Necessary and Sufficient Conditions for ZERO-Rate Density Estimation

Jorge F. Silva[†] and Milan S. Derpich [††]

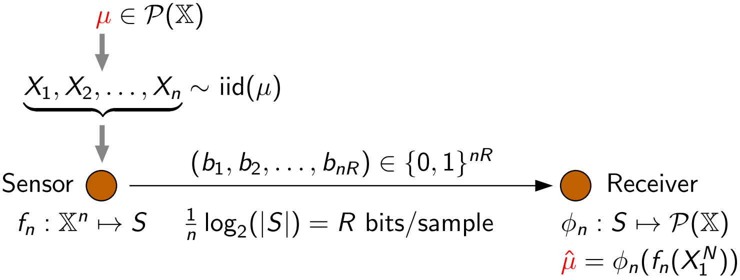
[†] Department of Electrical Engineering
Universidad de Chile
[††] Department of Electronic Engineering
Universidad Técnica Federico Santa María

Wireless Communications Symposium 2011

Outline

FOCUS: Density estimation under an operational data-rate constraint (in bits-per-sample)

FOCUS: Density estimation under an operational data-rate constraint (in bits-per-sample)



FOCUS: Density estimation under an operational data-rate constraint (in bits-per-sample)

Applications

- Remote sensors: learning under communication constraint
- Fixed-rate universal lossy source coding (FR-USC)

FOCUS: Density estimation under an operational data-rate constraint (in bits-per-sample)

FR-USC: Raginsky joint coding and modeling

- *Raginsky IEEE IT 2008*[a], introduced a connection between FR-USC and consistent density estimation under the ZERO-rate regime.
- ZERO-rate consistent estimation $\Rightarrow$ weakly mini-max universal lossy source coding.
- Results obtained for bounded parametric family of densities under some "learnability" and "regularity conditions".

[a] IEEE Trans. on IT,vol 54, no 7, pp. 3059-3077, 2008

FOCUS: Density estimation under an operational data-rate constraint (in bits-per-sample)

Contributions of this Work

- Necessary and sufficient conditions for "ZERO-rate" density estimation
- non-parametric families covered (L_1-totally bounded)
- concrete coding scheme proposed for the achievability part (Skeleton estimate by Yatracos, 1985)
- optimality of the skeleton used to derive rate of convergence results

Basic Definitions

Let $\mathbb{X} \subset \mathbb{R}^d$, and let $\mathcal{P}(\mathbb{X})$ be the collection of probability measures in $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$.

Definition: Total Variation

For v and μ in $\mathcal{P}(\mathbb{X})$ the total variational distance is

$$V(\mu, v) = \sup_{A \in \mathcal{B}(\mathbb{X})} |\mu(A) - v(A)| \,. \tag{1}$$

...when the measures have densities:

$$V(\mu, v) = \frac{1}{2} \int_{\mathbb{X}} \left| \frac{d\mu}{d\lambda}(x) - \frac{dv}{d\lambda}(x) \right| d\lambda(x). \tag{2}$$

Let $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{AC}(\mathbb{X})$ be an indexed collection of interest.

Learning Rule

A *(n, M)-learning rule* of length n and size M is a pair (f, ϕ), with $f : \mathbb{X}^n \to S$ and $\phi : S \to \Theta$, where $|S| = M$.

- $\pi = \phi \circ f : \mathbb{X}^n \to \Theta$ defines its explicit learning rule,
- $\{\phi(s) : s \in S\} \subset \Theta$ defines its codebook,
- $R(\pi) \equiv \log_2(|S|)/n$ defines its rate of the rule in bits-per-sample.

Let $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{AC}(\mathbb{X})$ be an indexed collection of interest.

Learning Scheme

A finite description learning scheme Π with rate sequence $(R_n)_{n\geq 1}$ is a collection of learning rules $\Pi = \{(f_n, \phi_n) : n \geq 1\}$ such that

$$R(\pi_n) = R_n, \text{ for all } n \geq 1. \tag{3}$$

Let $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{AC}(\mathbb{X})$ be an indexed collection of interest.

