About non-asymptotic bounds for Ruppert-Polyak averaging

S. Gadat and F. Panloup

Université d'Angers

Workshop "Optimization and Learning"
IMT, 13/09/2018

Optimization : from deterministic to stochastic approaches

In whole the talk, the optimization problem is considered on $\mathbb{R}^d$.

- **General Objective** : For a given function h from $\mathbb{R}^d$ to $\mathbb{R}^d$, find some strategies to compute/approximate the zeros of h, say

$$\{\theta \in \mathbb{R}^d, h(\theta) = 0\}.$$

- **Important Case** : When $h = \nabla f$, this remains to find the critical points of f. In particular, if f is a coercive function which admits a unique critical point $\theta^\star$, the objective is to approximate the unique minimum of f.

- For the sake of simplicity, in what follows, $h = \nabla f$ and $\theta^\star$ unique minimum of f.

Optimization : from deterministic to stochastic approaches

A basic dynamics to approximate $\theta^\star$ is the gradient descent given by :

$$\dot{\mathbf{x}}_t = -\nabla f(\mathbf{x}_t)$$

and its (first order) discretized counterpart :

$$\bar{x}_{n+1} = \bar{x}_n - \gamma_{n+1}\nabla f(\bar{x}_n)$$

where $(\gamma_n)_{n\geqslant 1}$ is a (step) sequence of positive numbers (generally constant or decreasing)

- In the sequel, we assume that $(\gamma_n)_{n\geqslant 0}$ decreases to 0 and that

$$\Gamma_n = \sum_{k=1}^{n} \gamma_k \to +\infty.$$

- Such a procedure converges under appropriate assumptions but

requires the ability to compute ∇f.

Optimization : from deterministic to stochastic approaches

- **Starting point of stochastic optimization.** Assume that ∇f has the following representation :

$$\nabla f(\theta) = \mathbb{E}[\Lambda(\theta, Z)]$$

where Z is a given random variable and that the cost of computation of this expectation is too large.

- **Robbins-Monro :**Then, the strategy of stochastic optimization is to replace the dynamics $(\bar{x}_n)_{n\geqslant 0}$ by the **stochastic gradient** algorithm

$$\theta_{n+1} = \theta_n - \gamma_{n+1}\Lambda(\theta_n, Z_{n+1})$$

where (Z_n) is a sequence of $i.i.d.$ random variables. This can be rewritten

$$\theta_{n+1} = \theta_n - \gamma_{n+1}\nabla f(\theta_n) + \gamma_{n+1}\Delta M_{n+1}$$

with

$$\Delta M_n = \Lambda(\theta_{n-1}, Z_n) - \mathbb{E}[\Lambda(\theta_{n-1}, Z_n)|\mathcal{F}_{n-1}].$$

Some typical situations/applications

- Quantile estimate : For a given $\alpha > 0$ and a given distribution $\mathbb{P}_Z$ find $q_\alpha := \inf\{t \geqslant 0, F_Z(t) \geqslant \alpha\}$, *i.e.* find "the" zero of $h(\theta) = \mathbb{E}[\Lambda(\theta, Z)]$ with $\Lambda(\theta, Z) = 1_{Z \leqslant \theta} - \alpha$.

- Quantization/K-Means : For a given distribution $\mathbb{P}_X$ on a space E, and a given integer K, find the best approximation of $\mathbb{P}_X$ (in a Wasserstein sense) by a probability defined for any $\mathbf{x} = (x^{(1)}, \ldots, x^{(K)}) \in E^K$ by

$$\nu(\mathbf{x}) = \sum_{i=1}^{K} \pi_i(\mathbf{x}) \delta_{x^{(i)}}$$

where $\pi(\mathbf{x}) = \mathbb{P}(x^{(i)}$, nearest point of $X)$. This problem can be viewed as the minimization of the *distorsion* function

$$D(\mathbf{x}) = \mathbb{E}[\min_{i=1}^{K} \|X - x^{(i)}\|^p] \quad (= \mathcal{W}_p(\mathbb{P}_X, \nu(\mathbf{x}))^p).$$

Remark : Once again, optimization of a function defined as an expectation.

