2011/70

■

Stochastic first order methods in smooth convex optimization

Olivier Devolder

DISCUSSION PAPER

Center for Operations Research
and Econometrics

Voie du Roman Pays, 34
B-1348 Louvain-la-Neuve
Belgium
http://www.uclouvain.be/core

CORE DISCUSSION PAPER
2011/70

Stochastic first order methods in smooth convex optimization

Olivier DEVOLDER[1]

December 2011

Abstract

In this paper, we are interested in the development of efficient first-order methods for convex optimization problems in the simultaneous presence of smoothness of the objective function and stochasticity in the first-order information. First, we consider the Stochastic Primal Gradient method, which is nothing else but the Mirror Descent SA method applied to a smooth function and we develop new practical and efficient stepsizes policies. Based on the machinery of estimates sequences functions, we develop also two new methods, a Stochastic Dual Gradient Method and an accelerated Stochastic Fast Gradient Method. Convergence rates on average, probabilities of large deviations and accuracy certificates are studied. All of these methods are designed in order to decrease the effect of the stochastic noise at an unimprovable rate and to be easily implementable in practice (the practical efficiency of our method is confirmed by numerical experiments). Furthermore, the biased case, when the oracle is not only stochastic but also affected by a bias is considered for the first time in the literature.

Keywords: stochastic optimization, stochastic approximation methods, smooth convex optimization, first-order methods, fast gradient method, complexity bounds, probability of large deviations.

---

[1] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium.
E-mail: Olivier.devolder@uclouvain.be.

The author is a F.R.S.-FNRS Research Fellow. This text presents research results obtained during a research stay at the H. Milton Stewart School of Industrial and System Engineering at the Georgia Institute of Technology during the fall of 2011. The author would like to thanks warmly professor Arkadi Nemirovski for all his support, the highly rewarding discussions and the previous advices that have clearly improved the content of this paper. This research stay has been financed thanks to a F.R.S.-FNRS grant.

This paper presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the author.

1 Introduction

This paper is devoted to the development of efficient first-order methods for convex optimization problems of the form $\min_{x\in Q} f(x)$ where f is a smooth convex function but endowed with a stochastic first-order oracle.

In the deterministic convex case, smoothness is a highly desirable property. Indeed, for a non-smooth Lipschitz-continuous function (with constant M), the best convergence rate for $f(y_k)-f^*$ (where k is the iteration counter and y_k the approximate solution generated after k iterations) that we can expect , using a first-order method, has the form $O\left(\frac{MR}{\sqrt{k}}\right)$ where R represents the distance between the initial iterate and the optimal solution. This slow rate is achieved for example by subgradient type methods (see for example [14, 6]).

On the other hand, when the objective function is smooth with a Lipschitz-continuous gradient (with constant L), the convergence rate of the (sub)gradient method becomes $O\left(\frac{LR^2}{k}\right)$ and it is even possible to obtain a convergence rate $O\left(\frac{LR^2}{k^2}\right)$ (optimal for deterministic smooth problem) using the fast gradient methods developed in various variants by Nesterov since 1983 ([12, 13, 14, 15]).

In the stochastic convex case, when the first-order information is affected by a random noise, the most classical first order methods are the Stochastic Approximation (SA) methods that mimic the subgradient method, replacing the exact gradient by the stochastic one. In the modern SA methods, like the Mirror Descent SA method (see [11]), the function endowed with a stochastic oracle is typically assumed to be non-smooth and the obtained convergence rate is $O\left(\frac{MR}{\sqrt{k}}+\frac{\sigma R}{\sqrt{k}}\right)$ where σ is the level of the stochastic noise. This rate has an optimal dependence in M (since the problem is non-smooth) but also an optimal dependence in σ. Indeed, it has been proved in [10] that the effect of the stochastic noise cannot be decreased, by a first-order method, with a better rate than $\frac{1}{\sqrt{k}}$ and this limitation is also valid when the function is smooth. This result has led to the common belief that in the presence of a stochastic oracle, the smoothness of the objective function is useless. It does not matter that the function is smooth or not, in any case we come back to a slow convergence rate $O\left(\frac{1}{\sqrt{k}}\right)$ like in the deterministic non-smooth case. However when the Lipschitz constant of the gradient L is big as compare to the stochastic noise σ, and when we are interested in solution with moderate accuracy, a convergence rate of the form $O\left(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}\right)$ or $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$, exploiting the smoothness of f in its first term, can be significantly better than $O\left(\frac{MR}{\sqrt{k}}+\frac{\sigma R}{\sqrt{k}}\right)$.

First-order methods in the stochastic smooth case has been considered for the first time by Lan in [7]. In this paper, he adapts the Mirror descent SA method, designed initially for non smooth problem, to the smooth case, obtaining a convergence rate of the form $O\left(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}\right)$ and adapts one variant of the fast gradient methods, initially designed for deterministic smooth problems to the stochastic case, obtaining a convergence rate of the form $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$. The development of fast gradient methods in the smooth stochastic case with applications in machine learning problems has also been considered recently in [3, 9].

