Least-Squares Regression for Non-stationary Designs

David Barrera[1]

Centre de Mathématiques Appliquées (CMAP)
École Polytechnique
(Palaiseau, France)

October 27th, 2017

[1] Joint work with E.Gobet (Polytechnique, CMAP) and G.Fort (Toulouse, IMT)

Outline

Outline

The outline of the talk is the following:

Outline

The outline of the talk is the following:

- Introductory Question.

Outline

The outline of the talk is the following:

- Introductory Question.
- **Part I:** A theorem of Convergence for i.i.d. Samples.

Outline

The outline of the talk is the following:

- Introductory Question.
- **Part I:** A theorem of Convergence for i.i.d. Samples.
- **Part II:** What happens for non i.i.d. designs? (with an illustration)

Outline

The outline of the talk is the following:

- Introductory Question.
- **Part I:** A theorem of Convergence for i.i.d. Samples.
- **Part II:** What happens for non i.i.d. designs? (with an illustration)
- **Part III:** Convergence in Distribution of LSR.

Outline

The outline of the talk is the following:

- Introductory Question.
- **Part I:** A theorem of Convergence for i.i.d. Samples.
- **Part II:** What happens for non i.i.d. designs? (with an illustration)
- **Part III:** Convergence in Distribution of LSR.

All random variables are defined on $(\Omega, \mathcal{A}, \mathbb{P})$.

The Question

Given a random vector $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, *how to approximate a "regressor"* f^* *of* Y *given* X*?*

$$f^* \in \arg \min_{\{f : f \circ X \in L^2_{\mathbb{P}}\}} E[|f \circ X - Y|^2] \tag{1}$$

(so that $f^* \circ X = E[Y|X]$ *if* $Y \in L^2_{\mathbb{P}}$*).*

Part I

A Theorem of Convergence for i.i.d. Samples

Hypotheses

Hypotheses

- $\mathcal{F}$ a family of (measurable) functions $\mathbb{R}^d \to \mathbb{R}$.

Hypotheses

- $\mathcal{F}$ a family of (measurable) functions $\mathbb{R}^d \to \mathbb{R}$.
- **(New Goal)** To find, if possible

$$f^*(\mathcal{F}) \in \arg\min_{f \in \mathcal{F}} E[|f \circ X - Y|^2]$$

else $(f^{*,k})_k$ with $E[|f^{*,k} \circ X - Y|^2] \to_k \inf_{f \in \mathcal{F}} E[|f \circ X - Y|^2]$.

Hypotheses

- $\mathcal{F}$ a family of (measurable) functions $\mathbb{R}^d \to \mathbb{R}$.
- **(New Goal)** To find, if possible

$$f^*(\mathcal{F}) \in \arg\min_{f \in \mathcal{F}} E[|f \circ X - Y|^2]$$

else $(f^{*,k})_k$ with $E[|f^{*,k} \circ X - Y|^2] \to_k \inf_{f\in\mathcal{F}} E[|f \circ X - Y|^2]$.

- **(I.i.d Design)** $D_n := ((X_k, Y_k))_{k=1}^n$ an i.i.d. vector. $(X_k, Y_k) \sim (X, Y)$.

Hypotheses

- $\mathcal{F}$ a family of (measurable) functions $\mathbb{R}^d \to \mathbb{R}$.
- **(New Goal)** To find, if possible

$$f^*(\mathcal{F}) \in \arg\min_{f\in\mathcal{F}} E[|f \circ X - Y|^2]$$

else $(f^{*,k})_k$ with $E[|f^{*,k} \circ X - Y|^2] \to_k \inf_{f\in\mathcal{F}} E[|f \circ X - Y|^2]$.

- **(I.i.d Design)** $D_n := ((X_k, Y_k))_{k=1}^n$ an i.i.d. vector. $(X_k, Y_k) \sim (X, Y)$.
- **(LSR Strategy)** Given data $D_n(\omega) = ((X_k(\omega), Y_k(\omega)))_{k=1}^n$

$$\hat{f}^*(\mathcal{F}, D_n(\omega)) \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n}|f(X_k(\omega)) - Y_k(\omega)|^2. \qquad (2)$$

Heuristics

Heuristics

1. By the Law of Large Numbers

$$\frac{1}{n}\sum_{k=1}^{n}|f(X_k(\omega)) - Y_k(\omega)|^2 \approx E[|f \circ X - Y|^2] \tag{3}$$

for $\mathbb{P}-$a.e. ω and "large" $n \geq N(f,\omega)$.

Heuristics

1. By the Law of Large Numbers

$$\frac{1}{n}\sum_{k=1}^{n}|f(X_k(\omega)) - Y_k(\omega)|^2 \approx E[|f \circ X - Y|^2] \tag{3}$$

for $\mathbb{P}-$a.e. ω and "large" $n \geq N(f,\omega)$.