Some typical applications (sequel)

In statistics :

- **M-estimation** : for some observations $U_1, \ldots, U_N$, search for

$$\hat{\theta}_N := \arg\min_{\theta} \sum_{i=1}^{N} \rho(\theta, U_i)$$

where ρ is a given function (say a loss function). This can be reformulated as

$$\hat{\theta}_N = \arg\min_{\theta} \mathbb{E}[f(\theta, Z)]$$

by setting

$$f(\theta, z) = \rho(\theta, U_z) \text{and} \quad Z \sim \mathcal{U}(\{1, \ldots, N\}).$$

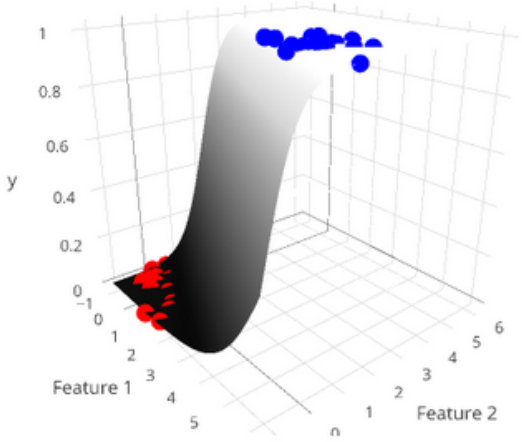
This covers a large class of statistical problems. Among them,

Some typical applications (sequel)

- (Parametric) regression (Linear Models for instance). $U_i = (X_i, Y_i)$, $Y_i = f_{\theta^\star}(X_i) + \varepsilon_i$ ($\theta^\star$ unknown true parameter)

$$\rho(\theta, U_i) = \|Y_i - f_\theta(X_i)\|^p.$$

- Supervised Classification. Logistic regression for instance.



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

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

Writing the log-likelihood to estimate $\theta^\star$, this yields

$$\rho(\theta, U_i) = \log\left(1 + e^{-Y_i <X_i, \theta>}\right).$$

- Neural Networks. . .

Convergence of Robbins-Monro-algorithms

Let us recall a classical result for RM-algorithms (in the SGD case).

Theorem (Convergence)

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be* $\mathcal{C}^2$ *and satisfy*

$$\lim_{|\theta| \to +\infty} f(\theta) = +\infty, \quad \sup_{\theta} \|D^2 f(\theta)\| < +\infty.$$

Set $\Theta = \{\theta \in \mathbb{R}^d, \nabla f(\theta) = 0\}$ *and assume that* $\mathbb{E}_{n-1}[|\Delta M_n|^2] \leqslant C(1 + f(\theta_n)$. *Then, if* $\sum \gamma_n^2 < +\infty$,

$$d(\theta_n, \Theta) \xrightarrow{n \to} 0.$$

Remark : 1. the condition $\sum \gamma_n^2 < +\infty$ may disappear under stringent assumptions on f (typically, strong convexity but not only).

2. In this talk, we only consider *smooth* problems.

Rate of convergence for Robbins-Monro-algorithms

Assume for simplicity the strong convex assumption :

$$\mathbf{SC}(\alpha) \quad D^2 f - \alpha I_d \geqslant 0 \quad \text{(in the sense of symmetric matrices).}$$

Then, denoting by $\theta^\star$ the unique mininimum, we have

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

Remark : At this stage, one remarks that the optimal order of rate of convergence can be attained with $\beta = 1$ but under a condition on γ and the unknown α.
"From a CLT point of view", we have

$$\sqrt{n^\beta}(\theta_n - \theta^\star) \Longrightarrow \mathcal{N}(0, \Sigma(\beta, \gamma, D^2 f(\theta^\star), \ldots))$$

without conditions when $\beta < 1$ but once again under a condition depending on γ and the eigenvalues of $D^2 f(\theta^\star)$. Furthermore, event though we were able to attain the optimal order of rate, we have to keep in mind that the variance depends on the choice of γ.

Remarks on the proof

- In the two previous results, the proof is based on a linearization of $\nabla f(\theta_n)$ around $\theta^\star$:

$$\nabla f(\theta_n) = D^2 f(\theta^\star)(\theta_n - \theta^\star) + r_n$$

and then to get a weakly perturbed linear dynamics. . .

- For the CLT, an interesting point of view (see $e.g.$ Duflo) is the *pseudo-diffusion* is to remark that after linearization $(Y_k)_{k \geqslant 1}$ defined by

$$Y_k = \frac{\theta_k - \theta^\star}{\sqrt{\gamma_k}}$$

is more or less the Euler scheme of an Ornstein-Uhlenbeck. In one-dimension and $\gamma_k = \gamma k^{-\beta}$ this yields

$$Y_{k+1} \approx Y_k \left(1 + \frac{\beta}{2} k^{-1} - \gamma_k f''(\theta^\star)\right) + \sqrt{\gamma_k} \Delta M_{k+1} + r_k.$$

. . . which implies that the limiting distribution can be viewed as the *invariant distribution* of an Ornstein-Uhlenbeck.

