Non-asymptotic bound for stochastic averaging

S. Gadat and F. Panloup

Toulouse School of Economics

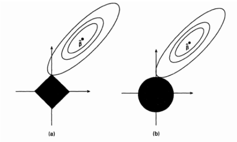
PGMO Days, November, 14, 2017

I - 1 Optimization - Motivations : Statistical problems

- Objective : Solve
$$\arg\min_{\theta\in\mathbb{R}^d} f(\theta)$$
- Motivation : minimization originates from a statistical estimation problem
- M-estimation point of view :
$$\hat{\theta}_N := \arg\min f_N(\theta)$$
where f_N is a stochastic approximation of the target function f.
- Among other, statistical problems like :
  - Supervised regression $(X_i, Y_i)_{1\leqslant i\leqslant N}$: Sum of squares in linear models
  $$f_N(\theta) = \sum_{i=1}^{N} \|Y_i - \langle X_i, \theta\rangle\|^2.$$
  - Supervised classification $(X_i, Y_i)_{1\leqslant i\leqslant N}$: Logistic regression
  $$f_N(\theta) = \sum_{i=1}^{N} \log\left(1 + \exp(-Y_i\langle X_i, \theta\rangle\right).$$
  - Quantile estimation
- Cornerstone of the talk :
$$\frac{1}{N}\mathbb{E}[f_N(\theta)] = f(\theta) \qquad \text{or} \qquad \frac{1}{N}\mathbb{E}[\nabla f_N(\theta)] = \nabla f(\theta)$$

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

- A lot of observations that may be observed recursively : large n
- A large dimensional scaling : large d
  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
- Each iteration : use only one arrival of the channel (picked up uniformly)

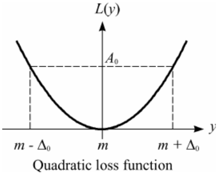
$$f_N(\theta) = \sum_{i=1}^{N} \ell_{(X_i,Y_i)}(\theta)$$

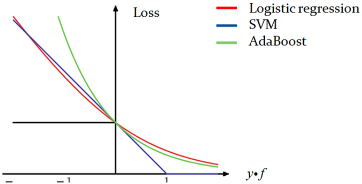
I - 2 Optimization - convexity

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

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

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





- Benchmark first order deterministic methods (with ∇f) :
  - when f is assumed to be convex, quadratic rates (NAGD) :

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

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

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

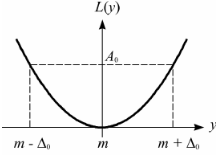
- Minimax paradigm : worst case in a class of functions within horizon t

I - 3 Stochastic Optimization - convexity

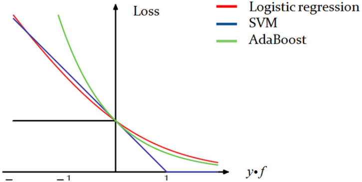
- Smooth minimization $\mathcal{C}^2$ problem

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

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 $\mathbb{E}[\xi] = 0$) :
  - when f is convex (Nemirovski-Yudin 83) :

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

  - when f is strongly convex (Cramer-Rao lower bound) :

$$O(1/t)$$

- Minimax paradigm : worst case in a class of functions within 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 ?
  Independent of some unknown quantities : $D^2 f(\theta^\star), \min_x \min Sp(D^2 f(x))$.
- Non asymptotic bound ? Exact/sharp constant ?

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

  $Tr(V)$: 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 ?

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

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

First order Markov chain stochastic approximation :

- Stochastic Gradient Descent (SGD for short) : $(\theta_n)_{n\geqslant 1}$

Second order Markov chain stochastic approximation :

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

II - 1 Stochastic Gradient Descent (SGD)

- Robbins-Monro algorithm 1951.
- Idea : use the steepest descent to produce a first order recursive method. 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

$$g_n(\theta_n) = \nabla f(\theta_n) + \xi_n,$$

  where $(\xi_n)_{n\geqslant 1}$ is a sequence of independent zero mean noise :

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

  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 $\mathbb{E}[\|\theta_n - \theta^\star\|^2] \leqslant C_\alpha n^{-1}$*

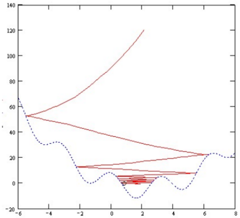
Pros : easy analysis, avoid local traps with probability 1 (Pemantle 1990, Benaïm 1996, Brandiere-Duflo 1996)
Cons : Not adaptive, no sharp inequality, no KL settings, ...