The new first-order methods that we develop in this paper exhibit the same kind of convergence rates but our methods are characterized by some common properties that extend their applicability in practice:

1. Our methods can be used with a general norm (not especially the Euclidean norm)

and can be adapted to the problem geometry, using a good setup and therefore auxiliary subproblems as easy as possible. This desirable property is not satisfied by the methods developed in [3, 9]. In these papers, the methods use auxiliary subproblems based on the squared norm and such quadratic function is sometimes difficult to minimize on the feasible set. We answer the question of the setup choice in the Section 2.2.

2. Our methods use stepsizes that does not need to know a priori the performed number of iterations. This property is highly desirable when we want to run a method for a given time and not for a given number of iterations (for example when we compare methods with different iteration costs). On the contrary, the methods developed in [7, 9] assume an a priori knowledge of the performed number of iterations N and use stepsizes based on this number. We discuss in details the stepsizes choice for each method in the Sections 3, 5, 6 and show in the Section 9, on the numerical experiments, the advantage of stepsizes policies not based on N.
3. Our methods can be applied, without modification of the convergence rate, to the composite case where we add to the smooth objective function f, an easy convex function h (potentially non-smooth) that can be kept in the auxiliary subproblems used by the first-order methods. This composite case, when f is endowed with a stochastic oracle and h is easy, has been already considered in [9] but not in [7, 3].

   Remark 1 Lan in [7] considers a different composite case where the non-smooth part of the function is also endowed with a stochastic black-box oracle. In our case, we use the explicit structure of the (possibly) non-smooth component, avoiding in this way that h slows down the convergence rate.

Furthermore, to the best of our knowledge, this paper considers for the first time, the biased case i.e. when the smooth function f is endowed with an oracle which is not only stochastic (with stochastic noise σ) but also biased (with bias δ), meaning that on average, the stochastic first-order information does not coincide with the exact one.

The paper is organized as follows. In Section 2, we present in a more formal form our problem class, introduce different possible setups that can be used by the first-order methods and present three simple examples of smooth convex problems with stochastic oracle. In some cases, the stochasticity is in the problem since the beginning. In other cases, we introduce ourself the stochasticity via a randomization of the first-order information in order to reduce the computational cost of the first-order methods. In Section 3, we develop new practical stepsizes policy for the Stochastic Primal Gradient Method (SPGM) which is nothing else than the Mirror-Descent SA method (see [11, 7]) but applied to a smooth convex problem. In Section 4, we introduce the machinery of estimate functions and generalize it to the stochastic case. Based on this principle, we develop and study the average behavior of two new methods, a Stochastic Dual Gradient Method (SDGM) with convergence rate $O\left(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}\right)$ (Section 5) and a Stochastic Fast Gradient Method (SFGM) with convergence rate $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$ (Section 6). All these methods decrease the effect of the stochastic noise at the unimprovable rate $O\left(\frac{\sigma R}{\sqrt{k}}\right)$ where R represents the distance between the initial iterate x_0 and the optimal solution x^* and k is the iteration counter. In Section 7 and 8, we study the probabilities of large deviations for these methods and develop accuracy certificates. The last section is devoted to numerical experiments. We consider quadratic problems on the simplex when the gradient is affected by a stochastic noise and compare our methods (using different possible stepsizes policies) with the existing methods.

2 Smooth convex problem with stochastic oracle

2.1 Problem class and biased stochastic oracle

Let E be a finite dimensional vector space endowed with the norm $\|.\|$ and E^*, the dual space of E, with the dual norm $\|g\|_* = \sup_{y\in E}\{|\langle g, y\rangle| : \|y\| \leq 1\}$ where $\langle ., .\rangle$ denotes the dual pairing. We consider the convex optimization problem:

$$\phi^* = \min_{x\in Q} \phi(x) \tag{2.1}$$

where $Q \subset E$ is a closed convex set, $\phi = f + h$ and

- $f : Q \to \mathbb{R}$ is a convex function, typically smooth but endowed with a stochastic first-order oracle (possibly biased)
- $h : Q \to \mathbb{R}$ is an easy convex function. Easy means that we can easily minimize the sum of h and a well-chosen model of f on the set Q. This property will be explained in more details later in the next subsetion.

The stochastic first-order oracle available for f is characterized by two-levels of inexactness:

- The function f is endowed with a (δ, L)-oracle (this notion of oracle with deterministic error has been introduced recently in [2]) i.e. that for each $x \in Q$, we could potentially compute $f_{\delta,L}(x) \in \mathbb{R}$ and $g_{\delta,L}(x) \in E^*$ such that:

$$0 \leq f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x\rangle \leq \frac{L}{2}\|x - y\|^2 + \delta, \quad \forall y \in Q. \tag{2.2}$$