Definition: R-rate consistent estimate

The rate $R \geq 0$ is asymptotically achievable for $\mathcal{F}$, if, if there is a scheme $\Pi = \{(f_n, \phi_n) : n \geq 1\}$, with $\limsup_{n\to\infty} R(\pi_n) \leq R$ and

$$lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n}(V(\mu_{\pi_n(X_1^n)}, \mu)) = 0. \tag{4}$$

- Π is an R-rate uniformly consistent estimate for $\mathcal{F}$.

Totally Bounded Classes

$\mathcal{F}$ is L_1-totally bounded if $\forall\epsilon > 0$, there is a finite covering $\mathcal{G}_\epsilon = \{\mu_i : i = 1, .., N\}$ in $\mathcal{F}$ such that

$$\mathcal{F} \subset \bigcup_{i=1}^{N} B_\epsilon^V(\mu_i), \tag{5}$$

with $B_\epsilon^V(\mu) \equiv \left\{ v \in \mathcal{AC}(\mathbb{X}) : \frac{1}{2} \int \left| \frac{d\mu}{d\lambda}(x) - \frac{dv}{d\lambda}(x) \right| d\lambda(x) < \epsilon \right\}$ is the L_1 ball of radius ϵ centered at μ.

Definitions

- Let N_ϵ denotes the smallest integer that achieves (5).
- $\mathcal{K}(\epsilon) \equiv \log_2(N_\epsilon)$ denotes the *Kolmogorov's ϵ-entropy* of $\mathcal{F}$.

Totally Bounded Classes

$\mathcal{F}$ is L_1-totally bounded if $\forall\epsilon > 0$, there is a finite covering $\mathcal{G}_\epsilon = \{\mu_i : i = 1, .., N\}$ in $\mathcal{F}$ such that

$$\mathcal{F} \subset \bigcup_{i=1}^{N} B_\epsilon^V(\mu_i), \tag{5}$$

with $B_\epsilon^V(\mu) \equiv \left\{ v \in \mathcal{AC}(\mathbb{X}) : \frac{1}{2} \int \left| \frac{d\mu}{d\lambda}(x) - \frac{dv}{d\lambda}(x) \right| d\lambda(x) < \epsilon \right\}$ is the L_1 ball of radius ϵ centered at μ.

Definitions

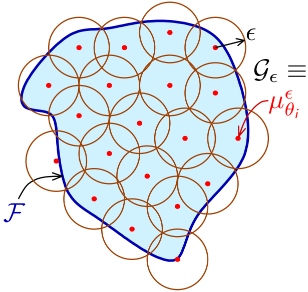
- Let N_ϵ denotes the smallest integer that achieves (5).
- $\mathcal{K}(\epsilon) \equiv \log_2(N_\epsilon)$ denotes the *Kolmogorov's ϵ-entropy* of $\mathcal{F}$.

Theorem 1

There is a "ZERO-rate" uniformly consistent scheme Π for the class $\mathcal{F}$ if, and only if, $\mathcal{F}$ is "*L_1-totally bounded*".

Idea of the proof:

Encoding



$$\mathcal{G}_\epsilon \equiv \{\mu^{\epsilon}_{\theta_i}\}_{i=1}^{2^{nR}}$$

$$\mu \in \mathcal{F}$$

$$\downarrow$$

$$X_1, X_2, \ldots, X_n \sim \mu$$

$$\downarrow$$

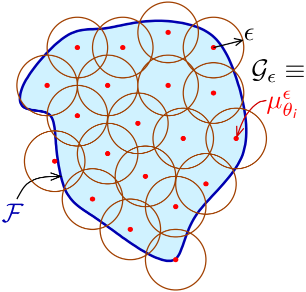
$\hat{\mu}_n$ (empirical distribution)

$$i^* \equiv \arg\min_{i \in \{1,2,\ldots,2^{nR}\}} d(\hat{\mu}_n, \mu_{\theta_i^\epsilon}) = f_n(X_1^n)$$

$$(b_1, b_2, \ldots, b_{nR})$$

Idea of the proof:

