Non-asymptotic bound for stochastic averaging

and some other related stuff

S. Gadat and F. Panloup

Toulouse School of Economics

CMAP Seminar, April 2018

I - Introduction
- I - 1 Motivations
- I - 2 Optimization
- I - 3 Stochastic Optimization
- I - 4 No novelty in this talk, as usual !

II Well known algorithms
- II - 1 Stochastic Gradient Descent
- II - 2 Heavy Ball with Friction
- II - 3 Polyak-Ruppert Averaging
- II - 4 In this talk

III Stochastic Heavy Ball with Friction
- III - 1 Almost sure convergence
- III - 2 Rates of convergence

IV Polyak averaging
- IV - 1 Almost sure convergence
- IV - 2 Beyond strong convexity ?
- IV - 3 Averaging analysis
- IV - 4 Linearisation and moments
- IV - 5 Averaging - Main result

I - 1 Optimization - Motivations : Statistical problems

- Objective : minimize a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}_+$

$$f(\theta) := \mathbb{E}[f(\theta, X)] = \int_{\mathcal{X}} f(\theta, x) d\mathbb{Q}(x)$$

- Motivation : minimization originates from a statistical estimation problem
- M-estimation point of view : observations $X_1, \ldots, X_N$ and

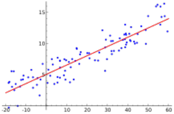
$$\hat{\theta}_N := \arg\min f_N(\theta)$$

  f_N may be seen as a stochastic approximation of a hidden f.
- Among others, classical statistical problems :
  - Supervised regression : Linear Models
  - Supervised classification : Logistic regression
  - Other problems : Quantile estimation
  - ...
- Important way of thinking :
  - the situations we are expecting to deal with are on-line
  - Why ? May be the core of the application (!)
  - Why ? Too much observations to handle all of them in a single pass

I - 1 Optimization - Motivations : Supervised regression

Assume $(X_i, Y_i)_{1 \leqslant i \leqslant N}$ comes from the statistical model

$$\forall i \in \{1 \dots N\} \qquad Y_i = \langle X_i, \theta^\star \rangle + \epsilon_i.$$



You observe $(X_i, Y_i)_{1 \leqslant i \leqslant N}$. $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$. $\theta^\star$ is unknown.

You assume that $(\epsilon_i)_{1 \leqslant i \leqslant N}$ are centered and i.i.d.

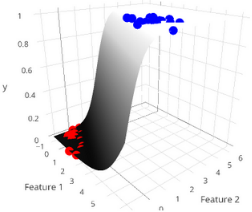
- Gaussian settings : if $(\epsilon_i)_{1 \leqslant i \leqslant N}$ are $\mathcal{N}(0,1)$, the log-likelihood leads to the minimization of the sum of squares :

$$f_N(\theta) = \sum_{i=1}^{N} \|Y_i - \langle X_i, \theta \rangle\|^2.$$

- You can choose to minimize f_N regardless any assumption on ϵ.

Important point :

$$\mathbb{E}\left[\frac{f_N(\theta)}{N}\right] = \underbrace{\mathbb{E}_{X,Y}[\|Y - \langle X, \theta \rangle\|^2]}_{:=f(\theta)} \qquad \text{and} \qquad \theta^\star = \arg\min f.$$



Assume $(X_i, Y_i)_{1 \leqslant i \leqslant N}$ comes from the statistical model :

- X_i are i.i.d. whose distribution is $\mathbb{Q}$ over $\mathbb{R}^p$ (p=2 on the left)
- $Y_i \in \{-1, +1\}$ and

$$\mathbb{P}[Y_i = +1 \,|X = x] = \frac{1}{1 + e^{-<x,\theta^\star>}}.$$

You observe $(X_i, Y_i)_{1 \leqslant i \leqslant N}$. $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$. $\theta^\star$ is unknown.
Write the log-likelihood to estimate $\theta^\star$:

$$f_N(\theta) = \sum_{i=1}^{N} \log\left(1 + e^{-Y_i <X_i,\theta>}\right)$$

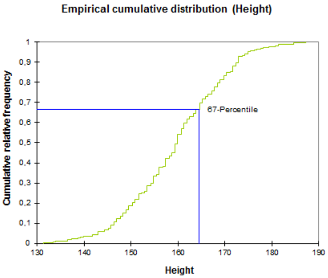
Important point :

$$\mathbb{E}\left[\frac{f_N(\theta)}{N}\right] = \underbrace{\mathbb{E}_{X,Y} \log\left(1 + e^{-Y<X,\theta>}\right)}_{:=f(\theta)} \qquad \text{and} \qquad \theta^\star = \arg\min f.$$

I - 1 Optimization - Motivations : Recursive quantile estimation

We observe $(X_i)_{1 \leqslant i \leqslant N}$ distributed according to $\mathbb{Q}$ over $\mathbb{R}$.