2. Therefore, if $N(f,\omega) = N$ is "uniform"

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n}|f(X_k(\omega)) - Y_k(\omega)|^2 \approx \inf_{f\in\mathcal{F}} E|f \circ X - Y|^2. \tag{4}$$

But are the "$\arg\inf$***'s"*** ***close also?:*** the problem of generalization.

Remarks

Remarks

1. Within $\mathcal{F}$, one cannot do better than

$$\inf_{f\in\mathcal{F}} E|f\circ X - Y|^2 - \min_{\{f:f\circ X\in L^2_{\mathbb{P}}\}} E|f\circ X - Y|^2 =$$

$$\inf_{f\in\mathcal{F}}(E|f\circ X - Y|^2 - E|E[Y|X] - Y|^2) = \inf_{f\in\mathcal{F}}(E|f\circ X - E[Y|X]|^2) \tag{5}$$

(*approximation error*).

2. The "uniformity" of N means (either)

2. The "uniformity" of N means (either)

   - The **uniform law of large numbers** (consistency)

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n}(|f\circ X_k - Y_k|^2 - E|f\circ X_k - Y_k|^2) = 0, \quad \mathbb{P}-a.s. \quad (6)$$

2. The “uniformity” of N means (either)

- The **uniform law of large numbers** (consistency)

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n}(|f\circ X_k - Y_k|^2 - E|f\circ X_k - Y_k|^2) = 0, \quad \mathbb{P}-a.s. \quad (6)$$

- **Uniform concentration inequalities** (speed of convergence)

$$\mathbb{P}[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{k=1}^{n}(|f\circ X_k - Y_k|^2 - E[|f\circ X_k - Y_k|^2])| > \delta] \leq \epsilon(n,\delta). \quad (7)$$

$\epsilon(n,\delta) \to 0$ as $n\to\infty$ for all $\delta > 0$.

2. The "uniformity" of N means (either)
   - The **uniform law of large numbers** (consistency)

$$\lim_{n\to\infty}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}(|f\circ X_k - Y_k|^2 - E|f\circ X_k - Y_k|^2) = 0, \quad \mathbb{P}-a.s. \quad (6)$$

   - **Uniform concentration inequalities** (speed of convergence)

$$\mathbb{P}[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{k=1}^{n}(|f\circ X_k - Y_k|^2 - E[|f\circ X_k - Y_k|^2])| > \delta] \leq \epsilon(n,\delta). \quad (7)$$

$\epsilon(n,\delta)\to 0$ as $n\to\infty$ for all $\delta > 0$.

This leads to assumptions on the distribution of D_n and on $\mathcal{F}$.

A Classical Result for i.i.d. Samples

Theorem ([GKKM03], Theorem 11.5)

A Classical Result for i.i.d. Samples

Theorem ([GKKM03], Theorem 11.5)

- $B \geq 1$, $||Y||_{\mathbb{P},\infty} \leq B$.

A Classical Result for i.i.d. Samples

Theorem ([GKKM03], Theorem 11.5)

- $B \geq 1$, $||Y||_{\mathbb{P},\infty} \leq B$.
- $D_n = ((X_k, Y_k))_{k=1}^{n}$ *is i.i.d.* $(X_k, Y_k) \sim (X, Y)$ *with distribution* μ_∞.

A Classical Result for i.i.d. Samples

Theorem ([GKKM03], Theorem 11.5)

- $B \geq 1$, $||Y||_{\mathbb{P},\infty} \leq B$.
- $D_n = ((X_k, Y_k))_{k=1}^n$ *is i.i.d.* $(X_k, Y_k) \sim (X, Y)$ *with distribution* μ_∞.
- $\lambda > 1$.

A Classical Result for i.i.d. Samples

Theorem ([GKKM03], Theorem 11.5)

- $B \geq 1, \ ||Y||_{\mathbb{P},\infty} \leq B.$
- $D_n = ((X_k, Y_k))_{k=1}^n$ *is i.i.d.* $(X_k, Y_k) \sim (X, Y)$ *with distribution* μ_∞.
- $\lambda > 1.$

then

$$E \int |(\hat{f}^* I_{[\hat{f}^* \leq B]}(x)) - y|^2 d\mu_\infty(x, y) \leq$$

$$C(\lambda) B^4 V_{\mathcal{F}} \frac{(1 + \log n)}{n} + \lambda \inf_{f \in \mathcal{F}} E|f \circ X - Y|^2. \quad (8)$$

$V_{\mathcal{F}} =$ *VC- dimension associated to* $\mathcal{F}$.

Remark

Remark

The estimate (8) is a consistency estimate with speed of convergence:

Remark

The estimate (8) is a consistency estimate with speed of convergence:

1. **(Consistency of the generalization)** It implies that if (X', Y') is an independent copy of (X, Y) (independent from D_n)

$$\lim_n E|\hat{f}^* \circ X' - Y'|^2 = \inf_{f \in \mathcal{F}} E|f \circ X - Y|^2,$$

(fix $\lambda > 1$, let $n \to \infty$, then let $\lambda \to 1$, then let $B \to \infty$).