Encoding



$\mathcal{G}_\epsilon \equiv \{\mu^{\epsilon}_{\theta_i}\}_{i=1}^{2^{nR}}$

$\mu \in \mathcal{F}$

$\downarrow$

$X_1, X_2, \ldots, X_n \sim \mu$

$\downarrow$

$\hat{\mu}_n$ (empirical distribution)

$i^* \equiv \arg\min_{i \in \{1,2,\ldots,2^{nR}\}} d(\hat{\mu}_n, \mu_{\theta_i^\epsilon}) = f_n(X_1^n)$

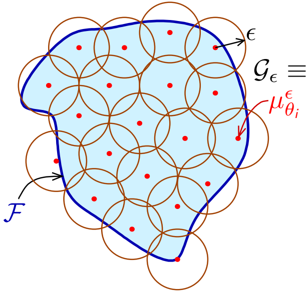
$(b_1, b_2, \ldots, b_{nR})$

Decoding

$i^* \to \hat{\mu} = \mu^{\epsilon}_{\theta_{i^*}} \in \mathcal{G}_\epsilon$

Idea of the proof:

Encoding



$\mathcal{G}_\epsilon \equiv \{\mu^\epsilon_{\theta_i}\}_{i=1}^{2^{nR}}$

$\mu \in \mathcal{F}$

$\downarrow$

$X_1, X_2, \ldots, X_n \sim \mu$

$\downarrow$

$\hat{\mu}_n$ (empirical distribution)

$i^* \equiv \arg\min_{i\in\{1,2,\ldots,2^{nR}\}} d(\hat{\mu}_n, \mu_{\theta_i^\epsilon}) = f_n(X_1^n)$

$(b_1, b_2, \ldots, b_{nR})$

Decoding

$i^* \rightarrow \hat{\mu} = \mu^\epsilon_{\theta_{i^*}} \in \mathcal{G}_\epsilon$

Error Analysis: $V(\mu, \hat{\mu})$

- $V(\mu, \tilde{\mu})$ Approximation Error
- $V(\tilde{\mu}, \hat{\mu})$ Estimation Error

$\tilde{\mu} \equiv \arg\min_{\mu_{\theta_i^\epsilon}\in\mathcal{G}_\epsilon} V(\mu, \mu_{\theta_i^\epsilon})$ is the ORACLE solution

the skeleton estimate (Yatracos, 1985)

1. Let $\mathcal{G}_\epsilon = \left\{ \mu_{\theta_i^\epsilon} : i = 1, .., N_\epsilon \right\}$ denote the ϵ-skeleton of $\mathcal{F}$.
2. Let $\Theta_\epsilon \equiv \{\theta_i^\epsilon : i = 1, .., N_\epsilon\}$ the index set of $\mathcal{G}_\epsilon$ in Θ.
3. Let us define the Yatracos class of $\mathcal{G}_\epsilon$ by $\mathcal{A}_\epsilon \equiv \left\{ A_{i,j}^\epsilon, A_{j,i}^\epsilon : 1 \leq i < j \leq N_\epsilon \right\}$, with:

$$A_{i,j}^\epsilon \equiv \left\{ x \in \mathbb{X} : \frac{d\mu_{\theta_i^\epsilon}}{d\lambda}(x) > \frac{d\mu_{\theta_j^\epsilon}}{d\lambda}(x) \right\} \subset \mathbb{X}.$$

