Non Parametric Stochastic Approximation

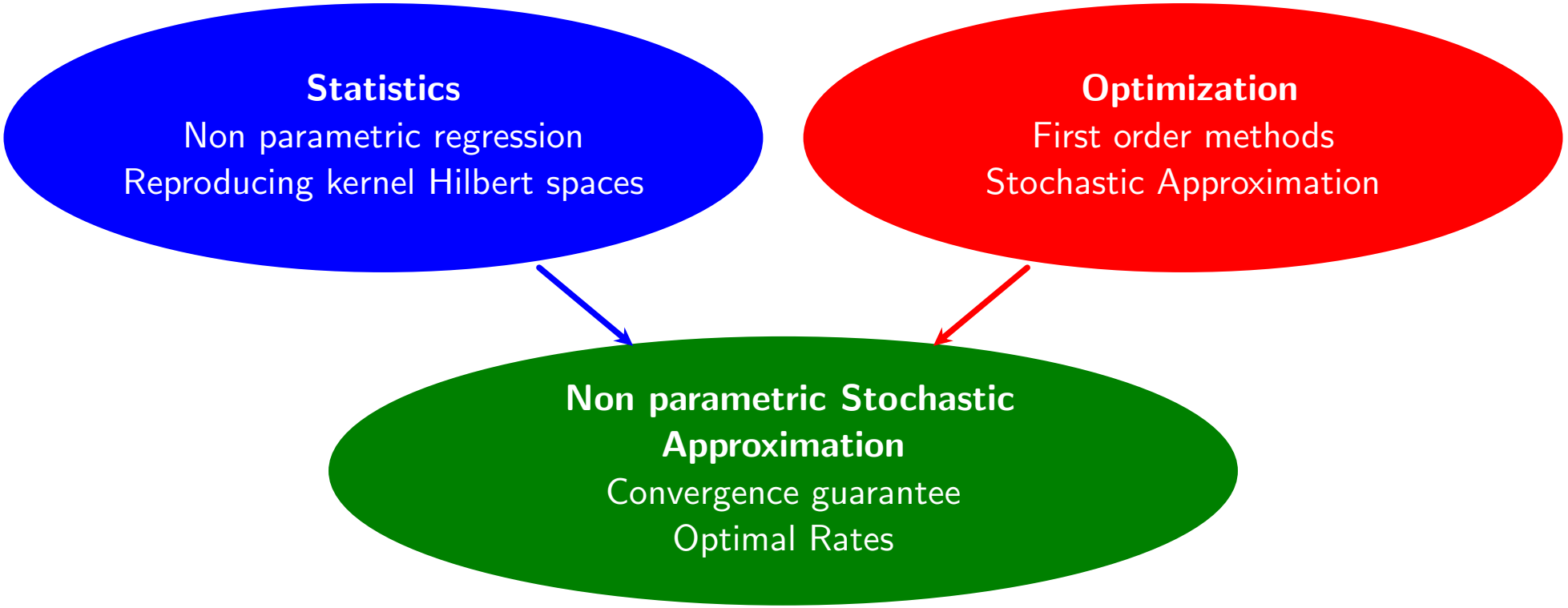
Aymeric Dieuleveut

INRIA - Ecole Normale Supérieure, Paris, France

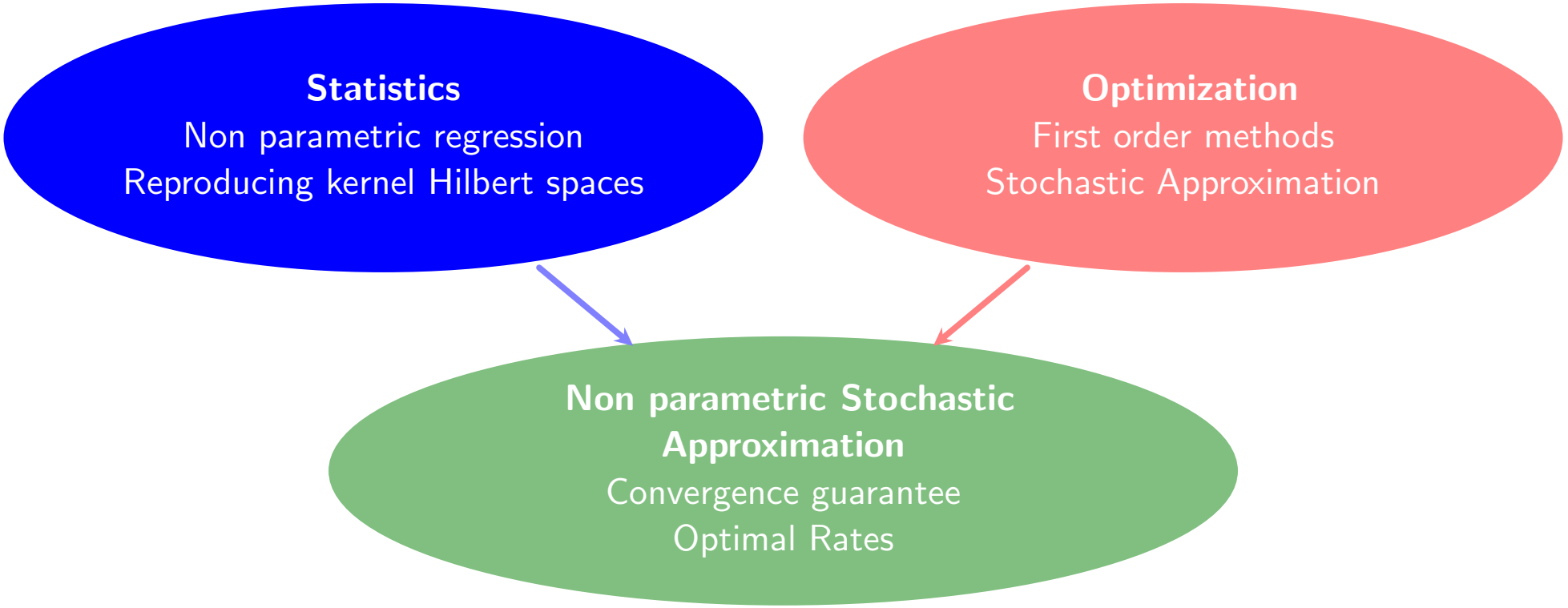
Joint work with Francis Bach

RdJS - Porquerolles, April 2017

Outline


Statistics
Non parametric regression
Reproducing kernel Hilbert spaces
Optimization
First order methods
Stochastic Approximation
Non parametric Stochastic Approximation
Convergence guarantee
Optimal Rates

Outline


Statistics
Non parametric regression
Reproducing kernel Hilbert spaces
Optimization
First order methods
Stochastic Approximation
Non parametric Stochastic Approximation
Convergence guarantee
Optimal Rates

Non parametric regression

Space $\mathcal{X}$; a random variable (X, Y) in $\mathcal{X} \times \mathbb{R}$.
ρ its distribution , and $\rho_{\mathcal{X}}$ the marginal distribution on $\mathcal{X}$.

Least squares regression: Find function with minimal prediction error

$$R(\theta) := \mathbb{E}_\rho[(Y - \theta(X))^2].$$

$$\arg\min_{\theta \in \mathcal{L}^2} R(\theta)$$

$\theta : \mathcal{X} \to \mathbb{R}$ function!

Bayes predictor $\theta_\rho(X) = \mathbb{E}[Y|X]$, $\theta_\rho \in \mathcal{L}^2$. with

$$\mathcal{L}^2 = \left\{ f : \mathcal{X} \to \mathbb{R} / \int f^2(t) d\rho_{\mathcal{X}}(t) < \infty \right\}.$$

Observe n i.i.d. samples $(X_i, Y_i)_{1 \leq i \leq n} \sim \rho^{\otimes n}$.

Possible estimators

Some classical estimators:

- Nadaraya-Watson estimator