- Remark : one retrieves the constraint when $\gamma_n = \gamma n^{-1}$.

Ruppert-Polyak Algorithm

- The RP-algorithm is only defined as the average of the standard SGD :

$$\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_k.$$

- This average has the effect to overcome the limitation of Robbins-Monro. More precisely,

Theorem (Polyak-Juditsky)

Assume $\mathbf{SC}(\alpha)$, *that* $\gamma_n = \gamma n^{-\rho}$ *with* $\rho \in (1/2, 1)$,..., *and that*

$$\lim_{n\longrightarrow+\infty} \mathbb{E}[\Delta M_{n+1}\Delta M_{n+1}^T|\mathcal{F}_n] = S^\star,$$

Then,

$$\sqrt{n}(\bar{\theta}_n - \theta^\star) \Longrightarrow \mathcal{N}(0, \Sigma^\star)$$

where

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

Important feature : The variance-covariance matrix $\Sigma^\star$ is optimal in the following sense : the trace of this matrix (which appears when one takes the Euclidean norm) attains the Cramer-Rao bound.

What about (non-asymptotic) L^2-bounds

Regarding the CLT, one expects some L^2-bounds of the following form :

$$\mathbb{E}[\|\bar{\theta}_n - \theta^\star\|^2] \leqslant \frac{C_1}{n} \dots . \tag{1}$$

or sharper bounds (optimal at the first order) as :

$$\mathbb{E}[\|\bar{\theta}_n - \theta^\star\|^2] \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} \dots + \frac{C_2}{n^{r_\beta}}. \tag{2}$$

Such bounds exist in the literature in the following settings :

- Strong Convex (Bach-Moulines) : (2) with $r_\beta = \frac{7}{6}$ when $\beta = 2/3$.
- Logistic Regression/Quantile (Bach14,Cardot& *al.*) : Bounds of (1)-type.

Remark : non-exhaustive list. Nevertheless, it seems that non general result holds in the non-uniformly convex setting.

Objectives of this work

- Using a second order dynamical point of view to study the Ruppert-Polyak dynamics.
- Providing a sharp bound of type (2) under general conditions (depending only on the moments).
- Getting rid of the (uniform) strong convexity assumption.

General Assumptions

- **Assumptions on** f : f is $\mathcal{C}^2$ with a unique minimum in $\theta^\star$, $\lim_{|x|\to+\infty} f(x) = +\infty$, $D^2 f(\theta^\star)$ invertible, $x \mapsto D^2 f(x)$ Lipschitz continuous.

- $(L^p, \sqrt{\gamma_n})$**-consistency**, assumption on the original procedure : for $p > 0$, the sequence $(\theta_n)_{n\geqslant 1}$ is said to satisfy the $(L^p, \sqrt{\gamma_n})$-consistency (convergence rate condition) if $\left(\frac{\theta_n - \theta^\star}{\sqrt{\gamma_n}}\right)_{n\geqslant 1}$ is bounded in L^p, *i.e.*, if :

$$\exists\, c_p > 0 \quad \forall n \geqslant 1 \qquad \mathbb{E}|\theta_n - \theta^\star|^p \leqslant c_p \{\gamma_n\}^{\frac{p}{2}}.$$

- **Assumption** $(\mathbf{H_S})$: The covariance matrix of the martingale increment satisfies :

$$\forall \theta_n \in \mathbb{R}^d \qquad \mathbb{E}\left[\Delta M_{n+1} \Delta M_{n+1}^t | \mathcal{F}_n\right] = S(\theta_n) \qquad a.s.$$

where $S : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ is a Lipschitz continuous function :

$$\exists L > 0 \quad \forall (\theta_1, \theta_2) \in \mathbb{R}^d \qquad \|S(\theta_1) - S(\theta_2)\| \leqslant L|\theta_1 - \theta_2|.$$

General result

Theorem (L^2-non-asymptotic bound (optimal at the first order))

Let $\gamma_n = \gamma n^{-\beta}$ *with* $\beta \in (1/2, 1)$. *Assume the previous assumptions with* $p = 4$. *Then, a constant* C *exists such that for any* $n \geqslant 1$,