Remark

The estimate (8) is a consistency estimate with speed of convergence:

1. **(Consistency of the generalization)** It implies that if (X', Y') is an independent copy of (X, Y) (independent from D_n)
$$\lim_n E|\hat{f}^* \circ X' - Y'|^2 = \inf_{f \in \mathcal{F}} E|f \circ X - Y|^2,$$
(fix $\lambda > 1$, let $n \to \infty$, then let $\lambda \to 1$, then let $B \to \infty$).

2. **(Speed of Convergence)** If the elements of $\mathcal{F}$ are bounded by B, it gives a function $N(\epsilon)$ such that
$$0 \leq E|\hat{f}^* \circ X' - Y'|^2 - \inf_{f \in \mathcal{F}} E|f \circ X - Y|^2 < \epsilon$$
if $n \geq N(\epsilon)$.

   (Fix $\epsilon > 0$ and $\lambda = \lambda(\epsilon) > 1$ such that
$$(\lambda(\epsilon) - 1) \inf_{f \in \mathcal{F}} E|f \circ X - Y|^2 < \epsilon/2).$$

Part II

*What happens if $(X_k, Y_k)_k$ is **not** an i.i.d. sequence?*

Motivation: an MCMC Example

Motivation: an MCMC Example

([FGM17]): Given

Motivation: an MCMC Example

([FGM17]): Given $\tilde{X}, Y$ (accessible) random variables and a "rare" event $\tilde{A}$ for $\tilde{X}$

$$0 < \mathbb{P}[\tilde{X} \in \tilde{A}] << 1,$$

consider $X \sim \tilde{X} | [\tilde{X} \in \tilde{A}]$:

$$\mathbb{P}[X \in A] = \frac{\mathbb{P}[\tilde{X} \in A \cap \tilde{A}]}{\mathbb{P}[\tilde{X} \in \tilde{A}]}.$$

Motivation: an MCMC Example

([FGM17]): Given $\tilde{X}, Y$ (accessible) random variables and a "rare" event $\tilde{A}$ for $\tilde{X}$

$$0 < \mathbb{P}[\tilde{X} \in \tilde{A}] << 1,$$

consider $X \sim \tilde{X} | [\tilde{X} \in \tilde{A}]$:

$$\mathbb{P}[X \in A] = \frac{\mathbb{P}[\tilde{X} \in A \cap \tilde{A}]}{\mathbb{P}[\tilde{X} \in \tilde{A}]}.$$

Problem: how do we efficiently approximate

$$E[f(X, E[Y|X])]$$

supposing the knowledge of the *conditional* probability measures

$$Q(A, x) = \mathbb{P}(Y \in A | X = x) = \mathbb{P}(Y \in A | \tilde{X} = x)?$$

Strategy (sketch):

Strategy (sketch):

- Do a sample $D_n(\omega) := (X_k(\omega))_{k=1}^n$ from a Markov Chain $(X_k)_k$ with

$$X_k \Rightarrow_k X$$

(for any initial distribution or a convenient one).

Strategy (sketch):

- Do a sample $D_n(\omega) := (X_k(\omega))_{k=1}^n$ from a Markov Chain $(X_k)_k$ with

$$X_k \Rightarrow_k X$$

(for any initial distribution or a convenient one).

- Use $Q(\cdot, x)$ to sample a corresponding $(X_k(\omega), Y_k(\omega))_k$.

Strategy (sketch):

- Do a sample $D_n(\omega) := (X_k(\omega))_{k=1}^n$ from a Markov Chain $(X_k)_k$ with

$$X_k \Rightarrow_k X$$

(for any initial distribution or a convenient one).

- Use $Q(\cdot, x)$ to sample a corresponding $(X_k(\omega), Y_k(\omega))_k$.
- *(Regression Step)* Use the approximation ***(why?)***

$$E[Y|X = \cdot] \approx \hat{h}_\omega(\cdot) := \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{k=1}^{n} |X_k(\omega) - Y_k(\omega)|^2. \qquad (9)$$

Strategy (sketch):

- Do a sample $D_n(\omega) := (X_k(\omega))_{k=1}^n$ from a Markov Chain $(X_k)_k$ with

$$X_k \Rightarrow_k X$$

(for any initial distribution or a convenient one).

- Use $Q(\cdot, x)$ to sample a corresponding $(X_k(\omega), Y_k(\omega))_k$.
- *(Regression Step)* Use the approximation ***(why?)***

$$E[Y|X = \cdot] \approx \hat{h}_\omega(\cdot) := \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{k=1}^{n} |X_k(\omega) - Y_k(\omega)|^2. \qquad (9)$$

- Use the approximation ***(why?)***

$$Ef(X, E[Y|X]) \approx \frac{1}{n} \sum_{k=1}^{n} f(X_k(\omega), \hat{h}_\omega(X_k(\omega))). \qquad (10)$$

Does this work? What is the speed of convergence (if any) of this procedure?