- Locally polynomial (Tsybakov)
- Reproducing kernel Hilbert space.

Goal: Statistically robust and efficient to compute.

Reproducing kernel Hilbert space

We denote $\mathcal{H}_K$ a Hilbert space of function. $\mathcal{H}_K \subset \mathbb{R}^{\mathcal{X}}$.

Which is characterized by the kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

- for any x, $K_x : \mathcal{X} \to \mathbb{R}$ defined by $K_x(x') = K(x, x')$ is in $\mathcal{H}_K$.
- reproducing property: for all $\theta \in \mathcal{H}_K$ and $x \in \mathcal{X}$, $\theta(x) = \langle \theta, K_x \rangle_K$.

Other points of view:

- for any positive definite kernel (i.e., function s.t. $\forall (X_i)_{i=1..n} \in \mathcal{X}^n, a_i \in \mathbb{R}^n, \sum_{i=1..n, j=1..n} a_i a_j K(X_i, X_j) \geq 0$), there exists such an RKHS.

Why RKHS?

A few reasons:

- They are used in practice in many domains (image analysis, genomics) to transform non vectorial inputs into vectorial features.

- Analyzing stochastic approximation in infinite dimension allows to understand it in dimension $d \gg n$, which happens in many practical applications of machine learning.

- For regression, RKHS are suitable from
  - the statistical point of view (we can approximate many functions from functions in RKHS, especially for "universal" kernels)
  - the computational point of view (it is possible to compute inner products in the RKHS in finite time (one evaluation of the Kernel function)).

