estimator...

First Stage

How to estimate $f(x)$ at a given x?

- no unbiased estimator...
- first-stage approximation: consider $f_h = f * K_h$ instead, where K_h is some kernel with bandwidth h

First Stage

How to estimate $f(x)$ at a given x?

- no unbiased estimator...
- first-stage approximation: consider $f_h = f * K_h$ instead, where K_h is some kernel with bandwidth h

Example

When $K_h(x) = \frac{1}{2h}\mathbb{1}_{[-h,h]}(x)$, we have

$$f_h(x) = \frac{1}{2h}\int_{x-h}^{x+h} f(y)dy$$

First Stage (Cont'd)

Advantages of f_h:

- close to f for small bandwidth: $\|f_h - f\|_p \lesssim h^s$

First Stage (Cont'd)

Advantages of f_h:

- close to f for small bandwidth: $\|f_h - f\|_p \lesssim h^s$
- admits an unbiased estimator:

$$\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

First Stage (Cont'd)

Advantages of f_h:

- close to f for small bandwidth: $\|f_h - f\|_p \lesssim h^s$
- admits an unbiased estimator:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\right]$$

First Stage (Cont'd)

Advantages of f_h:

- close to f for small bandwidth: $\|f_h - f\|_p \lesssim h^s$
- admits an unbiased estimator:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[K_h(x - X_i)]$$

First Stage (Cont'd)

Advantages of f_h:

- close to f for small bandwidth: $\|f_h - f\|_p \lesssim h^s$
- admits an unbiased estimator:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[K_h(x - X_i)] = \frac{1}{n}\sum_{i=1}^{n} \int K_h(x-y)f(y)dy$$

First Stage (Cont'd)

Advantages of f_h:

- close to f for small bandwidth: $\|f_h - f\|_p \lesssim h^s$
- admits an unbiased estimator:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[K_h(x - X_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int K_h(x-y)f(y)dy = f * K_h(x)\end{aligned}$$

First Stage (Cont'd)

Advantages of f_h:

- close to f for small bandwidth: $\|f_h - f\|_p \lesssim h^s$
- admits an unbiased estimator:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[K_h(x - X_i)]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \int K_h(x-y)f(y)dy = f * K_h(x)$$

First-stage approximation

Estimate $h(f_h)$ instead of $h(f)$!

Second Stage

There exists unbiased estimator for $f_h(x)^k$ for any $k = 1, 2, \cdots, n$

Second Stage

There exists unbiased estimator for $f_h(x)^k$ for any $k = 1, 2, \cdots, n$

$$K_h(x - X_1)K_h(x - X_2)\cdots K_h(x - X_k)$$

Second Stage

There exists unbiased estimator for $f_h(x)^k$ for any $k = 1, 2, \cdots, n$

$$\mathbb{E}\left[K_h(x - X_1)K_h(x - X_2)\cdots K_h(x - X_k)\right]$$

Second Stage

There exists unbiased estimator for $f_h(x)^k$ for any $k = 1, 2, \cdots, n$

$$\mathbb{E}\left[K_h(x - X_1)K_h(x - X_2)\cdots K_h(x - X_k)\right]$$
$$= \int \cdots \int K_h(x - y_1)\cdots K_h(x - y_k)f(y_1)\cdots f(y_k)dy_1\cdots dy_k$$

Second Stage

There exists unbiased estimator for $f_h(x)^k$ for any $k = 1, 2, \cdots, n$

$$
\begin{aligned}
&\mathbb{E}\left[K_h(x - X_1)K_h(x - X_2)\cdots K_h(x - X_k)\right] \\
&= \int \cdots \int K_h(x - y_1)\cdots K_h(x - y_k) f(y_1)\cdots f(y_k) dy_1 \cdots dy_k \\
&= \left(\int K_h(x - y) f(y) dy \right)^k
\end{aligned}
$$

Second Stage

There exists unbiased estimator for $f_h(x)^k$ for any $k = 1, 2, \cdots, n$

$$\begin{aligned}
&\mathbb{E}\left[K_h(x - X_1)K_h(x - X_2)\cdots K_h(x - X_k)\right] \\
&= \int \cdots \int K_h(x - y_1)\cdots K_h(x - y_k)f(y_1)\cdots f(y_k)dy_1\cdots dy_k \\
&= \left(\int K_h(x - y)f(y)dy\right)^k = f_h(x)^k
\end{aligned}$$

