Non-asymptotic bound for stochastic averaging

S. Gadat and F. Panloup

Toulouse School of Economics

PGMO Days, November, 14, 2017
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 Polyak averaging
  III - 1 Almost sure convergence
  III - 2 Strong convexity ?
  III - 3 Averaging analysis
  III - 4 Linearisation and moments
  III - 5 Averaging - Main result
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 \((X_i, Y_i)_{1 \leq i \leq N}\) : Sum of squares in linear models

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

  - Supervised classification \((X_i, Y_i)_{1 \leq i \leq N}\) : Logistic regression

\[
f_N(\theta) = \sum_{i=1}^{N} \log (1 + \exp(-Y_i \langle X_i, \theta \rangle)).
\]

  - Quantile estimation

- **Cornerstone of the talk**:

\[
\frac{1}{N} \mathbb{E}[f_N(\theta)] = f(\theta) \quad \text{or} \quad \frac{1}{N} \mathbb{E}[\nabla f_N(\theta)] = \nabla f(\theta)
\]
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 $\implies$ 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 $C^2$ 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 $\nabla f$) :
  - 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 $C^2$ problem

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

Generally, $f$ is also assumed to be convex/strongly convex

Quadratic loss/Logistic loss:

- First order 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 $C^2$ problem

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

Build a recursive optimization method $(\theta_n)_{n \geq 1}$ 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: $D^2f(\theta^*)$, $\min_x \min S p(D^2f(x))$.

- Non asymptotic bound? Exact/sharp constant?

\[ \forall n \geq N \quad \mathbb{E}\|\theta_n - \theta^*\|^2 \leq \frac{\text{Tr}(V)}{n} + A/n^{1+\epsilon}, \]

$\text{Tr}(V)$: incompressible variance (Cramer-Rao lower bound.)

- Large deviations?

\[ \forall n \geq N \quad \forall t \geq 0 \quad \mathbb{P}\left(\|\theta_n - \theta^*\| \geq b(n) + t\right) \leq e^{-R(t,n)} \]

- $L^p$ loss?

\[ \mathbb{E}\|\theta_n - \theta^*\|^{2p} \leq \frac{A_p}{n^p} + B_p/n^{p+\epsilon} \]
I - 4 No novelty in this talk, as usual!

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

First order Markov chain stochastic approximation:
- Stochastic Gradient Descent (SGD for short): \((\theta_n)_{n \geq 1}\)

Second order Markov chain stochastic approximation:
- Polyak Averaging: \((\overline{\theta}_n)_{n \geq 1}\)
- Heavy Ball with Friction (HBF)
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 Polyak averaging
   III - 1 Almost sure convergence
   III - 2 Strong convexity ?
   III - 3 Averaging analysis
   III - 4 Linearisation and moments
   III - 5 Averaging - Main result
II - 1 Stochastic Gradient Descent (SGD)

- Robbins-Monro algorithm 1951.
- Idea: use the steepest descent to produce a first order recursive method. Homogenization all along the iterations
- Build the sequence \((\theta_n)_{n \geq 1}\) 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 \((\xi_n)_{n \geq 1}\) is a sequence of independent zero mean noise:
    \[\mathbb{E} [\xi_n | \mathcal{F}_n] = 0,\]
  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
    \[\mathbb{E} [\|\theta_n - \theta^*\|_2^2] \leq C_{\alpha} \gamma_n\]
  - If \(\gamma_n = \gamma n^{-1}\) with \(\gamma \alpha > 1/2\), then
    \[\mathbb{E} [\|\theta_n - \theta^*\|_2^2] \leq C_{\alpha} n^{-1}\]

*Pros*: easy analysis, avoid local traps with probability 1 (Pemantle 1990, Benâaïm 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):

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

- 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:

\[
X_{n+1} = X_n - \gamma_{n+1} Y_n \quad \text{and} \quad Y_{n+1} = Y_n + r_n \gamma_{n+1} (g_n(X_n) - Y_n)
\]
II - 3 Polyak-Ruppert Averaging

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

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

- Idea: Cesaro averaging all along the sequence

\[
\bar{\theta}_n = \frac{1}{n} \sum_{j=1}^{n} \theta_j
\]

- Typical state of the art result

Theorem (PJ92)

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

\[
\sqrt{n}(\bar{\theta}_n - \theta^*) \longrightarrow N(0, V) \quad \text{as} \quad n \longrightarrow +\infty.
\]

\(V\) *possesses an optimal trace and* \((\bar{\theta}_n)_{n \geq 1}\) *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") :

\[
\mathbb{E}\|\bar{\theta}_n - \theta^*\|^2 \leq \frac{C}{n}
\]
We propose two contributions:

- Relax the convexity assumption (Kurdyka-Łojasiewicz inequality)?
  - very mild assumption on the data/problem
  - convex semi-algebraic, recursive quantile, logistic regression, strongly convex functions, . . .
  - Incidentally easy $\mathbb{L}^p$ consistency rate of SGD( !)