Assume that $\mathbb{Q}$ has a density q w.r.t. λ (not necessarily compactly supported and lower bounded on this compact set).



Given any $\alpha > 0$, find q_α such that

$$\int_{-\infty}^{q_\alpha} p = 1 - \alpha.$$

Find the minimum of f such that $f'(\theta) = \int_{q_\alpha}^{\theta} p$:

$$f(\theta) := \int_{q_\alpha}^{\theta} \left[\int_{q_\alpha}^{u} p(s) ds \right] du$$

I - 1 Optimization - Motivations : large scale estimation problems ?

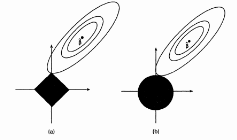
$$f(\theta) := \mathbb{E}[f(\theta, X)] = \int_{\mathcal{X}} f(\theta, x) d\mathbb{Q}(x)$$

- A lot of observations that may be observed recursively : large N

  Goal : manageable from a computational point of view.

- We handle in this talk only smooth problems :

  f is assumed to be differentiable $\Longrightarrow$ no composite problems



- Noisy/stochastic minimization :
  - the N observations are i.i.d. and are gathered in a channel of information
  - they feed the computation of the target function f_N, that mimics f
  - Idea : use at each iteration **only one arrival** in the channel

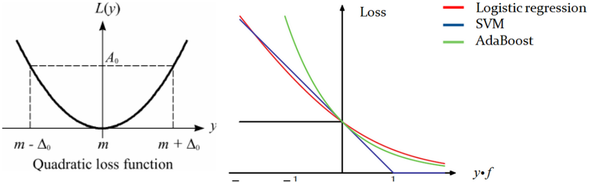
$$f_N(\theta) = f_{N-1}(\theta) + \ell_{(X_N, Y_N)}(\theta) \Longrightarrow \theta_N = \theta_{N-1} - \gamma_N g(\theta_{N-1}, X_N, Y_N)$$

I - 2 Optimization - convexity

- Smooth minimization $\mathcal{C}^2$ problem

$$\arg\min_{\mathbb{R}^d} f.$$

Generally, f is also assumed to be convex/strongly convex
Quadratic loss/Logistic loss :



- **First order deterministic methods** (with t evaluations of ∇f) :
  - when f is assumed to be convex, polynomial rates (NAGD) :

$$O(1/t^2)$$

  - when f is strongly convex, linear rates (NAGD) :

$$O(e^{-\rho t})$$

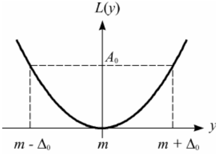
- Last observation : minimax paradigm. Worst case in the class of functions with a fixed horizon t

I - 3 Stochastic Optimization - convexity

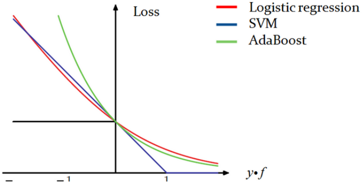
- Smooth minimization $\mathcal{C}^2$ problem

$$\arg\min_{\mathbb{R}^d} f.$$

Generally, f is also assumed to be convex/strongly convex
Quadratic loss/Logistic loss :



Quadratic loss function



- **First order stochastic methods** (with $\nabla f + \xi$ with $\mathbb{E}[\xi] = 0$) (NY83) :
  - when f is assumed to be convex :

$$O(1/\sqrt{t})$$

  - when f is strongly convex :

$$O(1/t)$$

- Last observation : minimax paradigm. Worst case in the class of functions with a fixed horizon t

I - 3 Stochastic Optimization - convexity

Smooth minimization $\mathcal{C}^2$ problem

$$\theta^\star := \arg\min_{\mathbb{R}^d} f.$$

Build a recursive optimization method $(\theta_n)_{n\geqslant 1}$ with noisy gradients and ...