Second Stage

There exists unbiased estimator for $f_h(x)^k$ for any $k = 1, 2, \cdots, n$

$$\begin{aligned}
&\mathbb{E}\left[K_h(x - X_1)K_h(x - X_2)\cdots K_h(x - X_k)\right] \\
&= \int \cdots \int K_h(x - y_1)\cdots K_h(x - y_k) f(y_1)\cdots f(y_k) dy_1 \cdots dy_k \\
&= \left(\int K_h(x - y) f(y) dy\right)^k = f_h(x)^k
\end{aligned}$$

U-statistics

$$U_k(x) = \frac{1}{\binom{n}{k}} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \prod_{j=1}^{k} K_h(x - X_{i_j})$$

- efficiently computable via Newton's identity

Second Stage (Cont'd)

How to estimate $-f_h(x)\log f_h(x)$ at a given x?

Second Stage (Cont'd)

How to estimate $-f_h(x) \log f_h(x)$ at a given x?

- no unbiased estimator either...

Second Stage (Cont'd)

How to estimate $-f_h(x) \log f_h(x)$ at a given x?

- no unbiased estimator either...
- but we have unbiased estimator for all polynomials of $f_h(x)$!

Second Stage (Cont'd)

How to estimate $-f_h(x)\log f_h(x)$ at a given x?

- no unbiased estimator either...
- but we have unbiased estimator for all polynomials of $f_h(x)$!

Second-stage Approximation

Write the objective functional as

$$-f_h(x)\log f_h(x) \approx \sum_{k=0}^{K} a_k f_h(x)^k$$

then $\hat{H}(x) = \sum_{k=0}^{K} a_k U_k(x)$ is an unbiased estimator for the polynomial approximation.

Estimator Construction

- choose a suitable kernel K_h with bandwidth h, and $f_h \triangleq f * K_h$

Estimator Construction

- choose a suitable kernel K_h with bandwidth h, and $f_h \triangleq f * K_h$
- for every x, we aim to estimate $-f_h(x)\log f_h(x)$

Estimator Construction

- choose a suitable kernel K_h with bandwidth h, and $f_h \triangleq f * K_h$
- for every x, we aim to estimate $-f_h(x) \log f_h(x)$
  1. if $f_h(x)$ is small, apply the previous unbiased estimator of the polynomial approximation of $y \mapsto -y \log y$

Estimator Construction

- choose a suitable kernel K_h with bandwidth h, and $f_h \triangleq f * K_h$
- for every x, we aim to estimate $-f_h(x) \log f_h(x)$
  1. if $f_h(x)$ is small, apply the previous unbiased estimator of the polynomial approximation of $y \mapsto -y \log y$
  2. if $f_h(x)$ is large, just plug in $-\hat{f}_h(x) \log \hat{f}_h(x)$

Estimator Construction

- choose a suitable kernel K_h with bandwidth h, and $f_h \triangleq f * K_h$
- for every x, we aim to estimate $-f_h(x) \log f_h(x)$
  1. if $f_h(x)$ is small, apply the previous unbiased estimator of the polynomial approximation of $y \mapsto -y \log y$
  2. if $f_h(x)$ is large, just plug in $-\hat{f}_h(x) \log \hat{f}_h(x)$
- integrate the pointwise estimate

Error Decomposition

$$\mathbb{E}_f|\hat{h} - h(f)|$$

Error Decomposition

$$\mathbb{E}_f|\hat{h} - h(f)| \leq |h(f) - h(f_h)| + \mathbb{E}_f|\hat{h} - h(f_h)|$$

Error Decomposition

$$\begin{aligned}\mathbb{E}_f|\hat{h} - h(f)| &\leq |h(f) - h(f_h)| + \mathbb{E}_f|\hat{h} - h(f_h)| \\ &\leq |h(f) - h(f_h)| + |\mathbb{E}_f\hat{h} - h(f_h)| + \sqrt{\mathrm{Var}_f(\hat{h})}\end{aligned}$$