4. Minimun distance estimate:

$$\hat{\theta}_\epsilon(X_1^n) \equiv \arg \min_{\theta_i^\epsilon \in \Theta_\epsilon} \sup_{B \in \mathcal{A}_\epsilon} \left| \mu_{\theta_i^\epsilon}(B) - \hat{\mu}_n(B) \right|,$$

with $\hat{\mu}_n$ the standard empirical measure.

the skeleton estimate (Yatracos, 1985)

1. Let $\mathcal{G}_\epsilon = \left\{ \mu_{\theta_i^\epsilon} : i = 1, .., N_\epsilon \right\}$ denote the ϵ-skeleton of $\mathcal{F}$.
2. Let $\Theta_\epsilon \equiv \{\theta_i^\epsilon : i = 1, .., N_\epsilon\}$ the index set of $\mathcal{G}_\epsilon$ in Θ.
3. Let us define the Yatracos class of $\mathcal{G}_\epsilon$ by $\mathcal{A}_\epsilon \equiv \left\{ A_{i,j}^\epsilon, A_{j,i}^\epsilon : 1 \leq i < j \leq N_\epsilon \right\}$, with:

$$A_{i,j}^\epsilon \equiv \left\{ x \in \mathbb{X} : \frac{d\mu_{\theta_i^\epsilon}}{d\lambda}(x) > \frac{d\mu_{\theta_j^\epsilon}}{d\lambda}(x) \right\} \subset \mathbb{X}.$$

4. Minimun distance estimate:

$$\hat{\theta}_\epsilon(X_1^n) \equiv \arg \min_{\theta_i^\epsilon \in \Theta_\epsilon} \sup_{B \in \mathcal{A}_\epsilon} \left| \mu_{\theta_i^\epsilon}(B) - \hat{\mu}_n(B) \right|,$$

with $\hat{\mu}_n$ the standard empirical measure.

the skeleton estimate (Yatracos, 1985)

1. Let $\mathcal{G}_\epsilon = \left\{ \mu_{\theta_i^\epsilon} : i = 1, .., N_\epsilon \right\}$ denote the ϵ-skeleton of $\mathcal{F}$.
2. Let $\Theta_\epsilon \equiv \{\theta_i^\epsilon : i = 1, .., N_\epsilon\}$ the index set of $\mathcal{G}_\epsilon$ in Θ.
3. Let us define the Yatracos class of $\mathcal{G}_\epsilon$ by

$\mathcal{A}_\epsilon \equiv \left\{ A^\epsilon_{i,j}, A^\epsilon_{j,i} : 1 \leq i < j \leq N_\epsilon \right\}$, with:

$$A^\epsilon_{i,j} \equiv \left\{ x \in \mathbb{X} : \frac{d\mu_{\theta_i^\epsilon}}{d\lambda}(x) > \frac{d\mu_{\theta_j^\epsilon}}{d\lambda}(x) \right\} \subset \mathbb{X}.$$

4. Minimun distance estimate:

$$\hat{\theta}_\epsilon(X_1^n) \equiv \arg \min_{\theta_i^\epsilon \in \Theta_\epsilon} \sup_{B \in \mathcal{A}_\epsilon} \left| \mu_{\theta_i^\epsilon}(B) - \hat{\mu}_n(B) \right|,$$

with $\hat{\mu}_n$ the standard empirical measure.

the skeleton estimate (Yatracos, 1985)

1. Let $\mathcal{G}_\epsilon = \left\{\mu_{\theta_i^\epsilon} : i = 1, .., N_\epsilon\right\}$ denote the ϵ-skeleton of $\mathcal{F}$.
2. Let $\Theta_\epsilon \equiv \{\theta_i^\epsilon : i = 1, .., N_\epsilon\}$ the index set of $\mathcal{G}_\epsilon$ in Θ.
3. Let us define the Yatracos class of $\mathcal{G}_\epsilon$ by $\mathcal{A}_\epsilon \equiv \left\{A_{i,j}^\epsilon, A_{j,i}^\epsilon : 1 \leq i < j \leq N_\epsilon\right\}$, with:

$$A_{i,j}^\epsilon \equiv \left\{x \in \mathbb{X} : \frac{d\mu_{\theta_i^\epsilon}}{d\lambda}(x) > \frac{d\mu_{\theta_j^\epsilon}}{d\lambda}(x)\right\} \subset \mathbb{X}.$$