Current hot questions ?

- Beyond convexity/strong convexity ?
  Example : recursive quantile estimation problem.
  Use of KL functional inequality ? Multiple wells situations ?
- Adaptivity of the method ?
  Methods *independent on/robust to* some unknown quantities :
  Hessian at the target point.
- Non asymptotic bound ? Exact/sharp constant ?

$$\forall n \geqslant \mathbb{N} \qquad \mathbb{E}\|\theta_n - \theta^\star\|^2 \leqslant \frac{Tr(V)}{n} + A/n^{1+\epsilon},$$

  $Tr(V)$: asymptotic incompressible variance (Cramer-Rao lower bound.)
- Large deviations ?

$$\forall n \geqslant \mathbb{N} \quad \forall t \geqslant 0 \qquad \mathbb{P}\left(\|\theta_n - \theta^\star\| \geqslant b(n) + t\right) \leqslant e^{-R(t,n)}$$

- $\mathbb{L}^p$ loss ?

$$\mathbb{E}\|\theta_n - \theta^\star\|^{2p} \leqslant \frac{A_p}{n^p} + B_p/n^{p+\epsilon}$$

I - 4 No novelty in this talk, as usual !

We will consider some well known methods in this talk (! !)

- Stochastic Gradient Descent (SGD)
- Heavy Ball with Friction (HBF)
- Polyak Averaging $((\overline{\theta}_n)_{n\geqslant 1})$

I - Introduction
- I - 1 Motivations
- I - 2 Optimization
- I - 3 Stochastic Optimization
- I - 4 No novelty in this talk, as usual !

II Well known algorithms
- II - 1 Stochastic Gradient Descent
- II - 2 Heavy Ball with Friction
- II - 3 Polyak-Ruppert Averaging
- II - 4 In this talk

III Stochastic Heavy Ball with Friction
- III - 1 Almost sure convergence
- III - 2 Rates of convergence

IV Polyak averaging
- IV - 1 Almost sure convergence
- IV - 2 Beyond strong convexity ?
- IV - 3 Averaging analysis
- IV - 4 Linearisation and moments
- IV - 5 Averaging - Main result

II - 1 Stochastic Gradient Descent (SGD) - Robbins-Monro 1951.

$$f(\theta) = \mathbb{E}_{X\sim\mathbb{Q}}[f(\theta, X)] \qquad X_1, \ldots, X_n \qquad i.i.d. \qquad \mathbb{Q}.$$

- Idea : use the steepest descent with one observation each time.
- Homogeneization all along the iterations
- Build the sequence $(\theta_n)_{n\geqslant 1}$ as follows :
  - $\theta_0 \in \mathbb{R}^d$
  - Iterate $\theta_{n+1} = \theta_n - \gamma_{n+1} g_n(\theta_n)$ with

    $$g_n(\theta_n) = \nabla_\theta f(\theta_n, X_n) = \nabla f(\theta_n) + \xi_{n+1},$$

    where $(\xi_n)_{n\geqslant 1}$ is a sequence of Martingale increments :

    $$\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] = 0,$$

    where $\mathcal{F}_n = \sigma(\theta_0, \ldots, \theta_n)$.
- Typical state of the art result

Theorem

Assume f is strongly convex $SC(\alpha)$:

- *If $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (0,1)$ then $\mathbb{E}[\|\theta_n - \theta^\star\|^2] \leqslant C_\alpha \gamma_n$*
- *If $\gamma_n = \gamma n^{-1}$ with $\gamma\alpha > 1/2$, then $\mathbb{E}[\|\theta_n - \theta^\star\|^2] \leqslant C_\alpha n^{-1}$*

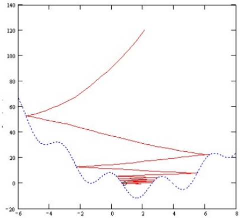
Pros : easy analysis, avoid traps (Pemantle 1990, Brandiere-Duflo 1996)
Cons : Not adaptive, no sharp inequality, no KL settings, . . .