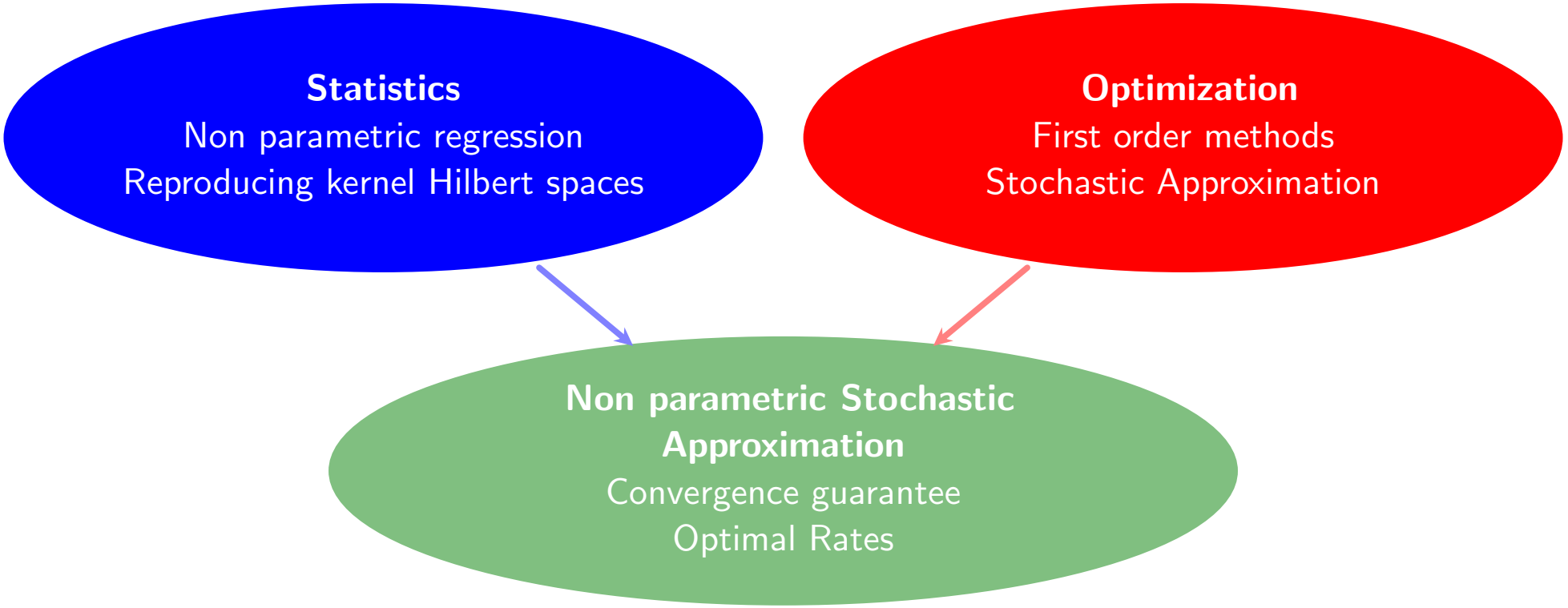
Possible estimators

What type of estimators can we consider?

- Locally polynomial
- Nadaraya-Watson estimator
- Reproducing kernel Hilbert space:
  - Empirical risk minimization $\hat{\theta} = \arg\min_{\theta \in \mathcal{H}_K} \sum_{i=1}^{n} (Y_i - \theta(X_i))^2$.
    $\leadsto$ Representer theorem: $\hat{\theta} = \sum_{i=1}^{n} a_i K_{X_i}$, with $a = K^{-1} Y$.
  - ERM + Tikhonov regularization.

Cost: $O(n^3) \rightarrow$:(

Outline


Statistics
Non parametric regression
Reproducing kernel Hilbert spaces
Optimization
First order methods
Stochastic Approximation
Non parametric Stochastic Approximation
Convergence guarantee
Optimal Rates

First order methods: use only gradients.

In the next few slides, think of $\theta \in \mathcal{H}$ as a vector, $\theta(X) = \langle \theta, K_X \rangle$.

With the reproducing property:

$$R(\theta) = \mathbb{E}_\rho[(Y - \langle \theta, K_X \rangle)^2].$$

We only observe the empirical loss:

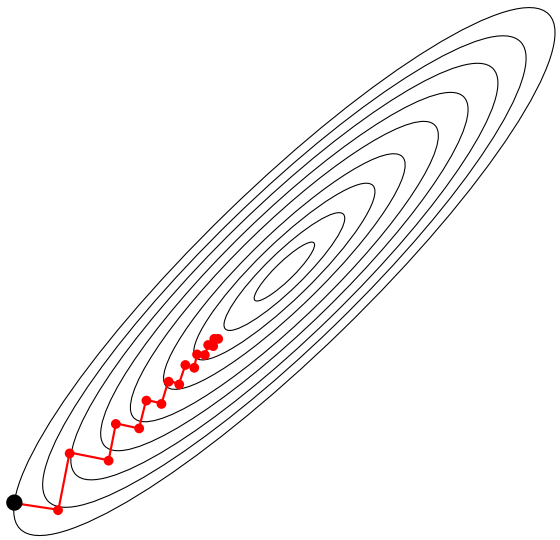
$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \theta, K_{X_i} \rangle)^2.$$

Rk: Generalization to any smooth loss ℓ ($R(\theta) = \mathbb{E}_\rho[\ell(Y, \langle \theta, K_X \rangle)]$).