Error Decomposition

$$
\begin{aligned}
\mathbb{E}_f|\hat{h} - h(f)| &\le |h(f) - h(f_h)| + \mathbb{E}_f|\hat{h} - h(f_h)| \\
&\le |h(f) - h(f_h)| + |\mathbb{E}_f\hat{h} - h(f_h)| + \sqrt{\mathrm{Var}_f(\hat{h})} \\
&= \text{approx. error} + \text{bias} + \text{std}
\end{aligned}
$$

Error Decomposition

$$
\begin{aligned}
\mathbb{E}_f|\hat{h} - h(f)| &\leq |h(f) - h(f_h)| + \mathbb{E}_f|\hat{h} - h(f_h)| \\
&\leq |h(f) - h(f_h)| + |\mathbb{E}_f\hat{h} - h(f_h)| + \sqrt{\mathrm{Var}_f(\hat{h})} \\
&= \text{approx. error} + \text{bias} + \text{std} \\
&\lesssim h^s + \frac{\log n}{nh^d K^2} + \frac{2^K}{n\sqrt{h^d}}
\end{aligned}
$$

Error Decomposition

$$
\begin{aligned}
\mathbb{E}_f|\hat{h} - h(f)| &\leq |h(f) - h(f_h)| + \mathbb{E}_f|\hat{h} - h(f_h)| \\
&\leq |h(f) - h(f_h)| + |\mathbb{E}_f \hat{h} - h(f_h)| + \sqrt{\mathrm{Var}_f(\hat{h})} \\
&= \text{approx. error} + \text{bias} + \text{std} \\
&\lesssim h^s + \frac{\log n}{nh^d K^2} + \frac{2^K}{n\sqrt{h^d}}
\end{aligned}
$$

Choosing $h \asymp (n \log n)^{-\frac{1}{s+d}}$, $K \asymp \log n$ completes the proof.

Key Lemma in Bounding $|h(f_h) - h(f)|$

Inequality of Fisher Information

Let $f \in C^2(\mathbb{R})$ be supported on $[0,1]$, and $f \geq 0$ everywhere. The following inequality holds:

$$J(f) \triangleq \int \frac{(f')^2}{f} \leq C_p \|f''\|_p$$

where $1 < p \leq \infty$.

Proof of Key Lemma

Non-negativity of f:

$$0 \leq f(x+h)$$

Proof of Key Lemma

Non-negativity of f:

$$0 \leq f(x+h) \leq f(x) + hf'(x) + h\int_{x}^{x+h} |f''(y)|dy$$

Proof of Key Lemma

Non-negativity of f:

$$0 \le f(x+h) \le f(x) + hf'(x) + h\int_x^{x+h} |f''(y)|dy$$

$$0 \le f(x-h) \le f(x) - hf'(x) + h\int_{x-h}^{x} |f''(y)|dy$$

Proof of Key Lemma

Non-negativity of f:

$$0 \leq f(x+h) \leq f(x) + hf'(x) + h\int_x^{x+h} |f''(y)|dy$$

$$0 \leq f(x-h) \leq f(x) - hf'(x) + h\int_{x-h}^{x} |f''(y)|dy$$

Rearranging:

$$|f'(x)| \leq \inf_{h>0} \left[\frac{2f(x)}{h} + 2h \cdot \frac{1}{2h}\int_{x-h}^{x+h} |f''(y)|dy\right]$$

Proof of Key Lemma

Non-negativity of f:

$$0 \leq f(x+h) \leq f(x) + hf'(x) + h \int_x^{x+h} |f''(y)| dy$$

$$0 \leq f(x-h) \leq f(x) - hf'(x) + h \int_{x-h}^{x} |f''(y)| dy$$

Rearranging:

$$|f'(x)| \leq \inf_{h>0} \left[\frac{2f(x)}{h} + 2h \cdot \frac{1}{2h} \int_{x-h}^{x+h} |f''(y)| dy \right]$$

$$\leq \inf_{h>0} \left[\frac{2f(x)}{h} + 2h \cdot \sup_{r>0} \frac{1}{2r} \int_{x-r}^{x+r} |f''(y)| dy \right]$$