Does this work? What is the speed of convergence (if any) of this procedure?

Answers:

- See [FGM17] for some answers under convenient hypotheses.

Does this work? What is the speed of convergence (if any) of this procedure?

Answers:

- See [FGM17] for some answers under convenient hypotheses.
- For the ***regression step:***

Does this work? What is the speed of convergence (if any) of this procedure?

Answers:

- See [FGM17] for some answers under convenient hypotheses.
- For the ***regression step:***

  (Contribution) *Generalize [GKKM03], Theorem 11.5 using* $\beta-$**mixing coefficients** *associated to* $(X_k, Y_k)_k$.

Definition ($\beta-$mixing Coefficients.)

For sub sigma-algebras $\mathcal{A}_1$ and $\mathcal{A}_2$ of $\mathcal{A}$,

$$BETA(\mathcal{A}_1, \mathcal{A}_2) = E[\sup_{A_2 \in \mathcal{A}_2} |\mathbb{P}(A_2) - \mathbb{P}[A_2|\mathcal{A}_1]|]. \quad (11)$$

A General Theorem for LSR with Bounded Response.

A General Theorem for LSR with Bounded Response.

Setting:

A General Theorem for LSR with Bounded Response.

Setting:

- $B \geq 1$.

A General Theorem for LSR with Bounded Response.

Setting:

- $B \geq 1$.
- $(X_k, Y_k)_k$ a sequence of random vectors (**maybe *not* i.i.d**).

A General Theorem for LSR with Bounded Response.

Setting:

- $B \geq 1$.
- $(X_k, Y_k)_k$ a sequence of random vectors (**maybe *not* i.i.d**).
- $\sup_k ||Y_k||_{\mathbb{P},\infty} \leq B$.

A General Theorem for LSR with Bounded Response.

Setting:

- $B \geq 1$.
- $(X_k, Y_k)_k$ a sequence of random vectors (**maybe *not* i.i.d**).
- $\sup_k ||Y_k||_{\mathbb{P},\infty} \leq B$.
- ρ_k the distribution of (X_k, Y_k). $\mu_n := (\rho_1 + \cdots + \rho_n)/n$.
- $I_1, \ldots, I_L$ a fixed (arbitrary) **partition** of $\{1, \ldots, n\}$, $|I_k| \leq |I_{k+1}|$.

A General Theorem for LSR with Bounded Response.

Setting:

- $B \geq 1$.
- $(X_k, Y_k)_k$ a sequence of random vectors (**maybe *not* i.i.d**).
- $\sup_k ||Y_k||_{\mathbb{P},\infty} \leq B$.
- ρ_k the distribution of (X_k, Y_k). $\mu_n := (\rho_1 + \cdots + \rho_n)/n$.
- $I_1, \ldots, I_L$ a fixed (arbitrary) **partition** of $\{1, \ldots, n\}$, $|I_k| \leq |I_{k+1}|$.
- $\beta(k, j) := BETA(\sigma((X_{j'}, Y_{j'})_{j' \in I_k \cap \{1:j-1\}}), \sigma(X_j, Y_j))$: the $\beta-$mixing coefficient between time j and its past **within** I_k.

A General Theorem for LSR with Bounded Response.

Setting:

- $B \geq 1$.
- $(X_k, Y_k)_k$ a sequence of random vectors (**maybe *not* i.i.d**).
- $\sup_k ||Y_k||_{\mathbb{P},\infty} \leq B$.
- ρ_k the distribution of (X_k, Y_k). $\mu_n := (\rho_1 + \cdots + \rho_n)/n$.
- $I_1, \ldots, I_L$ a fixed (arbitrary) **partition** of $\{1, \ldots, n\}$, $|I_k| \leq |I_{k+1}|$.
- $\beta(k, j) := BETA(\sigma((X_{j'}, Y_{j'})_{j' \in I_k \cap \{1:j-1\}}), \sigma(X_j, Y_j))$: the $\beta-$mixing coefficient between time j and its past **within** I_k.
- $\mathcal{F}$ a family of functions with associated VC dimension $V_{\mathcal{F}}$.

Theorem (A Rate of Convergence for LSR with bounded Response)

In the setting of the previous slide, let

$$\hat{f}^* = \hat{f}^*(\mathcal{F}, D_n) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{k=1}^{n} |f \circ X_k - Y_k|^2,$$

Theorem (A Rate of Convergence for LSR with bounded Response)

In the setting of the previous slide, let

$$\hat{f}^* = \hat{f}^*(\mathcal{F}, D_n) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{k=1}^{n} |f \circ X_k - Y_k|^2,$$

then

$$E \int |\hat{f}^* I_{[\hat{f}^* \leq B]}(x) - y|^2 d\mu_n(x,y) \leq C(\lambda) B^4 V_{\mathcal{F}} \frac{(1 + \log L + \log |I_1|)}{|I_1|} +$$

$$8B^2(\lambda + 1) \sum_{k=1}^{L} \sum_{j \in I_k} \beta(k,j) + \lambda \inf_{f \in \mathcal{F}} \int |f(x) - y|^2 d\mu_n.$$

Example: independent, non i.d. case ($L = 1$)

Example: independent, non i.d. case ($L = 1$)

Here $\beta(k, j) = 0$ for every k, j.