Gradient descent $\rightarrow$ Stochastic gradient descent $\rightarrow$ Stochastic approximation

Gradient descent

- **Setting**: $\hat{R}$ convex and L-smooth on $\mathbb{R}^d$
- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma_t \nabla \hat{R}(\theta_{t-1})$
- **Rate**: $\hat{R}(\theta_t) - \hat{R}(\theta_*) \leqslant \frac{1}{t} L \|\theta_0 - \theta_*\|^2$
- **Computation:** at each step, compute $\sum_{i=1}^n \nabla \ell(Y_i, \langle \theta_{t-1}, K_{X_i} \rangle)$.
- **Cost:** $O(nd)$ per step. **Minimizes** $\hat{R}$.

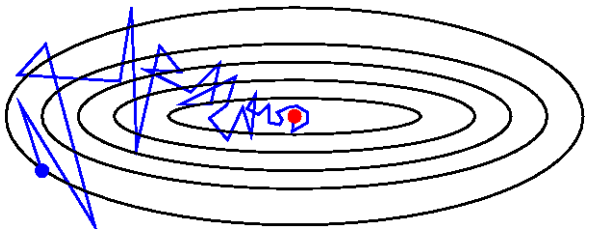


Stochastic gradient descent

Too slow! Let's take random gradients:

- **Stochastic gradient descent**: at each step t,
  - sample $i_t \sim \mathcal{U}[|1; n|]$,
  - use $\nabla \ell(y_{i_t}, \langle \theta_{t-1}, K_{x_{i_t}} \rangle)$.
- **Cost:** $O(d)$ per step. Minimizes $\hat{R}$, as

$$\mathbb{E}_{i_t}[\nabla \ell(y_{i_t}, \langle \theta_{t-1}, K_{x_{i_t}} \rangle)] = \nabla \hat{R}(\theta_{t-1}).$$



Stochastic Approximation 1/2

↬No need to optimize below statistical error

↬Testing error is more important than training error

- **Idea:** Minimizing the function R directly!
  Given only unbiased estimates $\nabla R_n(\theta_{n-1})$ of its gradients $\nabla R(\theta_{n-1})$ at certain points $\theta_{n-1} \in \mathbb{R}^d$:
  If (X_n, Y_n) is independent of θ_{n-1},

$$\nabla R_n(\theta_{n-1}) := \nabla \ell(Y_n, \langle \theta_{n-1}, X_n \rangle)$$

  is an unbiased estimate of gradient $\nabla R(\theta_{n-1})$ (conditionally to θ_{n-1}).

- **Same as SGD, but only one pass on the data to keep independence.**

Stochastic Approximation 2/2

- **Key algorithm:** Stochastic Approximation (a.k.a., Robbins-Monro)

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n \nabla R_n(\theta_{n-1})}$$

  - Averaging: $\bar{\theta}_n = \frac{1}{n+1} \sum_{k=0}^{n} \theta_k$
  - Learning rate sequence $\boxed{\gamma_n = C n^{-\zeta}}$

- **Running-time** $O(d)$ per iteration, $O(nd)$ after single pass through the data.

Take home messages:

Minimizes R

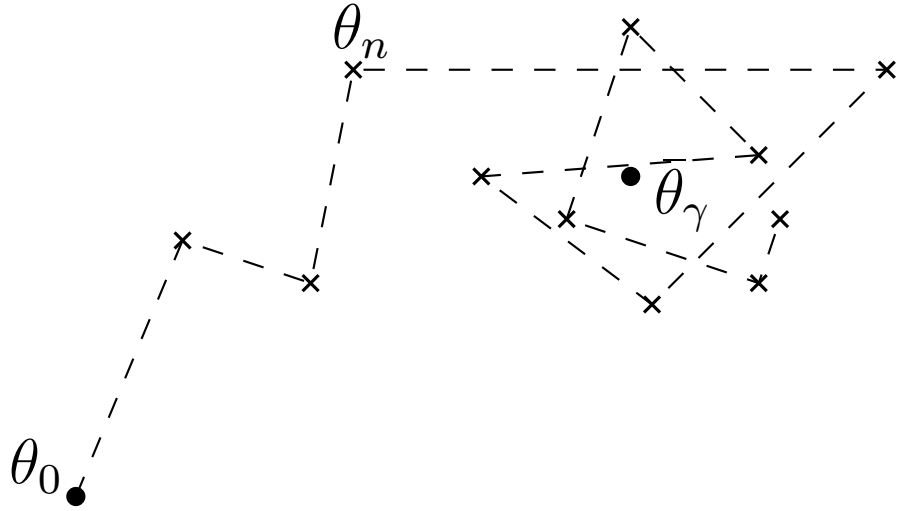
SA avoids over-fitting.