Proof of Key Lemma

Non-negativity of f:

$$0 \leq f(x+h) \leq f(x) + hf'(x) + h\int_x^{x+h} |f''(y)|dy$$

$$0 \leq f(x-h) \leq f(x) - hf'(x) + h\int_{x-h}^{x} |f''(y)|dy$$

Rearranging:

$$\begin{aligned}
|f'(x)| &\leq \inf_{h>0}\left[\frac{2f(x)}{h} + 2h \cdot \frac{1}{2h}\int_{x-h}^{x+h} |f''(y)|dy\right] \\
&\leq \inf_{h>0}\left[\frac{2f(x)}{h} + 2h \cdot \sup_{r>0}\frac{1}{2r}\int_{x-r}^{x+r} |f''(y)|dy\right] \\
&= 2\sqrt{f(x)M[|f''|](x)}
\end{aligned}$$

Maximal Function

Definition (Hardy–Littlewood Maximal Function)

For non-negative function h, the Hardy–Littlewood maximal function $M[h]$ is defined as

$$M[h](x) \triangleq \sup_{r>0} \frac{1}{|B(x;r)|} \int_{B(x;r)} h(y)dy.$$

Maximal Function

Definition (Hardy–Littlewood Maximal Function)

For non-negative function h, the Hardy–Littlewood maximal function $M[h]$ is defined as

$$M[h](x) \triangleq \sup_{r>0} \frac{1}{|B(x;r)|} \int_{B(x;r)} h(y)dy.$$

Theorem (Hardy–Littlewood Maximal Inequality)

The following tail bound holds:

$$\mathrm{Vol}\left\{x \in \mathbb{R}^d : M[h](x) > t\right\} \le \frac{C_d}{t} \int h(x)dx.$$

Consequently, $\|M[h]\|_p \le C_p \|h\|_p$ *for any* $p \in (1, \infty]$.

Proof of Key Lemma (Cont'd)

Recall

$$|f'(x)| \leq 2\sqrt{f(x)M[|f''|](x)}.$$

Proof of Key Lemma (Cont'd)

Recall

$$|f'(x)| \le 2\sqrt{f(x)M[|f''|](x)}.$$

Consequently,

$$\int \frac{(f')^2}{f} \le 4\|M[f'']\|_1 \le 4\|M[f'']\|_p \le 4C_p\|f''\|_p.$$

Applications of Maximal Function

- Doob's martingle inequality
- Lebesgue differentiation theorem
- Birkhoff's pointwise ergodic theorem

Summary

- two-stage approximation is optimal for differential entropy estimation
- polynomial-time estimator
- need to tune parameters h, K in practice
- requires the knowledge of s

Another View of Differential Entropy

$$h(f) = \int -f(x) \log f(x) dx$$

Another View of Differential Entropy

$$
\begin{aligned}
h(f) &= \int -f(x)\log f(x)dx \\
&= \mathbb{E}_f[-\log f(X)]
\end{aligned}
$$

Another View of Differential Entropy

$$
\begin{aligned}
h(f) &= \int -f(x)\log f(x)dx \\
&= \mathbb{E}_f[-\log f(X)] \\
&\approx \frac{1}{n}\sum_{i=1}^{n} -\log f(X_i)
\end{aligned}
$$

Another View of Differential Entropy

$$
\begin{aligned}
h(f) &= \int -f(x)\log f(x)dx \\
&= \mathbb{E}_f[-\log f(X)] \\
&\approx \frac{1}{n}\sum_{i=1}^{n} -\log f(X_i) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} -\log \hat{f}(X_i)
\end{aligned}
$$

Another View of Differential Entropy

$$
\begin{aligned}
h(f) &= \int -f(x)\log f(x)dx \\
&= \mathbb{E}_f[-\log f(X)] \\
&\approx \frac{1}{n}\sum_{i=1}^{n} -\log f(X_i) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} -\log \hat{f}(X_i)
\end{aligned}
$$

Question

How to find a good estimator $\hat{f}(X_i)$?