Example: independent, non i.d. case ($L = 1$)

Here $\beta(k, j) = 0$ for every k, j. We get, as before, the convergence (up to a exchange of limits)

$$E \int |\hat{f}^*(x) - y|^2 d\mu_n(x, y) - \inf_{f \in \mathcal{F}} \int |f(x) - y|^2 d\mu_n(x, y) \rightarrow_{n \rightarrow \infty} 0,$$

with speed ($\sup_{(f,x) \in \mathcal{F} \times \mathbb{R}^d} |f(x)| \leq B$)

$$C(\lambda) V_{\mathcal{F}} B^4 \frac{(1 + \log n)}{n}.$$

Illustration

Illustration

$U \sim unif[-1, 1]$,

Illustration

$U \sim unif[-1, 1]$, $X = \arctan U$,

Illustration

$U \sim unif[-1, 1]$, $X = \arctan U$, $N \sim N(0, \sigma^2)$ (truncated) independent of U, N

Illustration

$U \sim unif[-1,1]$, $X = \arctan U$, $N \sim N(0,\sigma^2)$ (truncated) independent of U, N

$$Y^{(-1)} := X^2 \sin X + N, \quad Y^{(1)} := -X^2 + N$$

Illustration

$U \sim unif[-1,1]$, $X = \arctan U$, $N \sim N(0,\sigma^2)$ (truncated) independent of U, N

$$Y^{(-1)} := X^2 \sin X + N, \quad Y^{(1)} := -X^2 + N$$

$$D_n = D_{n_{-1}} \cup D_{n_1}, \quad n_{-1} + n_1 = n,$$

Illustration

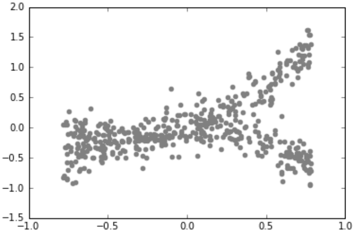
$U \sim unif[-1,1]$, $X = \arctan U$, $N \sim N(0,\sigma^2)$ (truncated) independent of U, N

$$Y^{(-1)} := X^2 \sin X + N, \quad Y^{(1)} := -X^2 + N$$

$$D_n = D_{n_{-1}} \cup D_{n_1}, \quad n_{-1} + n_1 = n,$$

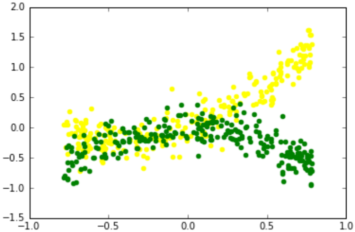
$D_{n_k} = n_k$ independent copies of $(X, Y^{(k)})$.

Unclassified Data



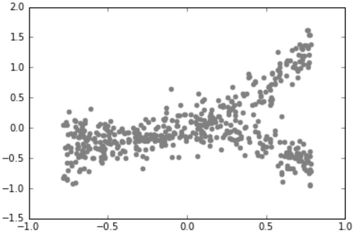
$$D_n = D_{n_{-1}} \cup D_{n_1}$$

Classified Data

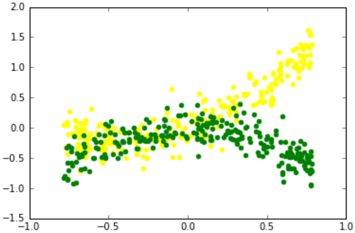


$$D_{n_1},\ D_{n_2}$$

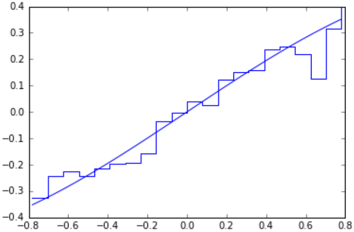
Unclassified Data



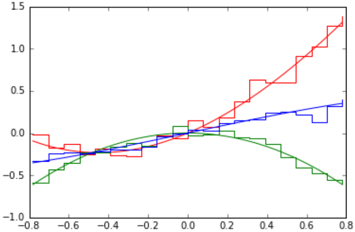
Classified Data



$D_n = D_{n_{-1}} \cup D_{n_1}$



D_{n_1}, D_{n_2}



(The blue empirical approximation is at least "as good" as the other ones).