Markov chain interpretation of constant step sizes

- SA for LS, with $R_n(\theta) = \frac{1}{2}\big(Y_n - \langle K_{X_n}, \theta\rangle\big)^2$

$$\theta_n = \theta_{n-1} - \gamma\big(\langle K_{X_n}, \theta_{n-1}\rangle - Y_n\big)K_{X_n}$$

- The sequence $(\theta_n)_n$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \overset{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$

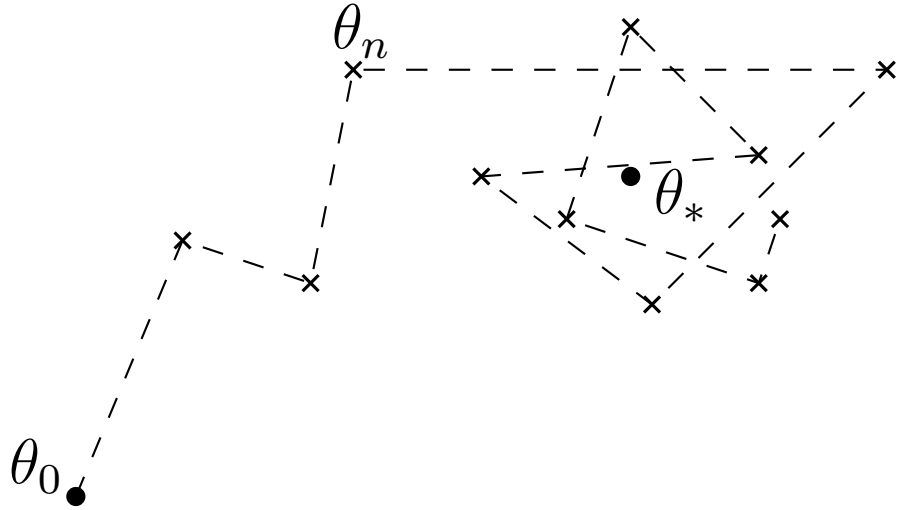


Markov chain interpretation of constant step sizes

- SA for LS, with $R_n(\theta) = \frac{1}{2}\big(Y_n - \langle K_{X_n}, \theta\rangle\big)^2$

$$\theta_n = \theta_{n-1} - \gamma\big(\langle K_{X_n}, \theta_{n-1}\rangle - Y_n\big) K_{X_n}$$

- The sequence $(\theta_n)_n$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$
- **For least-squares,** $\bar{\theta}_\gamma = \theta_*$

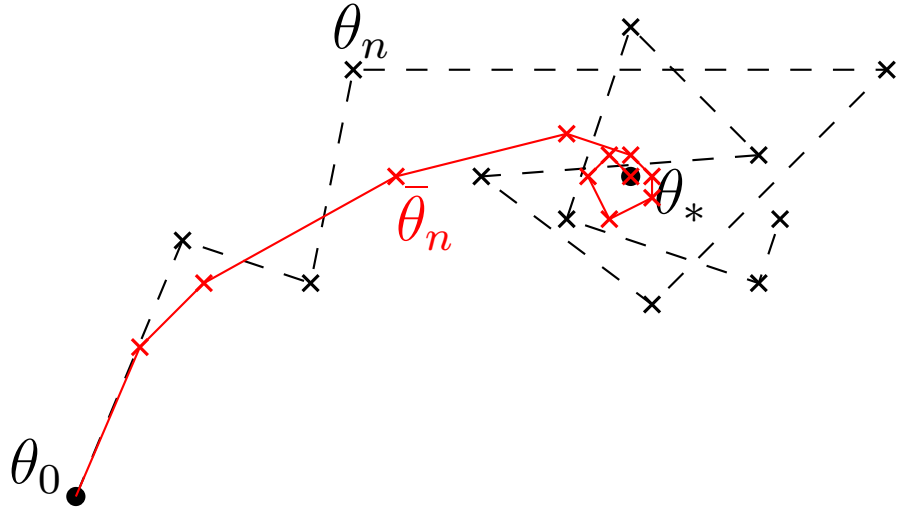


Markov chain interpretation of constant step sizes

- SA for LS, with $R_n(\theta) = \frac{1}{2}\big(Y_n - \langle K_{X_n}, \theta\rangle\big)^2$

$$\theta_n = \theta_{n-1} - \gamma\big(\langle K_{X_n}, \theta_{n-1}\rangle - Y_n\big)K_{X_n}$$

- The sequence $(\theta_n)_n$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \overset{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$
- **For least-squares,** $\bar{\theta}_\gamma = \theta_*$