Nearest Neighbor Estimator

Let h_i be the distance of X_i to its nearest neighbor, we set

$$\hat{f}(X_i) \cdot \mathrm{Vol}(B(X_i; h_i)) = \frac{1}{n}.$$

Nearest Neighbor Estimator

Let h_i be the distance of X_i to its nearest neighbor, we set

$$\hat{f}(X_i) \cdot \mathrm{Vol}(B(X_i; h_i)) = \frac{1}{n}.$$

Kozachenko–Leonenko (KL) Nearest Neighbor Estimator

$$\hat{h}_{\mathrm{KL}} = \frac{1}{n} \sum_{i=1}^{n} \log \left[n \mathrm{Vol}(B(X_i; h_i)) \right] + \gamma$$

where $\gamma \approx 0.577$ is Euler's constant.

Insights behind KL Estimator

Key Observation

For each i, $\int_{B(X_i,h_i)} f(y)dy \sim \mathrm{Beta}(1, n-1)$.

Insights behind KL Estimator

Key Observation

For each i, $\int_{B(X_i,h_i)} f(y)dy \sim \mathrm{Beta}(1, n-1)$.

Consequence

Define

$$f_h(x) = \frac{1}{\mathrm{Vol}(B(x;h))} \int_{B(x;h)} f(y)dy$$

Insights behind KL Estimator

Key Observation

For each i, $\int_{B(X_i,h_i)} f(y)dy \sim \text{Beta}(1, n-1)$.

Consequence

Define

$$f_h(x) = \frac{1}{\text{Vol}(B(x;h))} \int_{B(x;h)} f(y)dy$$

we have

$$\mathbb{E}_f[\hat{h}_{\text{KL}}] - h(f) = \mathbb{E}_f\left[\log \frac{f(X)}{f_{h(X)}(X)}\right] + \underbrace{\mathbb{E}\log[n \cdot \text{Beta}(1, n-1)] + \gamma}_{=O(n^{-1})}$$

Main Result

Theorem

For any $d > 0$ and $s \in (0, 2]$, the KL estimator satisfies

$$\sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f |\hat{h}_{\mathrm{KL}} - h(f)| \lesssim n^{-\frac{s}{s+d}} \log n + n^{-\frac{1}{2}}$$

Main Result

Theorem

For any $d > 0$ *and* $s \in (0, 2]$*, the KL estimator satisfies*

$$\sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f |\hat{h}_{\mathrm{KL}} - h(f)| \lesssim n^{-\frac{s}{s+d}} \log n + n^{-\frac{1}{2}}$$

Significance

- optimal up to logarithmic factor

Main Result

Theorem

For any $d > 0$ and $s \in (0, 2]$, the KL estimator satisfies

$$\sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f |\hat{h}_{\mathrm{KL}} - h(f)| \lesssim n^{-\frac{s}{s+d}} \log n + n^{-\frac{1}{2}}$$

Significance

- optimal up to logarithmic factor
- does not use extra assumptions (e.g., boundedness of f)

Main Result

Theorem

For any $d > 0$ and $s \in (0, 2]$, the KL estimator satisfies

$$\sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f |\hat{h}_{\mathrm{KL}} - h(f)| \lesssim n^{-\frac{s}{s+d}} \log n + n^{-\frac{1}{2}}$$

Significance

- optimal up to logarithmic factor
- does not use extra assumptions (e.g., boundedness of f)
- adaptive in smoothness s

Main Result

Theorem

For any $d > 0$ *and* $s \in (0, 2]$*, the KL estimator satisfies*

$$\sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f |\hat{h}_{\mathrm{KL}} - h(f)| \lesssim n^{-\frac{s}{s+d}} \log n + n^{-\frac{1}{2}}$$

Significance

- optimal up to logarithmic factor
- does not use extra assumptions (e.g., boundedness of f)
- adaptive in smoothness s
- do not need to tune parameter

Error Analysis

- Variance of $\hat{h}_{\mathrm{KL}}$: ☺

Error Analysis

- Variance of $\hat{h}_{\text{KL}}$: ☺
- Bias of $\hat{h}_{\text{KL}}$: suffices to upper bound $\left|\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)}\right|$

Error Analysis

- Variance of $\hat{h}_{\mathrm{KL}}$: ☺
- Bias of $\hat{h}_{\mathrm{KL}}$: suffices to upper bound $\left|\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)}\right|$
    1. Upper bound $\mathbb{E}_f \log \frac{f_{h(X)}(X)}{f(X)} = \int f(x)\mathbb{E} \log \frac{f_{h(x)}(x)}{f(x)} dx$: ☺