II - 2 Heavy Ball with Friction

- Produce a second order discrete recursion from the HBF ODE of Polyak (1987) and Antipin (1994) :

$$\ddot{x}_t + a_t \dot{x}_t + \nabla f(x_t) = 0 \qquad a_t = \frac{2\alpha+1}{t} \quad \text{or} \quad a_t = a > 0$$

- Mimic the displacement of a ball rolling on the graph of the function f.



- Up to a time scaling modification, equivalent system to the NAGD (CEG09, SBC12, AD17) that may be rewritten as

$$X'_t = -Y_t \quad \text{and} \quad Y'_t = r(t)(\nabla f(X_t) - Y_t)dt \quad \text{with} \quad r(t) = \frac{\alpha+1}{t} \quad \text{or} \quad r(t) = r > 0.$$

- Stochastic version, two sequences :

$$X_{n+1} = X_n - \gamma_{n+1} Y_n \qquad \text{and} \qquad Y_{n+1} = Y_n + r_n \gamma_{n+1}(g_n(X_n) - Y_n)$$

II - 3 Polyak-Ruppert Averaging

- Start from a SGD sequence $(\theta_n)_{n\geqslant 1}$

$$\theta_{n+1} = \theta_n - \gamma_{n+1} g_n(\theta_n) \qquad \text{with} \qquad \gamma_n = \gamma n^{-\beta}, \beta \in (0,1).$$

- Idea : Cesaro averaging all along the sequence
- Build the mean sequence over the past iterations :

$$\overline{\theta}_n = \frac{1}{n} \sum_{j=1}^{n} \theta_j$$

- Typical state of the art result

Theorem (PJ92)

If f is strongly convex $SC(\alpha)$ and $C^1_L(\mathbb{R}^d)$ and $\beta \in (1/2, 1)$:

$$\sqrt{n}(\overline{\theta}_n - \theta^\star) \longrightarrow N(0, V) \qquad \text{as} \qquad n \longrightarrow +\infty.$$

V possesses an optimal trace and $(\overline{\theta}_n)_{n\geqslant 1}$ attains the Cramer-Rao lower bound asymptotically.

Theorem (BM11,B14,G16)

For several particular cases of convex minimization problems (logistic, least squares, quantile with "convexity") :

$$\mathbb{E}\|\overline{\theta}_n - \theta^\star\|^2 \leqslant \frac{C}{n}$$

II - 4 In this talk

We propose two contributions on the previous second order methods :

- For the stochastic HBF (joint work with S. Saadane and F. Panloup) :
  - Almost sure consistency, multiple wells study
  - $\mathbb{L}^2$ rates of convergence (not optimal)
  - Spectral explanation of "why not adaptive ?"
- For the Polyak-Ruppert averaging algorithm (joint work with F. Panloup) :
  - Relax the convexity assumption (KL inequality instead), very mild assumption on the data
    Works for any convex semi-algebric function, recursive quantile, logistic regression, strongly convex functions, . . .
  - Incidentally easy $\mathbb{L}^p$ consistency rate of SGD (!)
  - **Sharp non asymptotic minimax $\mathbb{L}^2$ rate for $\overline{\theta}_n$**
  - Spectral explanation of "why it works ?"

I - Introduction
- I - 1 Motivations
- I - 2 Optimization
- I - 3 Stochastic Optimization
- I - 4 No novelty in this talk, as usual !

II Well known algorithms
- II - 1 Stochastic Gradient Descent
- II - 2 Heavy Ball with Friction
- II - 3 Polyak-Ruppert Averaging
- II - 4 In this talk

III Stochastic Heavy Ball with Friction
- III - 1 Almost sure convergence
- III - 2 Rates of convergence

IV Polyak averaging
- IV - 1 Almost sure convergence
- IV - 2 Beyond strong convexity ?
- IV - 3 Averaging analysis
- IV - 4 Linearisation and moments
- IV - 5 Averaging - Main result

III - 1 Almost sure convergence

General function f

- Recursive scheme :

$$X_{n+1} = X_n - \gamma_{n+1} Y_n \qquad \text{and} \qquad Y_{n+1} = Y_n + r_n \gamma_{n+1} (g_n(X_n) - Y_n)$$