4. Minimun distance estimate:

$$\hat{\theta}_\epsilon(X_1^n) \equiv \arg\min_{\theta_i^\epsilon \in \Theta_\epsilon} \sup_{B \in \mathcal{A}_\epsilon} \left|\mu_{\theta_i^\epsilon}(B) - \hat{\mu}_n(B)\right|,$$

with $\hat{\mu}_n$ the standard empirical measure.

Estimation-approximation error bound

Theorem (Yatracos, 1985)

$$V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| \,.$$

Estimation-approximation error bound

Theorem (Yatracos, 1985)

$$V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| \,.$$

Estimation error:

Theorem (from *Hoeffding's Inequality*, 1963)

$$\mathbb{E}_{\mathbb{P}_\mu^n} \left(\sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| \right) \leq \sqrt{\frac{\log(2N_\epsilon{}^2)}{2n}}, \ \forall \epsilon > 0, \ \forall \mu$$

Estimation-approximation error bound

Theorem (Yatracos, 1985)

$$V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| \, .$$

Then, $\sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n} \left\{ V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \right\} \leq 3\epsilon + \sqrt{\frac{8 \log(2 N_\epsilon{}^2)}{n}}$, for all $\epsilon > 0$

Estimation-approximation error bound

Theorem (Yatracos, 1985)

$$V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| .$$

Then, $\sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n} \left\{ V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \right\} \leq 3\epsilon + \sqrt{\frac{8 \log(2 N_\epsilon{}^2)}{n}}$, for all $\epsilon > 0$

Considering $\epsilon_n^* \equiv \inf \left\{ \epsilon > 0 : \log(2N_\epsilon^2) \leq \sqrt{n} \right\}$ (Devroye and Lugosi (1985))

$$\lim_{n \to \infty} \sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n} \left\{ V(\mu_{\hat{\theta}_{\epsilon_n^*}(X_1^n)}, \mu) \right\} = 0$$

$\mu_{\hat{\theta}_{\epsilon_n^*}(X_1^n)}$ uniformly consistent estimate in $\mathcal{F}$.

The ZERO-rate learning scheme

- Coding function:

$$\hat{f}_{n,\epsilon}(x_1^n) = \arg \min_{i\in\{1,..,N_\epsilon\}} \sup_{B\in\mathcal{A}_\epsilon} \left| \mu_{\theta_i^\epsilon}(B) - \hat{\mu}_n(B) \right|$$

- Decoding function: $\hat{\phi}_{n,\epsilon}(i) = \theta_i^\epsilon \in \Theta_\epsilon \subset \Theta$.

Then the Scheme $\hat{\Pi}((\epsilon_n^*)_{n\geq 1}) \equiv \left\{ (\hat{f}_{n,\epsilon_n^*}, \hat{\phi}_{n,\epsilon_n^*}) : n \geq 1 \right\}$ is ZERO-rate uniform consistent for $\mathcal{F}$

the rate is $R(\hat{\phi}_{n,\epsilon_n^*} \circ \hat{f}_{n,\epsilon_n^*}) = \frac{\log_2(N_{\epsilon_n^*})}{n}$ is $O(1/\sqrt{n})$

The ZERO-rate learning scheme

- Coding function:

$$\hat{f}_{n,\epsilon}(x_1^n) = \arg \min_{i \in \{1,..,N_\epsilon\}} \sup_{B \in \mathcal{A}_\epsilon} \left| \mu_{\theta_i^\epsilon}(B) - \hat{\mu}_n(B) \right|$$

- Decoding function: $\hat{\phi}_{n,\epsilon}(i) = \theta_i^\epsilon \in \Theta_\epsilon \subset \Theta$.