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

- Not novel (Ruppert 1988, Polyak-Juditsky 1992)
- Start from a SGD sequence $(\theta_n)_{n\geqslant 1}$ with slow step sizes

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

- Idea : Cesaro averaging all along the sequence

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

- 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 :

- Relax the convexity assumption (Kurdyka- Łojasiewicz inequality) ?
  $\hookrightarrow$ very mild assumption on the data/problem
  $\hookrightarrow$ convex semi-algebric, recursive quantile, logistic regression, strongly convex functions, . . .
  $\hookrightarrow$ Incidentally easy $\mathbb{L}^p$ **consistency rate of SGD**(!)
- Plug-in it in the Ruppert-Polyak averaging procedure ?
  $\hookrightarrow$ **Sharp non asymptotic minimax $\mathbb{L}^2$ rate for $\overline{\theta}_n$**
  $\hookrightarrow$ Spectral explanation of “why it works ?”

III - 1 Almost sure convergence

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

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

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

III - 2 Strong convexity ?

- Historically, plays a great role in optimization/stochastic optimization
- Generally : needs a strong convexity assumption to derive efficient rates
- Otherwise : each particular case is dealt with carefully

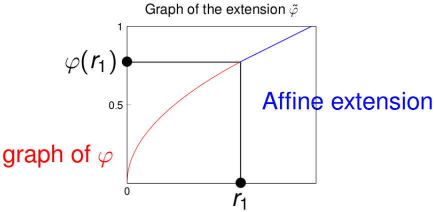
Definition (KL type inequality $\mathbf{H}_\varphi$)

$D^2 f(\theta^\star)$ invertible, an increasing asymptotically concave function ϕ exists s.t.

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

Implicitly :

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



If for a $\beta \in [0, 1]$:

$$\liminf_{|x| \to +\infty} f(x)^{-\beta}|\nabla f(x)|^2 > 0 \quad \text{and} \quad \limsup_{|x| \to +\infty} f(x)^{-\beta}|\nabla f(x)|^2 < +\infty.$$

Then, $\mathbf{H}_\varphi$ holds with $\varphi(x) = (1 + |x|^2)^{\frac{1-\beta}{2}}$.

III - 2 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 ($\varphi = 1$) and $\beta = 1$
- Online logistic regression and $\beta = 0$
- Recursive quantile problem and $\beta = 0$

Last assumption (for the sake of readability)

Assumption (Martingale noise)

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

Restrictive for the sake of readability.
Can be largely weakened with additional technicalities

III - 3 Averaging analysis Assume $\theta^\star = 0$

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 $\Lambda_n = \int_0^1 D^2 f(t\theta_n)dt$. 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 :

- We shall expect a behaviour of $(\overline{\theta}_n)_{n\geqslant 1}$ independent from $D^2 f(\theta^\star)$
- We shall 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

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

$\delta_{n,\beta}$ is an error term : variation of the eigenvectors from n to $n+1$.

If $\delta_{n,\beta}$ is small enough, then we obtain

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

Linearisation :

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

III - 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 $\mathbf{H}_\varphi$, a collection of constants C_p exists such that

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

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) :

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

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

III - 5 Averaging - Main result

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

Theorem

Under $\mathbf{H}_\varphi$*, a constant* C *exists such that*

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

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

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

- Non asymptotic
- Optimal variance term (Cramer-Rao lower bound)
- Adaptive to the unknown value of the Hessian
- Only requires invertibility of $D^2 f(\theta^\star)$
- $\beta = 3/4$ no real understanding on this optimality (just computations)
- Second order term seems to be of the good size

Conclusion

Conclusions :

- ~~In stochastic cases, Ruppert-Polyak is far 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 ? Other losses ?
- Non-smooth situations ?

Thank you for your attention !

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