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[|\overline{\theta}_n - \theta^\star|^2\right] \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} + Cn^{-r_\beta}, \tag{3}$$

where $\Sigma^\star = \{D^2 f(\theta^\star)\}^{-1} S(\theta^\star) \{D^2 f(\theta^\star)\}^{-1}$ *and*

$$r_\beta = \left(\beta + \frac{1}{2}\right) \wedge (2 - \beta).$$

In particular, $r_\beta > 1$ *for all* $\beta \in (1/2, 1)$ *and* $\beta \longmapsto r_\beta$ *attains its maximum for* $\beta = 3/4$, *which yields :*

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

Second order term

- Optimal at the first order but what about the second order ?
- The order $\frac{5}{4}$ improves the one of Bach/Moulines $\frac{7}{6}$.
- $\frac{5}{4}$ is in fact optimal : one can build some functions f such that the second order is lower-bounded by $n^{-\frac{5}{4}}$.
- However, one can theoretically improve the second order term is f is locally symmetric around $\theta^\star$ ($D^3 f(\theta^\star) = 0$).

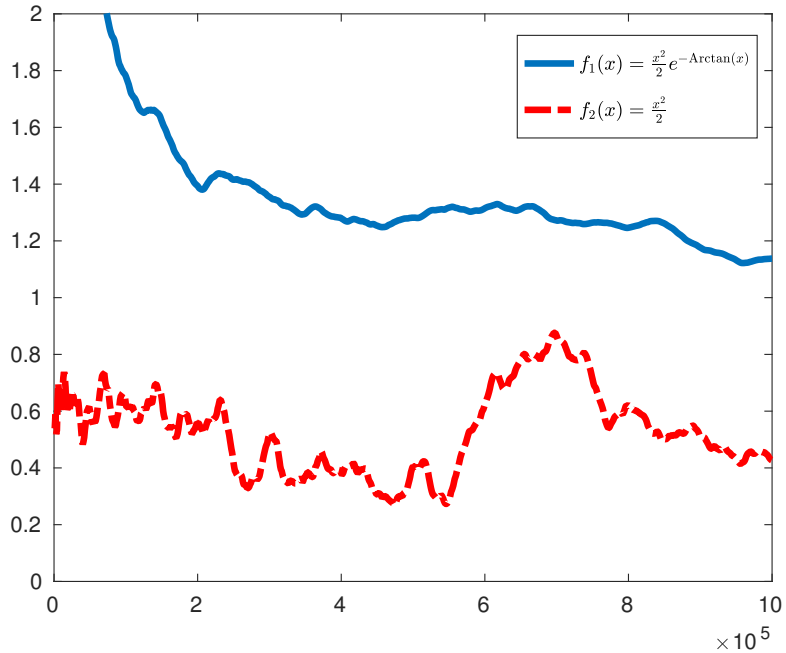


FIGURE: $n \mapsto n^\rho \left(\mathbb{E}[|\bar{\theta}_n - \theta^\star|^2] - \frac{\mathrm{Tr}(\Sigma^\star)}{n} \right)$. Blue curve : $\rho = \frac{5}{4}$ and $\beta = \frac{3}{4}$ for a non locally symmetric function f_1. Red curve : $\rho = \frac{4}{3}$ and $\beta = \frac{2}{3}$ for a locally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

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

Averaging analysis

Linear case :

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

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

where

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

$\delta_{n,\beta}$ 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

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

Linearisation :

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

Averaging analysis : cost of the linearisation

- We need to replace Λ_n by $D^2 f(\theta^\star)$
- At this stage, one needs some preliminary controls on the SGD $(\theta_n)_{n \geqslant 1}$ (moments) which correspond the L^p-consistency assumption.

Applications of the general result

- At this stage, the question is : in which setting can we apply the general result ?

- Regarding the assumptions, the problem is reduced to look for general conditions for the $(L^p, \sqrt{\gamma_n})$-consistency (with $p = 4$), that is

$$\exists\, c_p > 0 \quad \forall n \geqslant 1 \qquad \mathbb{E}|\theta_n - \theta^\star|^p \leqslant c_p \{\gamma_n\}^{\frac{p}{2}}.$$

- Under $SC(\alpha)$ ($D^2 f \geqslant \alpha I_d$), the problem is easy and only requires standard assumptions.