Then the Scheme $\hat{\Pi}((\epsilon_n^*)_{n\geq 1}) \equiv \left\{ (\hat{f}_{n,\epsilon_n^*}, \hat{\phi}_{n,\epsilon_n^*}) : n \geq 1 \right\}$ is ZERO-rate uniform consistent for $\mathcal{F}$

the rate is $R(\hat{\phi}_{n,\epsilon_n^*} \circ \hat{f}_{n,\epsilon_n^*}) = \frac{\log_2(N_{\epsilon_n^*})}{n}$ is $O(1/\sqrt{n})$ □

Definition

The *Yatracos class* of $\mathcal{F}$ be, $\mathcal{A}_\Theta \equiv \left\{A_{\theta,\bar{\theta}} : \theta, \bar{\theta} \in \Theta, \theta \neq \bar{\theta}\right\}$, with $A_{\theta,\bar{\theta}} \equiv \{x \in \mathbb{X} : d\mu_\theta/d\lambda(x) > d\mu_{\bar{\theta}}/d\lambda(x)\} \in \mathcal{B}(\mathbb{X})$.

Theorem 2

Let us assume that:

i) $\mathcal{F}$ is totally bounded,

ii) $\mathcal{A}_\Theta$ has a finite Vapnik and Chervonenkis dimension

iii) and $\log_2(N_{1/\sqrt{n}})$ is $o(n)$.

Then, the skeleton scheme $\hat{\Pi}((1/\sqrt{n})_{n\geq 1})$ has ZERO-rate and

$$\sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}^n_\mu} \left\{ V(\mu_{\hat{\theta}_{1/\sqrt{n}}(X_1^n)}, \mu) \right\} \text{ is } O(1/\sqrt{n}), \quad (6)$$

Definition

The *Yatracos class* of $\mathcal{F}$ be, $\mathcal{A}_\Theta \equiv \left\{A_{\theta,\bar\theta} : \theta, \bar\theta \in \Theta, \theta \neq \bar\theta\right\}$, with $A_{\theta,\bar\theta} \equiv \{x \in \mathbb{X} : d\mu_\theta/d\lambda(x) > d\mu_{\bar\theta}/d\lambda(x)\} \in \mathcal{B}(\mathbb{X})$.

Theorem 2

Let us assume that:

i) $\mathcal{F}$ is totally bounded,

ii) $\mathcal{A}_\Theta$ has a finite Vapnik and Chervonenkis dimension

iii) and $\log_2(N_{1/\sqrt{n}})$ is $o(n)$.

Then, the skeleton scheme $\hat\Pi((1/\sqrt{n})_{n\geq 1})$ has ZERO-rate and

$$\sup_{\mu\in\mathcal{F}} \mathbb{E}_{\mathbb{P}^n_\mu}\left\{V(\mu_{\hat\theta_{1/\sqrt{n}}(X_1^n)}, \mu)\right\} \text{ is } O(1/\sqrt{n}), \tag{6}$$

Sketch of the proof

Theorem (Yatracos, 1985)

$$V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| \, .$$

Sketch of the proof

Theorem (Yatracos, 1985)

$$V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} \left| \hat{\mu}_n(B) - \mu(B) \right| .$$

Estimation error:

VC Inequality

$$\mathbb{E}_{\mathbb{P}_\mu^n} \left(\sup_{B \in \mathcal{A}_\epsilon} \left| \hat{\mu}_n(B) - \mu(B) \right| \right) \leq \mathbb{E}_{\mathbb{P}_\mu^n} \left(\sup_{B \in \mathcal{A}_\Theta} \left| \hat{\mu}_n(B) - \mu(B) \right| \right) \leq c\sqrt{\frac{V}{n}},$$