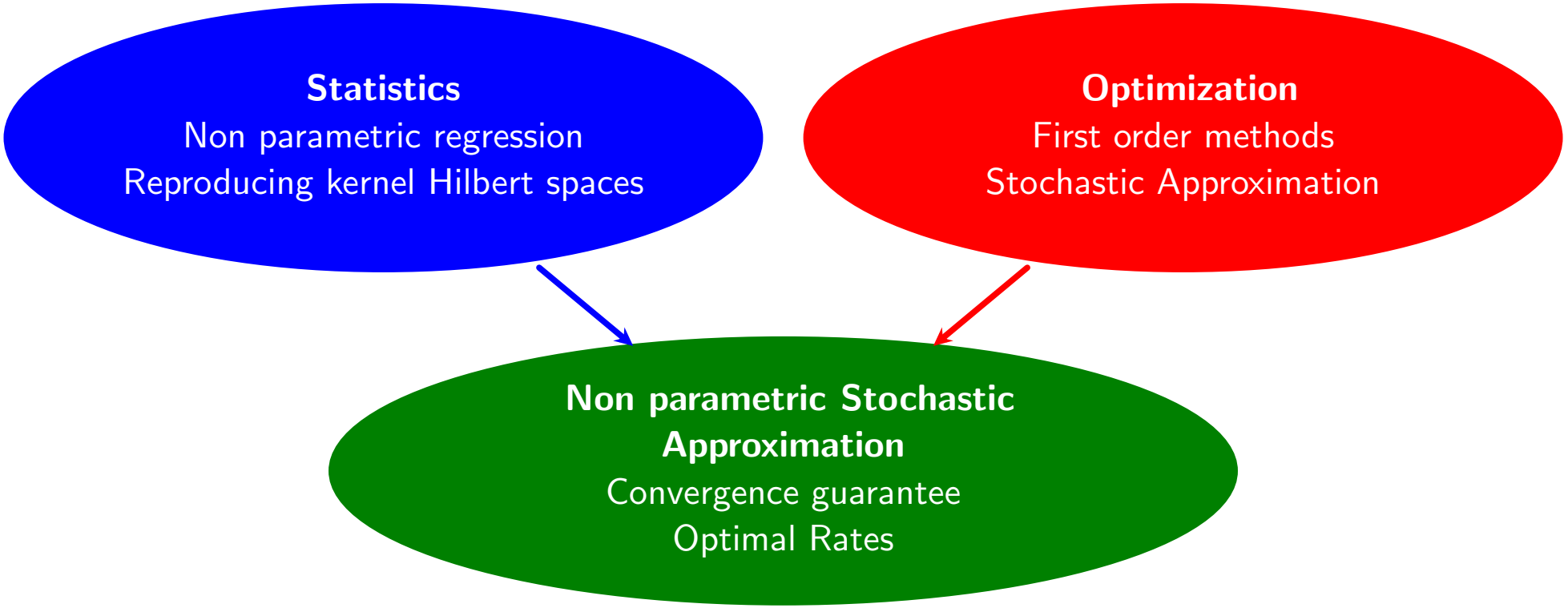
Markov chain interpretation of constant step sizes

- LMS recursion for $R_n(\theta) = \frac{1}{2}\big(Y_n - \langle K_{X_n}, \theta\rangle\big)^2$

$$\theta_n = \theta_{n-1} - \gamma\big(\langle K_{X_n}, \theta_{n-1}\rangle - Y_n\big)K_{X_n}$$

- The sequence $(\theta_n)_n$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$
- **For least-squares, $\bar{\theta}_\gamma = \theta_*$**
  - θ_n does not converge to θ_* but oscillates around it
  - oscillations of order $\sqrt{\gamma}$
- **Ergodic theorem:**
  - Averaged iterates converge to $\bar{\theta}_\gamma = \theta_*$ at rate $O(1/n)$

Outline


Statistics
Non parametric regression
Reproducing kernel Hilbert spaces
Optimization
First order methods
Stochastic Approximation
Non parametric Stochastic Approximation
Convergence guarantee
Optimal Rates

Stochastic Approximation in the RKHS

$$\begin{aligned} \theta_0 &\in \mathcal{H}_K \quad \text{(we often consider } \theta_0 = 0\text{)}, \\ \theta_n &= \sum_{i=1}^{n} a_i K_{X_i}, \end{aligned}$$

$(a_n)_n$ such that $a_n = -\gamma_n(\theta_{n-1}(X_n) - Y_n) = -\gamma_n \left(\sum_{i=1}^{n-1} a_i K(X_n, X_i) - Y_i \right)$.

$$\begin{aligned} \theta_n &= \theta_{n-1} - \gamma_n \left(\theta_{n-1}(X_n) - Y_n\right) K_{X_n} \\ &= \sum_{i=1}^{n} a_i K_{X_i} \quad \text{with } a_n \text{ defined as above.} \end{aligned}$$

$(\theta_{n-1}(X_n) - Y_n)K_{X_n}$ unbiased estimate of $\nabla R(\theta_{n-1})$.

SGD algorithm in the RKHS takes very simple form.