- Find a mean-reverting effect on the random dynamical system.
- Use former works on dissipative systems (H91, DV01, CEG09, ...) : construct a Lyapunov function as

$$V_n(x,y) = a_n f(x) + b_n \|y\|^2 - c\langle \nabla f(x), y\rangle$$

and prove that

$$\begin{aligned} \mathbb{E}\left[V_n(X_{n+1}, Y_{n+1}) \,|\, \mathcal{F}_n\right] \quad \leqslant \quad & (1 + C\gamma_{n+1}^2 r_n) V_n(X_n, Y_n) \\ & - c_1 \gamma_{n+1} \|Y_n\|^2 - c_2 \gamma_{n+1} r_n \|\nabla f(X_n)\|^2 + O(\gamma_{n+1}^2 r_n) \end{aligned}$$

Deduce that $\sum \left\{\gamma_{n+1} \|Y_n\|^2 + \gamma_{n+1} r_n \|\nabla f(X_n)\|^2\right\} < +\infty$ a.s.

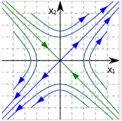
Theorem

If f is coercive with bounded hessian, if $\sup_{n\geqslant 1} \mathbb{E}\|\xi_n\|^2 < \infty$, and if the set of critical points is discrete, then X_n a.s. converges towards a critical point of f.

III - 1 Almost sure convergence Minimum/maximum

If f has several wells

- Well known fact : S.A. avoids local traps (result for SGD)
- Does-it hold for stochastic HBF ?
- Major difficulty : the martingale noise only acts on the Y coordinate
- Key result : Poincaré Lemma around hyperbolic equilibria.



Local maxima can be shown to be repulsive for the deterministic vector field. Then use/modify an argument of Pemantle to show that

Theorem

If the noise is elliptic (non negative variance in any direction of $\mathbb{R}^d$) and sub-Gaussian, then a.s. convergence towards a local minimum of f.

III - 2 Rates of convergence

If f is strongly convex with a unique minimizer $\theta^\star$

- Idea : study first the quadratic case for f (linear drift situation)
- Use a linearization argument to handle general functions f

$$\begin{cases} X_{n+1} = X_n - \gamma_{n+1} Y_n \\ Y_{n+1} = Y_n + \gamma_{n+1} r_n (S X_n - Y_n) + \gamma_{n+1} r_n \xi_{n+1}, \end{cases}$$

- Up to a change of basis (suitable for S), manage d $\{2 \times 2\}$ systems

$$Z_{n+1}^{(i)} = \left(I_2 + \gamma_{n+1} \begin{pmatrix} 0 & -1 \\ \lambda^{(i)} r_n & -r_n \end{pmatrix} \right) Z_n^{(i)} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \xi_n^{(i)}$$

- Characteristic polynomial :

$$\chi_{C_n}(t) = \left(t + \frac{r_n}{2} \right)^2 + \frac{r_n (4\lambda - r_n)}{4}.$$

III - 2 Rates of convergence - linear case

Theorem

If $Sp(S) \subset [\lambda, +\infty[$ and $r_n = r$. Assume that $\gamma_n = \gamma n^{-\beta}$. Set :

$$\alpha_r = \begin{cases} r\left(1 - \sqrt{1 - \frac{4\lambda}{r}}\right), & \text{if} \quad r \geqslant 4\lambda \\ r & \text{if} \quad r < 4\lambda, \end{cases} .$$

(i) *If $\beta < 1$, then a constant $c_{r,\lambda,\gamma}$ exists such that :*

$$\forall n \geqslant 1 \qquad \mathbb{E}\left[\|X_n\|^2 + \|Y_n\|^2\right] \leqslant c_{r,\lambda,\gamma}\gamma_n.$$

(ii) *If $\beta = 1$, then a constant $c_{r,\lambda,\gamma}$ exists such that :*

$$\forall n \geqslant 1 \qquad \mathbb{E}\left[\|X_n\|^2 + \|Y_n\|^2\right] \leqslant c_{r,\lambda,\gamma} n^{-(1 \wedge \gamma\alpha_r)} \log(n)^{\mathbf{1}_{\{\gamma\alpha_r = 1\}}}.$$

- Optimal rate n^{-1} possible when $\gamma\alpha_r > 1$
- $\max_r \alpha_r = 4\lambda > 2\lambda$
- When $r \longrightarrow +\infty$, $\alpha_r \longrightarrow 2\lambda$ (identical to a standard SGD)
- No adaptive procedure (optimality depends on λ), confirmed by a CLT
- can be generalized to strongly convex functions...