Error Analysis

- Variance of $\hat{h}_{\mathrm{KL}}$: ☺
- Bias of $\hat{h}_{\mathrm{KL}}$: suffices to upper bound $\left|\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)}\right|$
  1. Upper bound $\mathbb{E}_f \log \frac{f_{h(X)}(X)}{f(X)} = \int f(x) \mathbb{E} \log \frac{f_{h(x)}(x)}{f(x)} dx$: ☺
  2. Upper bound $\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)} = \int f(x) \mathbb{E} \log \frac{f(x)}{f_{h(x)}(x)} dx$

Error Analysis

- Variance of $\hat{h}_{\mathrm{KL}}$: ☺
- Bias of $\hat{h}_{\mathrm{KL}}$: suffices to upper bound $\left|\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)}\right|$
  1. Upper bound $\mathbb{E}_f \log \frac{f_{h(X)}(X)}{f(X)} = \int f(x)\mathbb{E} \log \frac{f_{h(x)}(x)}{f(x)} dx$: ☺
  2. Upper bound $\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)} = \int f(x)\mathbb{E} \log \frac{f(x)}{f_{h(x)}(x)} dx$
     - If $f_{h(x)}(x)$ is large: ☺

Error Analysis

- Variance of $\hat{h}_{\mathrm{KL}}$: ☺
- Bias of $\hat{h}_{\mathrm{KL}}$: suffices to upper bound $\left|\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)}\right|$
  1. Upper bound $\mathbb{E}_f \log \frac{f_{h(X)}(X)}{f(X)} = \int f(x) \mathbb{E} \log \frac{f_{h(x)}(x)}{f(x)} dx$: ☺
  2. Upper bound $\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)} = \int f(x) \mathbb{E} \log \frac{f(x)}{f_{h(x)}(x)} dx$
     - If $f_{h(x)}(x)$ is large: ☺
     - If $f_{h(x)}(x)$ is small: ☹

Error Analysis

- Variance of $\hat{h}_{\mathrm{KL}}$: ☺
- Bias of $\hat{h}_{\mathrm{KL}}$: suffices to upper bound $\left|\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)}\right|$
  1. Upper bound $\mathbb{E}_f \log \frac{f_{h(X)}(X)}{f(X)} = \int f(x)\mathbb{E} \log \frac{f_{h(x)}(x)}{f(x)} dx$: ☺
  2. Upper bound $\mathbb{E}_f \log \frac{f(X)}{f_{h(X)}(X)} = \int f(x)\mathbb{E} \log \frac{f(x)}{f_{h(x)}(x)} dx$
     - If $f_{h(x)}(x)$ is large: ☺
     - If $f_{h(x)}(x)$ is small: ☹

Question

For small $\varepsilon > 0$, find a good upper bound of

$$\mathbb{E}\left[\int f(x)\mathbb{1}(f_{h(x)}(x) \le \varepsilon)dx\right]$$

Minimal Function

Definition (Minimal Function)

For non-negative function f supported on $[0,1]^d$, the minimal function $m[f]$ is defined as

$$m[f](x) = \inf_{0<r\leq 1} \frac{1}{|\mathrm{Vol}(B(x;r))|} \int_{B(x;r)} f(y)dy.$$

Minimal Function

Definition (Minimal Function)

For non-negative function f supported on $[0,1]^d$, the minimal function $m[f]$ is defined as

$$m[f](x) = \inf_{0<r\leq 1} \frac{1}{|\mathrm{Vol}(B(x;r))|} \int_{B(x;r)} f(y)dy.$$

Observation

$$\mathbb{E}\left[\int f(x)\mathbb{1}(f_{h(x)}(x) \leq \varepsilon)dx\right] \leq \int f(x)\mathbb{1}(m[f](x) \leq \varepsilon)dx$$

Generalized Maximal Inequality

Theorem (Generalized Maximal Inequality)

Let μ_1, μ_2 be two Borel measures on metric space $\Omega \subset \mathbb{R}^d$, then for any $t > 0$,

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

Generalized Maximal Inequality

Theorem (Generalized Maximal Inequality)

Let μ_1, μ_2 be two Borel measures on metric space $\Omega \subset \mathbb{R}^d$, then for any $t > 0$,