Note:

Note: *Here* $n_1 = n_{-1}$, *and* $\hat{f}_B$ *is an estimator of*

$$E[Y|X] = \frac{1}{2}(E[Y^{(-1)}|X] + E[Y^{(1)}|X]).$$

where

- $Y := Y^{(-1)} I_{[R=-1]} + Y^{(1)} I_{[R=1]}$.
- $R =$Rademacher (independent from data).

Note: *Here* $n_1 = n_{-1}$, *and* $\hat{f}_B$ *is an estimator of*

$$E[Y|X] = \frac{1}{2}(E[Y^{(-1)}|X] + E[Y^{(1)}|X]).$$

where

- $Y := Y^{(-1)} I_{[R=-1]} + Y^{(1)} I_{[R=1]}$.
- R =Rademacher (independent from data).

Indeed:

$$\inf_{f\in\mathcal{F}} \frac{1}{n} \sum_{(X_k,Y_k)\in D_{n_{-1}}\cup D_{n_1}} E|f\circ X_k - Y_k|^2 =$$

$$\inf_{f\in\mathcal{F}} \frac{1}{2}(E|f\circ X - Y^{(-1)}|^2 + E|f\circ X - Y^{(1)}|^2) = \inf_{f\in\mathcal{F}} E|f\circ X - Y|^2.$$

Exponentially Mixing Sequences

Exponentially Mixing Sequences

Recap: Convergence of LSR for bounded Y with speed

$$C(\lambda)B^4 V_{\mathcal{F}} \frac{(1+\log L + \log |I_1|)}{|I_1|} + 8B^2(\lambda+1)\sum_{k=1}^{L}\sum_{j\in I_k}\beta(k,j).$$

Exponentially Mixing Sequences

Recap: Convergence of LSR for bounded Y with speed

$$C(\lambda)B^4 V_{\mathcal{F}} \frac{(1+\log L + \log |I_1|)}{|I_1|} + 8B^2(\lambda+1)\sum_{k=1}^{L}\sum_{j\in I_k}\beta(k,j).$$

Exercise: *Assume the* ***(sub)exponential mixing condition***

$$\beta(\sigma((X_{j'},Y_{j'})_{j'\leq j}),\sigma((X_{j+k},Y_{j+k}))) \leq ae^{-ck}, \;\; (a,c)\in[0,\infty)\times(0,\infty) \tag{12}$$

and consider the partition $I_1,\ldots,I_L$ *of* $\{1,\ldots n\}$ *where*

$$L = \left\lceil (1+\frac{1}{c})\log n \right\rceil, \quad I_k := \{jL+k\}_{j=0}^{m-1}$$

for $0\leq k < L$ *(adjust the necessary details) to prove the following:*

Theorem (Rate of convergence of LSR for Exponential Mixing Sequences)

Under (12) (and the rest of our working hypotheses):

$$E\int |\hat{f}^* I_{[\hat{f}^*\leq B]}(x) - y|^2 d\mu_n(x,y) \leq C(\lambda)B^4 V_{\mathcal{F}}(1+\frac{1}{c})^2 \log n \times$$

$$(\frac{(1+\log n)}{n} + a(1+\frac{\log n}{n})n^{-c}) + \lambda \inf_{f\in\mathcal{F}}\int |f(x)-y|^2 d\mu_n(x,y).$$

for $n \geq 2$ *such that* $e^n \geq n^{1+\frac{1}{c}}$.

First Conclusion (convergence of Averages)

Under mixing conditions (exponential, polynomial) on the data sequence $(X_k, Y_k)_k$*, and for the LSR estimator* $\hat{f}^*$ *(constructed from* $D_n = (X_k, Y_k)_{k=1}^n$*), one has the convergence (if* $\mathcal{F}$ *is a VC class and* $||Y_k||_{\mathbb{P},\infty} \leq B$*)*

$$\lim_{n\to\infty} \left(E \int |\hat{f}^*(x) - y|^2 d\mu_n(x,y) - \inf_{f\in\mathcal{F}} \int |f(x) - y|^2 d\mu_n(x,y) \right) = 0. \tag{13}$$

with an explicit rate (depending on $\lambda > 1$*) in the bounded case for an error less than*

$$(\lambda - 1) \inf_{f\in\mathcal{F}} \frac{1}{n} \sum_{k=0}^{n} E|f \circ X_k - Y_k|^2.$$

Part III

Convergence in Distribution of Least-Squares Regression

An Interpretation of the Previous Results

For (not necessarily i.i.d.) mixing data $D_n := \{(X_k, Y_k)\}_{k=1}^n$ *with uniformly bounded response* ($||Y_k||_{\mathbb{P},\infty} \leq B$)*, the LSR*

$$\hat{f}_{n,B} = \hat{f}^*(T_B\mathcal{F}, D_n)$$

is a L^2-***universally consistent*** *estimator of the* ***best*** L^2 ***approximation of*** Y ***as a function of*** X ***taken from*** $T_B\mathcal{F}$***:***

$$\hat{f}_{n,B} \approx f^*(T_B\mathcal{F}, D_n) \in \arg\min_{f \in T_B\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} E|f \circ X_k - Y_k|^2 =$$