Sketch of the proof

Theorem (Yatracos, 1985)

$$V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} \left| \hat{\mu}_n(B) - \mu(B) \right| .$$

Then, $\sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n} \left\{ V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \right\} \leq 3\epsilon + 4c\sqrt{\frac{V}{n}}$, for all $\epsilon > 0$

Sketch of the proof

Theorem (Yatracos, 1985)

$$V(\mu_{\hat{\theta}_\epsilon(X_1^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| .$$

Then, $\sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n} \left\{ V(\mu_{\hat{\theta}_{\epsilon_n}(X_1^n)}, \mu) \right\} \leq 3\epsilon_n + 4c\sqrt{\frac{V}{n}}$

Considering $\epsilon_n = (1/\sqrt{n})$

$$\sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n} \left\{ V(\mu_{\hat{\theta}_{\epsilon_n}(X_1^n)}, \mu) \right\} \text{ is } O(1/\sqrt{n}),$$

where $\frac{\log_2(N_{1/\sqrt{n}})}{n}$ is $o(1)$ by iii).

Raginsky's assumptions[1]:

1. Θ is a bounded set in $\mathbb{R}^k$ (parametric assumption)
2. the mapping $\Theta \to \mathcal{F}$ is locally uniformly Lipschitz (LUL)
3. the Yatracos classs $\mathcal{A}_\Theta$ has a finite VC dimension

[1] IEEE Trans. on IT,vol 54, no 7, pp. 3059-3077, 2008

Raginsky‘s assumptions[1]:

1. Θ is a bounded set in $\mathbb{R}^k$ (parametric assumption)
2. the mapping $\Theta \to \mathcal{F}$ is locally uniformly Lipschitz (LUL)
3. the Yatracos classs $\mathcal{A}_\Theta$ has a finite VC dimension

Locally Uniformly Lipschitz (Raginsky, 2008)

The mapping $\Theta \to \mathcal{F}$ is LUL, if there exists $r > 0$ and $m > 0$, such $\forall \theta \in \Theta$, $\forall \phi \in B_r(\theta)$,

$$V(\mu_\theta, \mu_\phi) \leq m \, ||\theta - \phi|| \, , \tag{7}$$

with $B_r(\theta) \subset \Theta$ the ball of radius r (with the Euclidean norm) centered at θ.

[1] IEEE Trans. on IT,vol 54, no 7, pp. 3059-3077, 2008

Raginsky's assumptions[1]:

1. Θ is a bounded set in $\mathbb{R}^k$ (parametric assumption)
2. the mapping $\Theta \to \mathcal{F}$ is locally uniformly Lipschitz (LUL)
3. the Yatracos classs $\mathcal{A}_\Theta$ has a finite VC dimension

Approximation result (from 1 y 2)

$\mathcal{F}$ is L_1-totally bounded.

[1] IEEE Trans. on IT,vol 54, no 7, pp. 3059-3077, 2008

Raginsky's assumptions[1]:

1. Θ is a bounded set in $\mathbb{R}^k$ (parametric assumption)
2. the mapping $\Theta \to \mathcal{F}$ is locally uniformly Lipschitz (LUL)
3. the Yatracos classs $\mathcal{A}_\Theta$ has a finite VC dimension

Approximation result (from 1 y 2)

In this setting, for all $\epsilon > o$ there is a uniform covering $\tilde{\Theta}_\epsilon$ of Θ, with $\tilde{N}_\epsilon \sim O(1/\epsilon^k)$, that induces an ϵ-covering $\tilde{\mathcal{G}}_\epsilon$ (in total variation) for $\mathcal{F}$.

Remark: The rate $\frac{\log_2 \tilde{N}_{1/\sqrt{n}}}{n}$ of the uniform covering associated with $\epsilon_n = 1/\sqrt{n}$ is $O(\log n/n)$ (bits-per-sample).