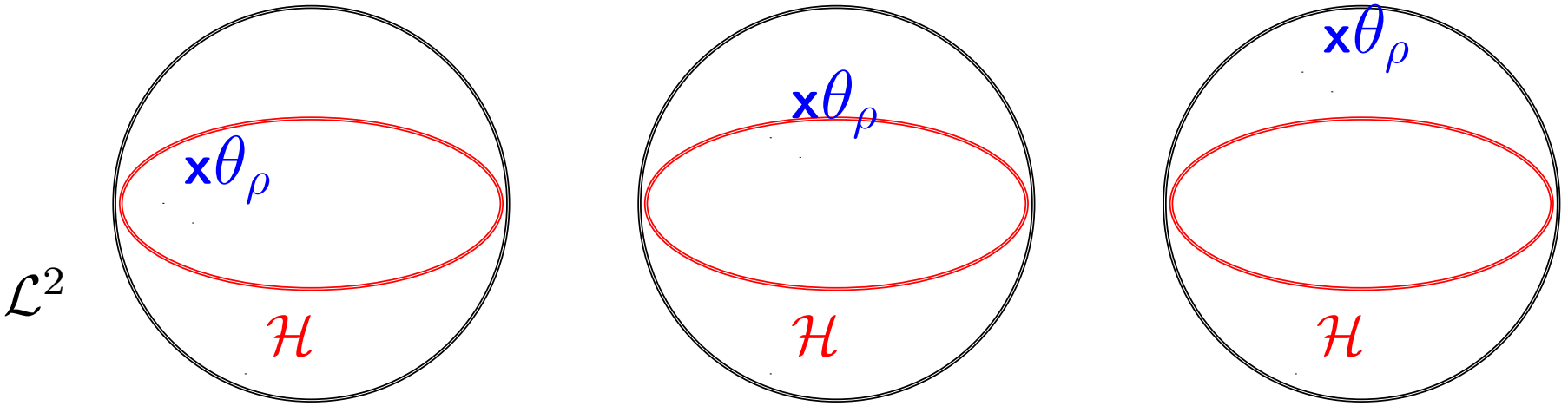
Assumptions 1/2

We define $\Sigma := \mathbb{E}\left[K_x \otimes K_x\right]$. Where $K_x \otimes K_x : g \mapsto \langle K_x, g \rangle K_x$.

Two important points characterize the difficulty of the problem:

1 The regularity of the objective function θ_ρ.

$$\theta_\rho \in \Sigma^r(\mathcal{L}^2), \text{ with } r \geq 0$$



$r > 0.5$ $\quad$ $r < 0.5$ $\quad$ $r \ll 0.5$

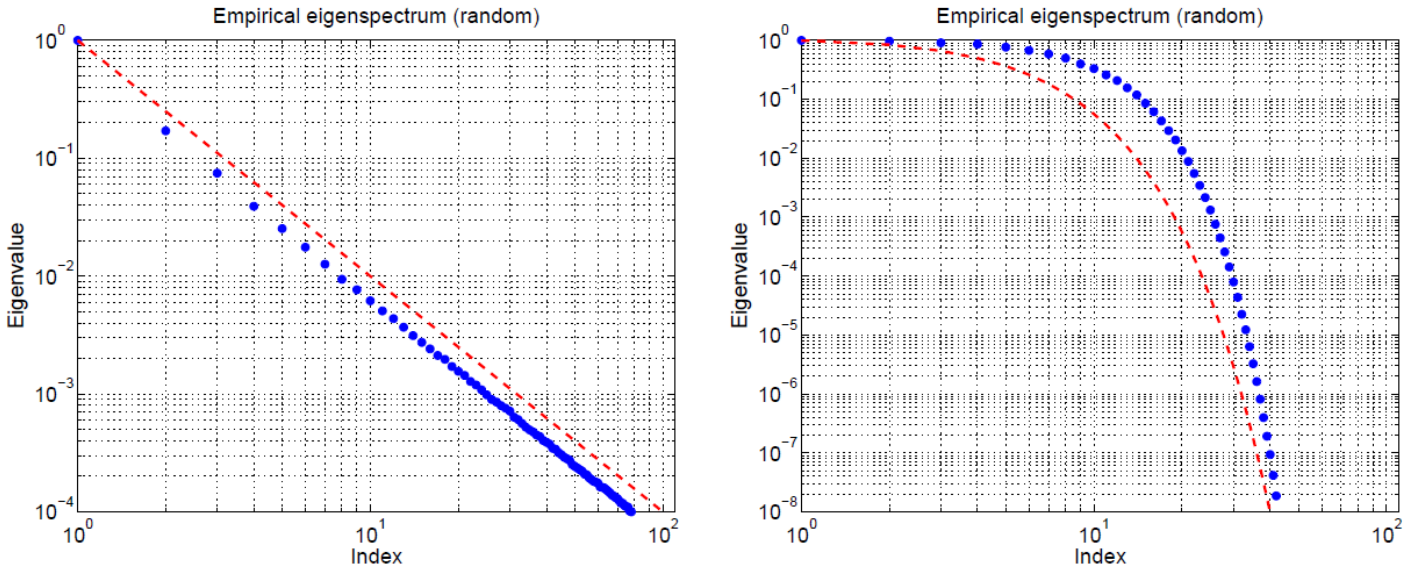
Assumptions 2/2

We define $\Sigma := \mathbb{E}\left[K_x \otimes K_x\right]$. Where $K_x \otimes K_x : g \mapsto \langle K_x, g \rangle K_x$.

Two important points characterize the difficulty of the problem:

2 The spectrum of the covariance operator:

$$\mathrm{tr}(\Sigma^\alpha) < \infty, \text{ for } \alpha \in [0; 1].$$



Main Result

Theorem 1. [Dieuleveut and Bach (2015)] *Under both assumptions, for a small enough constant step size γ, with averaging:*

$$\mathbb{E}\left[R\left(\bar{\theta}_n\right) - R(\theta_\rho)\right] \leq O\left(\frac{\sigma^2 \operatorname{tr}(\Sigma^\alpha)\gamma^\alpha}{n^{1-\alpha}}\right) + O\left(\frac{||\Sigma^{-r}(\theta_\rho - \theta_0)||^2}{(n\gamma)^{2r}}\right)$$

- Bias Variance decomposition
- Finite horizon result here but extends to online setting.