An Interpretation of the Previous Results

For (not necessarily i.i.d.) mixing data $D_n := \{(X_k, Y_k)\}_{k=1}^n$ *with uniformly bounded response* ($||Y_k||_{\mathbb{P},\infty} \leq B$), *the LSR*

$$\hat{f}_{n,B} = \hat{f}^*(T_B\mathcal{F}, D_n)$$

is a L^2-***universally consistent*** *estimator of the* ***best*** L^2 ***approximation of*** Y ***as a function of*** X ***taken from*** $T_B\mathcal{F}$***:***

$$\hat{f}_{n,B} \approx f^*(T_B\mathcal{F}, D_n) \in \arg\min_{f \in T_B\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} E|f \circ X_k - Y_k|^2 =$$

$$\arg\min_{f \in T_B\mathcal{F}} \int_{\mathbb{R}^d \times \mathbb{R}} |f(x) - y|^2 \, d\mu_n(x, y). \tag{14}$$

An Interpretation of the Previous Results

For (not necessarily i.i.d.) mixing data $D_n := \{(X_k, Y_k)\}_{k=1}^n$ *with uniformly bounded response* ($||Y_k||_{\mathbb{P},\infty} \leq B$)*, the LSR*

$$\hat{f}_{n,B} = \hat{f}^*(T_B\mathcal{F}, D_n)$$

is a L^2-***universally consistent*** *estimator of the* ***best*** L^2 ***approximation of*** Y ***as a function of*** X ***taken from*** $T_B\mathcal{F}$***:***

$$\hat{f}_{n,B} \approx f^*(T_B\mathcal{F}, D_n) \in \arg\min_{f \in T_B\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} E|f \circ X_k - Y_k|^2 =$$

$$\arg\min_{f \in T_B\mathcal{F}} \int_{\mathbb{R}^d \times \mathbb{R}} |f(x) - y|^2 \, d\mu_n(x, y). \tag{14}$$

Note: if $(X_k, Y_k) \sim (X, Y)$ (thus $\mu_n = \mu_\infty$), the r.h.s of (14) reduces to

$$\arg\min_{f \in T_B\mathcal{F}} E[|f \circ X - Y|^2].$$

Questions on Convergence

(When) is there a limit, as $n \to \infty$, *to*

$$\inf_{T_B \mathcal{F}} \int_{\mathbb{R}^d \times \mathbb{R}} |f(x) - y|^2 d\mu_n(x, y) \, ?$$

Questions on Convergence

(When) is there a limit, as $n \to \infty$, to

$$\inf_{T_B \mathcal{F}} \int_{\mathbb{R}^d \times \mathbb{R}} |f(x) - y|^2 d\mu_n(x, y)\,?$$

If *such limit exists, is there a speed of convergence?*

Asymptotic Consistency of LSR

Asymptotic Consistency of LSR

We have seen: under mixing conditions

$$0 = \lim_n (E \int |\hat{f}_{n,B}(x) - y|^2 \, d\mu_n(x,y) - \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\mu_n(x,y)).$$

Asymptotic Consistency of LSR

We have seen: under mixing conditions

$$0 = \lim_n (E \int |\hat{f}_{n,B}(x) - y|^2 \, d\mu_n(x,y) - \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\mu_n(x,y)).$$

Let μ be a measure. Assuming the "diagonal" convergence

$$0 = \lim_n (E \int |\hat{f}_{n,B}(x) - y|^2 \, d\mu_n(x,y) - E \int |\hat{f}_{n,B}(x) - y|^2 \, d\mu(x,y)), \quad (15)$$

we get $0 = \lim_n (E[\int |\hat{f}_{n,B}(x) - y|^2 \, d\mu] - \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\mu_n(x,y))$.

Asymptotic Consistency of LSR

We have seen: under mixing conditions

$$0 = \lim_n \Big(E \int |\hat{f}_{n,B}(x) - y|^2 \, d\mu_n(x,y) - \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\mu_n(x,y) \Big).$$

Let μ be a measure. Assuming the "diagonal" convergence

$$0 = \lim_n \Big(E \int |\hat{f}_{n,B}(x) - y|^2 \, d\mu_n(x,y) - E \int |\hat{f}_{n,B}(x) - y|^2 \, d\mu(x,y) \Big), \quad (15)$$

we get $0 = \lim_n (E[\int |\hat{f}_{n,B}(x) - y|^2 \, d\mu] - \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\mu_n(x,y))$.