I - Introduction
I - 1 Motivations
I - 2 Optimization
I - 3 Stochastic Optimization
I - 4 No novelty in this talk, as usual !

II Well known algorithms
II - 1 Stochastic Gradient Descent
II - 2 Heavy Ball with Friction
II - 3 Polyak-Ruppert Averaging
II - 4 In this talk

III Stochastic Heavy Ball with Friction
III - 1 Almost sure convergence
III - 2 Rates of convergence

IV Polyak averaging
IV - 1 Almost sure convergence
IV - 2 Beyond strong convexity ?
IV - 3 Averaging analysis
IV - 4 Linearisation and moments
IV - 5 Averaging - Main result

IV - 1 Almost sure convergence

- Use a SGD sequence $(\theta_n)_{n\geqslant 1}$ with step size $(\gamma_n)_{n\geqslant 1}$.
- Averaging

$$\overline{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_k, \quad n \geqslant 1$$

- $$\overline{\theta}_{n+1} = \overline{\theta}_n\left(1 - \frac{1}{n+1}\right) + \frac{1}{n+1}(\theta_n - \gamma_{n+1}g_n(\theta_n)).$$

Free result :
If unique minimizer of f (what is assumed below from now on), the a.s. convergence of $(\overline{\theta}_n)_{n\geqslant 1}$ comes from the one of $(\theta_n)_{n\geqslant 1}$.
Goals :

- Optimality
- Non asymptotic behaviour
- Adaptivity
- Weaken the convexity assumption

For deterministic problems : behaviour of f around $\theta^\star$ is important

For stochastic problems : behaviour of f around $\theta^\star$ and near ∞ are important

IV - 2 Beyond strong convexity ?

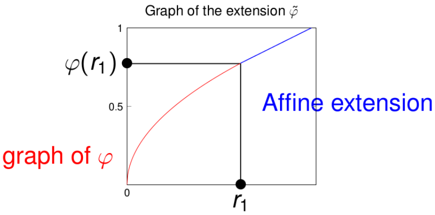
Definition (Kurdyka-Lojasiewicz type inequality $\mathbf{H}^{\mathbf{r}}_{\mathbf{kl}}$)

$f(\theta^\star) = 0$ is the unique (local/global) minimizer of f, $D^2 f(\theta^\star)$ invertible and

$$\exists\, r \in [0, 1/2] \qquad \liminf_{|x| \longrightarrow +\infty} f^{-r} |\nabla f| > 0 \qquad \text{and} \qquad \limsup_{|x| \longrightarrow +\infty} f^{-r} |\nabla f| > 0$$

Implicitly :

- Unique critical point
- Typically sub-quadratic situation (C^1_L)
- Desingularizes the function f near $\theta^\star$
- f **does not need to be convex**



Proposition

If $\mathbf{H}^{\mathbf{r}}_{\mathbf{kl}}$ holds for $r \in [0, 1/2]$, then define $\varphi(x) = (1 + |x|^2)^{\frac{1}{2} - r}$ and

$$\exists\, 0 < m < M \quad \forall x \in \mathbb{R}^d \backslash \{\theta^\star\} : \qquad m \leqslant \varphi'(f(x)) |\nabla f(x)|^2 + \frac{|\nabla f(x)|^2}{f(x)} \leqslant M.$$

Moreover, $\liminf_{|x| \to +\infty} f(x) |x|^{-\frac{1}{1-r}} > 0.$

IV - 2 Beyond Strong convexity ?

Few references :

- Seminal contributions of Kurdyka (1998) & Łojasiewicz (1958),
- Error bounds in many situations (see Bolte *et al.* linear convergence rate of the FoBa proximal splitting for the lasso)
- Many many functions satisfy KL : convex, coercive, semi-algebraic

For us, it makes it possible to handle :

- Recursive least squares problems $r = 1/2$
- Online logistic regression $r = 0$
- Recursive quantile problem $r = 0$

Last assumption (restrictive for the sake of readability)

Assumption (Martingale noise)

$$\sup_{n \geqslant 1} \|\xi_{n+1}\| < +\infty$$

Can be largely weakened with additional technicalities

IV - 3 Averaging analysis ($\theta^\star = 0$)

$$\overline{\theta}_{n+1} = \overline{\theta}_n \left(1 - \frac{1}{n+1}\right) + \frac{1}{n+1}(\theta_n - \gamma_{n+1} g_n(\theta_n)).$$