Main Result

Theorem 2. [Dieuleveut and Bach (2015)] *Under both assumptions, for a small enough constant step size γ, with averaging:*

$$\mathbb{E}\left[R\left(\bar{\theta}_n\right) - R(\theta_\rho)\right] \leq O\left(\frac{\sigma^2 \operatorname{tr}(\Sigma^\alpha)\gamma^\alpha}{n^{1-\alpha}}\right) + O\left(\frac{||\Sigma^{-r}(\theta_\rho - \theta_0)||^2}{(n\gamma)^{2r}}\right)$$

- Bias Variance decomposition
- Finite horizon result here but extends to online setting.

Corollary 1. *In finite dimension, $\alpha = 0$, $r = 1/2$:*

$$\mathbb{E}\left[R\left(\bar{\theta}_n\right) - R(\theta_\rho)\right] \leq O\left(\frac{\sigma^2 d}{n}\right) + O\left(\frac{||\theta_\rho - \theta_0||^2_{\mathbb{R}^d}}{n\gamma}\right)$$

I.e., Statistical rate + Rate of GD.

Corollary

Corollary 2. *If* $\frac{1-\alpha}{2} < r < \frac{2-\alpha}{2}$, *with* $\gamma = n^{-\frac{2r+\alpha-1}{2r+\alpha}}$ *we get the optimal rate:*

$$\mathbb{E}\left[R\left(\bar{g}_n\right) - R(g_\rho)\right] = O\left(n^{-\frac{2r}{2r+\alpha}}\right)$$

- **We get statistical optimal rate (Caponnetto and De Vito (2007)) of convergence for learning in RKHS with SA, with one pass.**

- We compare favorably to Ying and Pontil (2008); Tarrès and Yao (2011).

Current and future work

- Acceleration (Dieuleveut et al. (2016))
- Random features and Sketching (complexity $O(nd_n)$)
- Other Losses

References

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Adv. NIPS*, 2013.

A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Dieuleveut and F. Bach. Non-parametric Stochastic Approximation with Large Step sizes. *Annals of Statistics*, 2015.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. Technical Report 1602.05419, arXiv, 2016.

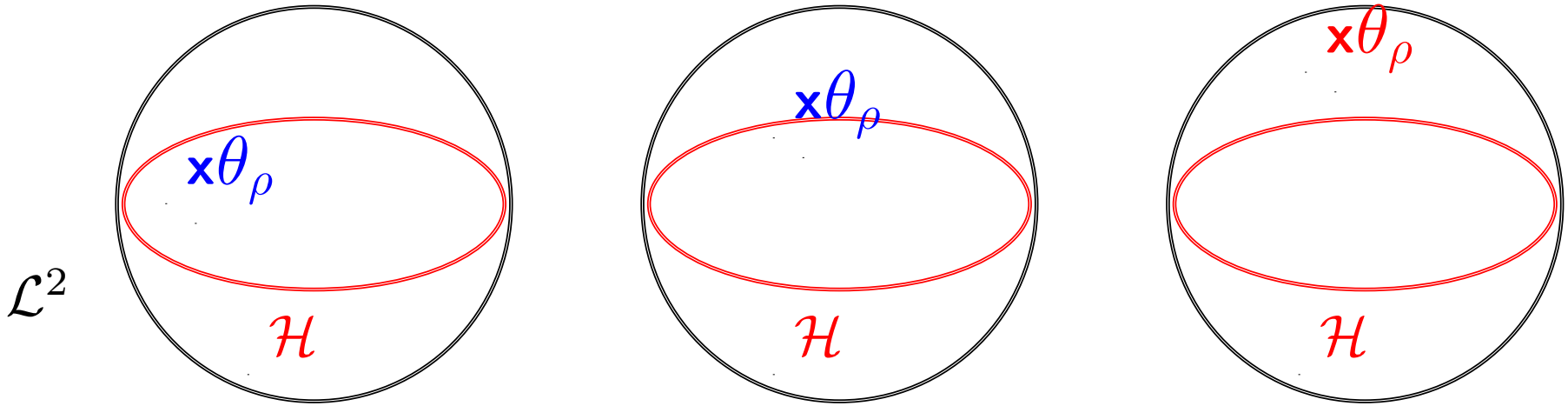
Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer, 2004.

P. Tarrès and Y. Yao. Online learning as stochastic approximation of regularization paths. ArXiv e-prints 1103.5538, 2011.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. ISBN 9780387790527.

Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 5, 2008.

Harder, Faster, Better, Stronger, Convergence Rates for Least Squares Regression



Averaged and accelerated (Nesterov (2004)) stochastic algorithm, to improve speed at which we forget initial conditions.