If in addition "$\lim_n \inf_{T_B \mathcal{F}} = \inf_{f \in T_B \mathcal{F}} \lim_n$" we arrive at

$$\lim_n E[\int |\hat{f}_{n,B}(x) - y|^2 \, d\mu(x,y)] = \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\mu(x,y). \quad (16)$$

Theorem (Convergence in Distribution of LSR)

Theorem (Convergence in Distribution of LSR)

Assume that $(X_k, Y_k)_k$, $\hat{f}_{n,B}$ *is as above* $((X_k, Y_k)$ $\beta-$*mixing,* $||Y_k||_{\mathbb{P},\infty} \leq B$, *etc.) and let*

$$\mu_n := \frac{1}{n}\sum_{k=1}^{n} \mu_{(X_k,Y_k)}$$

be the average measure at time n.

Theorem (Convergence in Distribution of LSR)

Assume that $(X_k, Y_k)_k$, $\hat{f}_{n,B}$ *is as above* ((X_k, Y_k) $\beta-$*mixing*, $||Y_k||_{\mathbb{P},\infty} \leq B$, *etc.) and let*

$$\mu_n := \frac{1}{n} \sum_{k=1}^{n} \mu_{(X_k, Y_k)}$$

be the average measure at time n. ***If*** μ_n ***converges to*** μ ***in total variation distance****, then*

$$\lim_n E \int |\hat{f}_{n,B}(x) - y|^2 d\mu(x,y) = \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\mu(x,y).$$

Theorem (Convergence in Distribution of LSR)

Assume that $(X_k, Y_k)_k$, $\hat{f}_{n,B}$ is as above ((X_k, Y_k) $\beta-$mixing, $||Y_k||_{\mathbb{P},\infty} \leq B$, etc.) and let

$$\mu_n := \frac{1}{n} \sum_{k=1}^{n} \mu_{(X_k, Y_k)}$$

be the average measure at time n. ***If μ_n converges to μ in total variation distance****, then*

$$\lim_n E \int |\hat{f}_{n,B}(x) - y|^2 d\mu(x,y) = \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\mu(x,y).$$

Proof: Convergence in TVD implies

$$\lim_n \sup_{f \in T_B \mathcal{F}} \left| \int |f(x) - y|^2 d\mu_n(x,y) - \int |f(x) - y|^2 d\mu(x,y) \right| = 0.$$

This implies the exchange of "lim" and "inf".

Speed of Convergence

Speed of Convergence

Speed of convergence can be obtained assuming control on $||\mu_n - \mu||_{TV}$.

Speed of Convergence

Speed of convergence can be obtained assuming control on $||\mu_n - \mu||_{TV}$.

Example: *A Markov Kernel Q satisfies the* ***Doeblin condition*** *if there exists* $(\delta, m) \in (0,1) \times \mathbb{N}^*$ *such that*

$$L_{Q^m} := \sup_{x_1 \neq x_2} ||Q^m(x_1, \cdot) - Q^m(x_2, \cdot)||_{TV} < \delta.$$

Speed of Convergence

Speed of convergence can be obtained assuming control on $||\mu_n - \mu||_{TV}$.

Example: *A Markov Kernel Q satisfies the* ***Doeblin condition*** *if there exists* $(\delta, m) \in (0,1) \times \mathbb{N}^*$ *such that*

$$L_{Q^m} := \sup_{x_1 \neq x_2} ||Q^m(x_1, \cdot) - Q^m(x_2, \cdot)||_{TV} < \delta.$$

Under the Doeblin condition, there exists a unique probability measure π with $Q\pi = \pi$ and for every probability measure π'

$$||\pi' Q^n - \pi||_{TV} \leq ||\pi' - \pi||_{TV} \, \delta^{\lfloor n/m \rfloor}.$$

Theorem (LSR under the Doeblin Condition)

Assume that $(X_k, Y_k)_k$ is an homogeneous (perhaps non-stationary) Markov chain satisfying the Doeblin Condition. Then if π is the unique stationary distribution of $(X_k, Y_k)_k$, there exists $(a, c) \in [0, \infty) \times (0, \infty)$ such that for all $\lambda > 1$

$$E \int |\hat{f}_{n,B}(x) - y|^2 \, d\pi(x, y) \leq$$

$$C(\lambda) B^4 V_{\mathcal{F}} (1 + \frac{1}{c})^2 \log n \times (\frac{(1 + \log n)}{n} + a(1 + \frac{\log n}{n}) n^{-c}) +$$

$$\lambda \inf_{f \in T_B \mathcal{F}} \int |f(x) - y|^2 d\,\pi(x, y).$$

Thank you!

References

Fort, G.; Gobet, E. and Moulines, E. (2017) MCMC Design-Based Non-Parametric Regression for Rare Event. Application for Nested Risk Computations. To appear in *Monte Carlo Methods Appl.*

Györfi, L; Kohler,M; Krzyzak, A and Walk, H (2002). A Distribution-Free Theory of Nonparametric Regression. *Springer Ser. Statist.*

Mojirsheibani, M. and Ren, Q. (2010) A Note on Nonparametric Regression with $\beta-$Mixing Sequences. *Commun Stat Theory Methods.* **39**. Pp. 2280–2287.