- Plug-in it in the Ruppert-Polyak averaging procedure?
  - **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 Polyak averaging
   III - 1 Almost sure convergence
   III - 2 Strong convexity?
   III - 3 Averaging analysis
   III - 4 Linearisation and moments
   III - 5 Averaging - Main result
III - 1 Almost sure convergence

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

\[
\bar{\theta}_n = \frac{1}{n} \sum_{k=1}^{n} \theta_k, \quad n \geq 1
\]

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

**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 $H_\phi$)

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

$$\exists 0 < m < M \ \forall x \in \mathbb{R}^d \setminus \{\theta^*\} : \quad m \leq \varphi'(f(x))|\nabla f(x)|^2 + \frac{|\nabla f(x)|^2}{f(x)} \leq M.$$

Implicitly:

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

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

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

Then, $H_\phi$ 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)

\[ \sup_{n \geq 1} \| \xi_{n+1} \| < +\infty \]

Restrictive for the sake of readability.
Can be largely weakened with additional technicalities
III - 3 Averaging analysis Assume $\theta^* = 0$

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

$$Z_{n+1} = \left(\begin{array}{cc}
I_d - \gamma_{n+1} \Lambda_n & 0 \\
\frac{1}{n+1} (I_d - \gamma_{n+1} \Lambda_n) & (1 - \frac{1}{n+1}) I_d
\end{array}\right) Z_n + \gamma_{n+1} \left(\begin{array}{c}
\xi_{n+1} \\
\frac{\xi_{n+1}}{n+1}
\end{array}\right),$$

where $\Lambda_n = \int_0^1 D^2 f(t \theta_n) dt$. Replace formally $\Lambda_n$ by $D^2 f(\theta^*)$.

**Key matrix**: for any $\mu > 0$ and any integer $n$:

$$E_{\mu, n} := \left(\begin{array}{cc}
1 - \gamma_{n+1} \mu & 0 \\
\frac{1 - \mu \gamma_{n+1}}{n+1} & 1 - \frac{1}{n+1}
\end{array}\right).$$

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

**Conclusion 1**:

- We shall expect a behaviour of $(\bar{\theta}_n)_{n \geq 1}$ independent from $D^2 f(\theta^*)$
- We shall expect a rate of $n^{-1}$

**Difficulties**:

- $E_{\mu, n}$ is not symmetric $\implies$ non orthonormal eigenvectors
- $E_{\mu, n}$ varies with $n$

Requires a careful understanding of the eigenvectors variations
Linear case:
How to produce a sharp upper bound? Derive an inequality of the form

$$E[\|\tilde{Z}_{n+1}\|^2 | \mathcal{F}_n] \leq \left(1 - \frac{1}{n+1} + \delta_{n,\beta}\right)^2 \|\tilde{Z}_n\|^2 + \frac{Tr(D^2 f(\theta^*))}{(n+1)^2}$$

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

$$E[\|\tilde{Z}_n\|^2] \leq \frac{Tr(D^2 f(\theta^*))}{n} + \underbrace{\epsilon_{n,\beta}}_{:=O(n^{-(1+\nu_\beta)})}$$

Linearisation:
We need to replace $\Lambda_n$ by $D^2 f(\theta^*)$ and we are done!
III - 4 Averaging analysis : cost of the linearisation

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

**Theorem**

For $\beta \in [0, 1]$, under $H_\varphi$, a collection of constants $C_p$ exists such that

$$\mathbb{E} \left[ \| \theta_n - \theta^* \|^p \right] \leq C_p \gamma_n^p$$

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):

$$\forall n \in \mathbb{N}^* \quad \mathbb{E} [V_p(\theta_{n+1}) \mid \mathcal{F}_n] \leq \left(1 - \frac{\alpha}{2} \gamma_{n+1} + c_1 \gamma_{n+1}^2 \right) V_p(\theta_n) + c_2 \{\gamma_{n+1}\}^{p+1}.$$  

Remarks:
- Important role of $\varphi$!
- Painful second order Taylor expansion . . .
We can state our main result with $\beta \in (1/2, 1), \gamma_n = \gamma_1 n^{-\beta}$:

**Theorem**

*Under $H_\varphi$, a constant $C$ exists such that*

$$\forall n \in \mathbb{N}^* \quad \mathbb{E} \left[ \| \bar{\theta}_n - \theta^* \|^2 \right] \leq \frac{Tr(V)}{n} + C n^{-\{(\beta+1/2) \wedge (2-\beta)\}}.$$

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

$$\forall n \in \mathbb{N}^* \quad \mathbb{E} \left[ \| \bar{\theta}_n - \theta^* \|^2 \right] \leq \frac{Tr(V)}{n} + C n^{-5/4}.$$

- Non asymptotic
- Optimal variance term (Cramer-Rao lower bound)
- Adaptive to the unknown value of the Hessian
- Only requires invertibility of $D^2 f(\theta^*)$
- $\beta = 3/4$ 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 $(\theta_n)_{n \geq 1}$? Good idea to use the spectral representation.
- Moments? Other losses?
- Non-smooth situations?

Thank you for your attention!