  More involved : without this strong convexity assumption ? (including Logistic regression/Quantile Estimation)

Contractivity under weak convexity

- Consider the function V defined by $V(x) = \exp(\phi(f(x)))$ where ϕ is a smooth function (to be calibrated). Remark that

$$\langle \nabla V(x), \nabla f(x) \rangle = \phi'(f(x)) |\nabla f(x)|^2 V(x).$$

- Then, if

$$\inf_{x \in \mathbb{R}^d} \phi'(f(x)) |\nabla f(x)|^2 > 0.$$

  one can hope to retrieve a contraction effect, that is something like

$$\mathbb{E}[V(\theta_{n+1})] \leqslant \mathbb{E}[V(\theta_n)](1 - \rho\gamma_{n+1}) + \dots$$

- example : if

$$0 < m \leqslant |\nabla f|^2 \leqslant M < +\infty \quad \text{(quantile estimation for instance)},$$

  and $\phi(x) = x$, then

$$\inf_{|x| \to +\infty} \phi'(f(x)) |\nabla f(x)|^2 = \inf_{x \in \mathbb{R}^d} |\nabla f(x)|^2 > 0.$$

Kurdyka-Lojasiewicz-type Assumptions

Assumption $(\mathbf{H}_\phi)$ - **Weakly reverting drift** The function f is $\mathcal{C}^2$ with $D^2 f$ bounded and Lipschitz, $D^2 f(\theta^\star)$ invertible and :

- $i)$ ϕ is $\mathcal{C}^2(\mathbb{R}_+, \mathbb{R}_+)$ non-decreasing and $\exists\, x_0 \geqslant 0 \,:\, \forall x \geqslant x_0,\, \phi''(x) \leqslant 0$.
- $ii)$ Two positive numbers m and M exist such that $\forall x \in \mathbb{R}^d \backslash \{\theta^\star\}$:

$$0 < m \leqslant \phi'(f(x))|\nabla f(x)|^2 + \frac{|\nabla f(x)|^2}{f(x)} \leqslant M. \tag{4}$$

- With power-functions ϕ ($\phi(x) = (1+|x|^2)^{\frac{1-2r}{2}}$), the assumption reads

$$\liminf_{|x|\longrightarrow+\infty} f^{-r}|\nabla f| > 0 \qquad \text{and} \qquad \limsup_{|x|\longrightarrow+\infty} f^{-r}|\nabla f| < +\infty \tag{5}$$

- This type of assumption lies in the family of Kurdyka-Łojasiewicz assumptions (coming from deterministic optimization)

Cost on the the noise ?

- Of course, working with exponential Lyapunov functions requires more constraining assumptions on the noise : to get the contraction property with the function $V_p(x) = f^p(x)\exp(\phi(f(x)))$, this yields the following technical condition :

$$\forall u \geqslant 0 \qquad \mathbb{E}[|\Delta M_n|^{2p+2} e^{\phi(u|\Delta M_n|^2)} | \mathcal{F}_n] \leqslant \rho_p(u) \qquad \text{a.s.} \qquad (6)$$

  where ρ_p is a locally bounded function.

- **Remark** : In the case of bounded stochastic gradient (quantile/logistic regression), this condition is always satisfied.

Theorem

Under $(\mathbf{H}_\phi)$ *and the noise condition,*

$$\mathbb{E}[f^p(\theta_n) e^{\phi(f(\theta_n))}] \leqslant C_p \{\gamma_n\}^p,$$

so that the RM-algorithm is L^{2p}*-consistent and the conclusion of the general theorem holds true.*

Some applications

- Logistic Regression : the Ruppert-Polyak algorithm related to

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \frac{Y_n X_n}{1 + e^{Y_n < \theta_n, X_n >}} = \theta_n - \gamma_{n+1} \nabla f(\theta_n) + \gamma_{n+1} \Delta M_{n+1}. \quad (7)$$

  satisfies the L^2-bound

- Same result for the quantile.

- Towards non-convex assumptions ? These results may apply to some "nice" non convex situations with at least linear increase for the function f (think to oscillating f'').

- It may be possible to extend to functions with sublinear increase at infinity : $|x|^r$ with $r < 1$ (Non-convex !) but under stringer and stringer assumptions on the noise (that is, the stochastic gradient decreases to 0 at infinity).

Conclusion

- 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 :

- Concentration inequalities $(\overline{\theta}_n)_{n\geqslant 1}$? 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 !