Linearisation : Introduce $Z_n = (\theta_n, \overline{\theta}_n)$ and

$$Z_{n+1} = \begin{pmatrix} I_d - \gamma_{n+1}\Lambda_n & 0 \\ \frac{1}{n+1}(I_d - \gamma_{n+1}\Lambda_n) & (1 - \frac{1}{n+1})I_d \end{pmatrix} Z_n + \gamma_{n+1} \begin{pmatrix} \xi_{n+1} \\ \frac{\xi_{n+1}}{n+1} \end{pmatrix},$$

where $\Lambda_n = \int_0^1 D^2 f(t\theta_n)dt$: $\Lambda_n Z_n = \nabla f(Z_n)$. Replace formally Λ_n by $D^2 f(\theta^\star)$
Key matrix : for any $\mu > 0$ and any integer n :

$$E_{\mu,n} := \begin{pmatrix} 1 - \gamma_{n+1}\mu & 0 \\ \frac{1-\mu\gamma_{n+1}}{n+1} & 1 - \frac{1}{n+1} \end{pmatrix}.$$

Obvious eigenvalues and ... $(0, \overline{\theta}_n)$ is living on the "good" eigenvector ;)

- Conclusion 1 : Expect a behaviour of $(\overline{\theta}_n)_{n \geqslant 1}$ independent from $D^2 f(\theta^\star)$
- Conclusion 2 : Expect a rate of n^{-1}

Difficulties :
$E_{\mu,n}$ is not symmetric $\Longrightarrow$ non orthonormal eigenvectors
$E_{\mu,n}$ varies with n
Requires a careful understanding of the eigenvectors variations

IV - 3 Averaging analysis : linear case

Linear case :

How to produce a sharp upper bound ? Derive an inequality of the form

$$\mathbb{E}[\|\widetilde{Z}_{n+1}\|^2 \mid \mathcal{F}_n] \leqslant \left(1 - \frac{1}{n+1} + \delta_{n,\beta}\right)^2 \|\widetilde{Z}_n\|^2 + \frac{Tr(V)}{(n+1)^2},$$

where

$$V = D^2 f(\theta^\star)^{-1} \Sigma^\star D^2 f(\theta^\star)^{-1}.$$

$\delta_{n,\beta}$ is an error term : variation of the eigenvectors from n to $n+1$.
If $\delta_{n,\beta}$ is shown to be small enough, then we obtain

$$\mathbb{E}[\|\widetilde{Z}_n\|^2] \leqslant \frac{Tr(V)}{n} + \underbrace{\epsilon_{n,\beta}}_{:=O(n^{-(1+\upsilon_\beta)})}$$

Linearisation :

We need to replace Λ_n by $D^2 f(\theta^\star)$ and we are done !

IV - 4 Averaging analysis : cost of the linearisation

- We need to replace Λ_n by $D^2 f(\theta^\star)$
- Needs some preliminary controls on the SGD $(\theta_n)_{n\geqslant 1}$ (moments)
- Known state of the art results when f SC or in particular situations

Theorem

For $\beta \in [0,1]$, under $\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}}$, a collection of constants $C_{p,r}$ exists such that

$$\mathbb{E}\left[\|\theta_n - \theta^\star\|^{2p}\right] \leqslant C_{p,r}\gamma_n^p$$

Key argument : define a Lyapunov function :

$$V_p(\theta) = f(\theta)^p e^{\varphi(f(\theta))}$$

and prove a mean reverting effect property (without any recursion on p) :

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[V_p(\theta_{n+1}) \,|\, \mathcal{F}_n\right] \leqslant \left(1 - \frac{\alpha}{2}\gamma_{n+1} + c_1\gamma_{n+1}^2\right) V_p(\theta_n) + c_2\{\gamma_{n+1}\}^{p+1}.$$

Remarks :
Important role of φ !
Painful second order Taylor expansion . . .