[1] IEEE Trans. on IT,vol 54, no 7, pp. 3059-3077, 2008

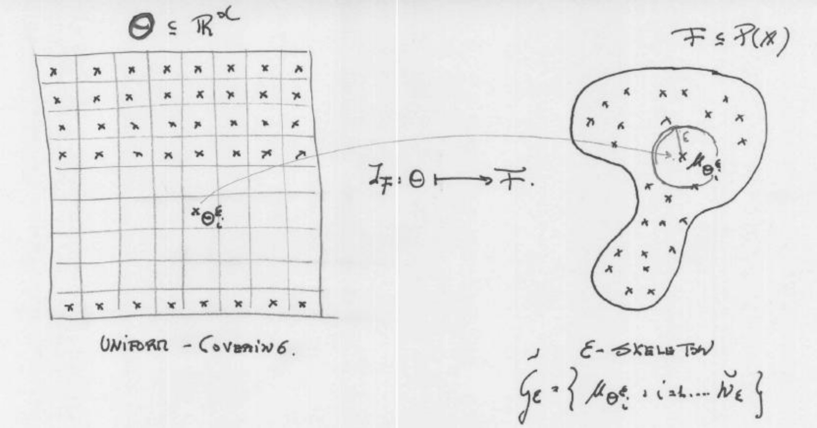


Figure: Locally uniformly Lipschitz mapping.

Adopting the practical Skeleton estimate:

$$\tilde{\theta}_\epsilon(X_1^n) = \arg\min_{\theta \in \tilde{\Theta}_\epsilon} \sup_{B \in \tilde{\mathcal{A}}_\epsilon} |\mu_\theta(B) - \hat{\mu}_n(B)|,$$

Theorem 3

The practical Skeleton scheme with $\epsilon_n = 1/\sqrt{n}$ satisfies that:

$$\sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n} \left\{ V(\mu_{\tilde{\theta}_{1/\sqrt{n}}(X_1^n)}, \mu) \right\} \text{ is } O(1/\sqrt{n}),$$

and

$$R(\tilde{\phi}_{n,1/\sqrt{n}} \circ \tilde{f}_{n,1/\sqrt{n}}) \text{ is } O(\log n/n),$$

Adopting the practical Skeleton estimate:

$$\tilde{\theta}_\epsilon(X_1^n) = \arg\min_{\theta \in \tilde{\Theta}_\epsilon} \sup_{B \in \tilde{\mathcal{A}}_\epsilon} |\mu_\theta(B) - \hat{\mu}_n(B)|,$$

Theorem 3

The practical Skeleton scheme with $\epsilon_n = 1/\sqrt{n}$ satisfies that:

$$\sup_{\mu \in \mathcal{F}} \mathbb{E}_{\mathbb{P}_\mu^n} \left\{ V(\mu_{\tilde{\theta}_{1/\sqrt{n}}(X_1^n)}, \mu) \right\} \text{ is } O(1/\sqrt{n}),$$

and

$$R(\tilde{\phi}_{n,1/\sqrt{n}} \circ \tilde{f}_{n,1/\sqrt{n}}) \text{ is } O(\log n/n),$$

Take home points...

- A Coding Theorem for the ZERO-rate density estimation is stablished.
- ZERO-rate is achievable for the large collection of L_1-totally bounded densities.
- The skeleton estimate offers a “concrete” learning-coding scheme for the problem.

Extensions (on going....)

- Study the mini-max optimality of the skeleton
- Formalize connections with universal lossy source coding
- Explore other coding-learning applications

Thank you!

Take home points...

- A Coding Theorem for the ZERO-rate density estimation is stablished.
- ZERO-rate is achievable for the large collection of L_1-totally bounded densities.
- The skeleton estimate offers a "concrete" learning-coding scheme for the problem.

Extensions (on going....)

- Study the mini-max optimality of the skeleton
- Formalize connections with universal lossy source coding
- Explore other coding-learning applications

Thank you!