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

Corollary

Choose $\mu_1 =$ Lebesgue measure, $\frac{d\mu_2}{d\mu_1} = f$, we have

$$\int f(x) \mathbb{1}(m[f](x) \leq \varepsilon) dx$$

Generalized Maximal Inequality

Theorem (Generalized Maximal Inequality)

Let μ_1, μ_2 be two Borel measures on metric space $\Omega \subset \mathbb{R}^d$, then for any $t > 0$,

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

Corollary

Choose $\mu_1 =$ Lebesgue measure, $\frac{d\mu_2}{d\mu_1} = f$, we have

$$\int f(x) \mathbb{1}(m[f](x) \leq \varepsilon) dx \leq \mu_2 \left\{ x \in [0,1]^d : \sup_{r>0} \frac{\mu_1(B(x;r))}{\mu_2(B(x;r))} > \frac{1}{\varepsilon} \right\}$$

Generalized Maximal Inequality

Theorem (Generalized Maximal Inequality)

Let μ_1, μ_2 be two Borel measures on metric space $\Omega \subset \mathbb{R}^d$, then for any $t > 0$,

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

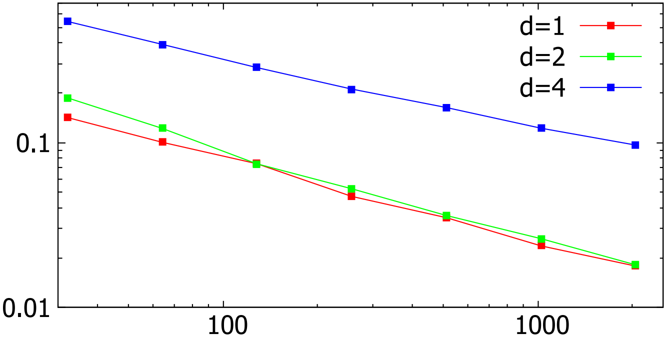
Corollary

Choose $\mu_1 =$ Lebesgue measure, $\frac{d\mu_2}{d\mu_1} = f$, we have

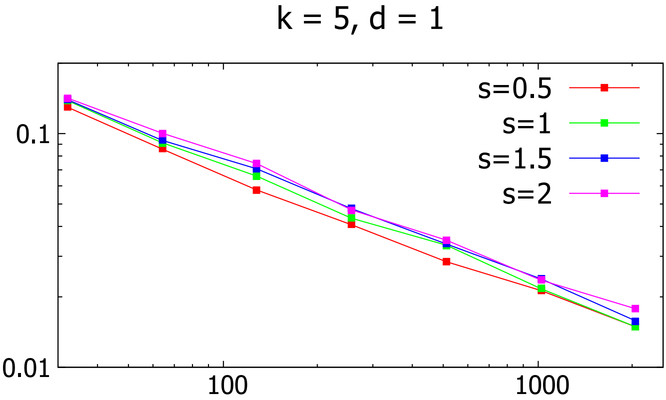
$$\int f(x)\mathbb{1}(m[f](x) \leq \varepsilon)dx \leq \mu_2\left\{x \in [0,1]^d : \sup_{r>0} \frac{\mu_1(B(x;r))}{\mu_2(B(x;r))} > \frac{1}{\varepsilon}\right\}$$
$$\leq C_d \cdot \varepsilon.$$

Dimensionality d

k = 5, s = 2



Smoothness s



Conclusion

Take-home message:

- two-stage approximation (first approximate the function, then approximate the functional) is optimal
- nearest neighbor estimator is near-optimal and adaptive to the smoothness parameter
- Hardy–Littlewood maximal inequality is crucial to deal with density close to zero

References

- Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu, "Optimal Rates of Entropy Estimation over Lipschitz Balls", *arXiv preprint, arXiv:1711.02141*
- Yanjun Han, Jiantao Jiao, Rajarshi Mukherjee, and Tsachy Weissman, "On Estimation of L_r-Norms in Gaussian White Noise Models", *arXiv preprint, arXiv:1710.03863*.
- Jiantao Jiao, Weihao Gao, and Yanjun Han, "The Nearest Neighbor Information Estimator is Adaptively Near Minimax Rate-Optimal", *arXiv preprint, arXiv:1711.08824*.