IV - 5 Averaging - Main result

We can state our main result with $\beta \in (1/2, 1), \gamma_n = \gamma_1 n^{-\beta}$:

Theorem

Under $\mathbf{H}_{\mathbf{KL}}^{\mathbf{r}}$, a constant C_r exists such that

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[\|\overline{\theta}_n - \theta^\star\|^2\right] \leqslant \frac{Tr(V)}{n} + C_r n^{-\{(\beta+1/2) \wedge (2-\beta)\}}.$$

The "optimal" choice $\beta = 3/4$ satisfies the upper bound :

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[\|\overline{\theta}_n - \theta^\star\|^2\right] \leqslant \frac{Tr(V)}{n} + C_r n^{-5/4}.$$

- Non asymptotic optimal variance term (Cramer-Rao lower bound)
- Adaptive to the unknown value of the Hessian
- Only requires invertibility of $D^2 f(\theta^\star)$
- No strong convexity
- $\beta = 3/4$ no real understanding on this optimality (just computations)
- Second order term seems to be of the good size (with simulations)
- State of the art : second order term only explicit in [BM11], of size $n^{-7/6}$

IV - 5 Averaging - Main result

	Setting	Cramer-Rao	2^{nd} order v_n
Our work	Strong. Convex Convex (Smooth KL) Logist. Reg. (KL) Recurs. Quantile (KL)	Yes : $\frac{Tr(V)}{n}$	$n^{-(\beta+\frac{1}{2})\wedge(2-\beta)}$, $v_n^\star = O(n^{-\frac{5}{4}})$
BM(11)	Strong. Convex	Yes : $\frac{Tr(V)}{n}$	$n^{-(\beta+\frac{1}{2})\wedge(\frac{3}{2}-\beta)}$, $v_n^\star = O(n^{-\frac{7}{6}})$
BM(11)	Convex Logist. Reg. Recurs. Quantile	No : $O(n^{-1/2})$ No : $O(n^{-1/2})$ $\varnothing$	$\varnothing$
B(14)	Logist. Reg.	No : $O\left(\frac{1}{n\lambda_{min}^2\{D^2 f(\theta^\star)\}}\right)$	$\varnothing$
CCGB(17)	Recurs. Quantile	No : $O\left(\frac{1}{n}\right)$	$n^{-(\beta+\frac{1}{2})\wedge(\frac{3}{2}-\beta)}$, $v_n^\star = O(n^{-\frac{7}{6}})$

TABLE : Overview of our results and comparisons with the literature. $v_n^\star$ refers to the optimal (smallest) size of the second-order term when β is chosen equal to $\beta^\star$.

IV - 5 Averaging - Second order term

We can theoretically improve the second order term when f is locally symmetric around $\theta^\star$ ($D^3 f(\theta^\star) = 0$)

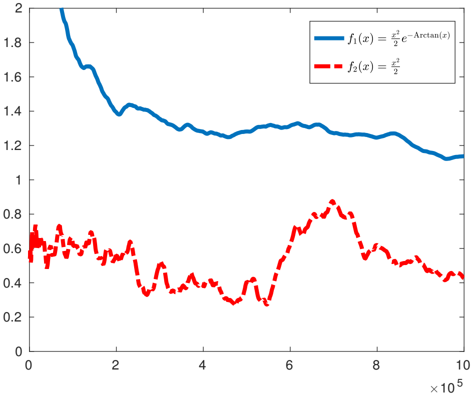


FIGURE : $n \mapsto n^\rho \left(\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] - \frac{\text{Tr}(\Sigma^\star)}{n} \right)$. Blue curve : $\rho = \frac{5}{4}$ and $\beta = \frac{3}{4}$ for a non locally symmetric function f_1. Red curve : $\rho = \frac{4}{3}$ and $\beta = \frac{2}{3}$ for a locally symmetric function f_2.

Conclusion

Conclusions :

- In stochastic cases, Ruppert-Polyak performs better than Nesterov/HBF systems
- May be shown to be optimal for quite general functions with a unique minimizer
- Conclusions may be different when dealing with multiple wells situations
- Tight bounds for recursive quantile, logistic regression, linear models,. . .

Developments :

- Sharp large deviation on $(\overline{\theta}_n)_{n\geqslant 1}$? Good idea to use the spectral representation.
- Moments of $(\overline{\theta}_n)_{n\geqslant 1}$? Other losses ?
- Non-smooth situations ?
- Improve the second order term with non-flat/uniform averaging ?

Thank you for your attention !

Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity, with F. Panloup, 2017