- We do not use $(f_{\delta,L}(x), g_{\delta,L}(x))$ but instead stochastic estimates $(F_{\delta,L}(x,\xi), G_{\delta,L}(x,\xi))$. More precisely, at all point $x \in Q$, we associate with x a random variable X whose probability distribution is supported on $\Xi \subset \mathbb{R}^d$ and such that:

$$E_{\xi\sim X}[F_{\delta,L}(x,\xi)] = f_{\delta,L}(x) \tag{2.3}$$

$$E_{\xi\sim X}[G_{\delta,L}(x,\xi)] = g_{\delta,L}(x) \tag{2.4}$$

$$E_{\xi\sim X}[\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*^2] \leq \sigma^2. \tag{2.5}$$

When $\delta = 0$, f is necessarily smooth with a Lipschitz-continuous gradient (with constant L i.e. $f \in F_L^{1,1}(Q)$) and the oracle is stochastic but unbiased: $E_{\xi\sim X}[F_{\delta,L}(x,\xi)] = f(x)$ and $E_{\xi\sim X}[G_{\delta,L}(x,\xi)] = \nabla f(x)$.
When $\delta \neq 0$, this kind of oracle can be seen as a biased stochastic oracle where σ represents the stochastic noise and δ the deterministic bias. Indeed, the notion of (δ, L) oracle (introduced in [2]) allows us to consider different natural notions of bias:

- $g_{\delta,L}(x)$ **is an approximate gradient of** f
  If $f \in F_{\overline{L}}^{1,1}(Q)$, $\|\nabla f(x) - g_{\delta,L}(x)\|_* \leq \Delta$ and Q is bounded with diameter $D = \max_{x\in Q, y\in Q}\|x - y\|$ then $(f_{\delta,L}(x) = f(x) - \Delta D, g_{\delta,L}(x))$ is a (δ, L) oracle with $\delta = 2\Delta D$ and $L = \overline{L}$.
- $g_{\delta,L}(x)$ **is a gradient of** f **computed at a shifted point** $\overline{x}$
  If $f \in F_{\overline{L}}^{1,1}(Q)$ and $g_{\delta,L}(x) = \nabla f(\overline{x})$ then
  $(f_{\delta,L}(x) = f(\overline{x}) + \langle \nabla f(\overline{x}), x - \overline{x}\rangle, g_{\delta,L}(x) = \nabla f(\overline{x}))$ is a (δ, L) oracle with $\delta = \overline{L}\|x - \overline{x}\|^2$ and $L = 2\overline{L}$.
- f **is in fact non-smooth and** $g_{\delta,L}(x)$ **is a subgradient of** f
  If f is non-smooth with bounded variations of subgradients i.e.:

$$\|g(x) - g(y)\|_* \leq M, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x), g(y) \in \partial f(y)$$

then $(f_{\delta,L}(x) = f(x), g_{\delta,L}(x) = g(x))$ is a (δ, L) oracle with δ an arbitrary positive constant and $L = \frac{M^2}{2\delta}$. (In this case, the bias δ correspond to the fact that the function is not as smooth as expected).

Remark 2 We use the denomination smooth convex problem for (2.1) even if the functions f and h can both be non-smooth. The reason is the fact that the component h does not play any role in the design and the convergence rate of the first-order methods that we will consider. Furthermore, the function f is typically a smooth convex function with Lipschitz-continuous gradient. A non-smooth f can be also considered but the non-smoothness is seen in this case as a bias with respect to the desired situation (using the notion of (δ, L) oracle). This generality is not the main goal of this paper, we are mainly interested in the minimization of a smooth convex function f endowed with stochastic oracle (augmented eventually by an easy non-smooth convex function h).

Remark 3 The first-order methods developed in this paper will use only stochastic estimates of the gradient $G_{\delta,L}(x_i, \xi_i)$ at different search points x_i, never the corresponding estimates of the function value. We need $F_{\delta,L}(x, \xi)$, only when we want to estimate the quality of a point $x \in Q$ for the objective function (see section 8).

2.2 Setup of first-order methods

In order to apply a first-order method to problem (2.1), we need to chose a metric i.e.:

1. a norm $\|.\|$ on E
2. a prox-function $d(x)$ i.e. a differentiable and strongly convex function on Q.

Let x_0 be the minimizer of d on Q. By translating and scaling d if necessary, we can always ensure that

$$d(x_0) = 0, \quad d(x) \geq \frac{1}{2}\|x - x_0\|^2, \quad \forall x \in Q. \tag{2.6}$$

We define also the corresponding Bregman distance:

$$V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle. \tag{2.7}$$

Due to the strong convexity of $d(x)$ with parameter 1, we have clearly:

$$V(x, z) \geq \frac{1}{2}\|x - z\|^2, \quad \forall x, z \in Q. \tag{2.8}$$

All the first-order methods that we will consider are based on subproblems of the forms:

$$\min_{x \in Q} \{\langle g, x \rangle + \beta d(x) + h(x)\}$$

with $g \in E^*$, $\beta \in \mathbb{R}_0^+$, or

$$\min_{x \in Q} \{\langle g, x \rangle + \beta V(x, z) + h(x)\}$$

with $g \in E^*$, $\beta \in \mathbb{R}_0^+$ and $z \in Q$.
The prox-function must be chosen such that these kinds of auxiliary subproblems can be solved easily. Of course it is possible to make these subproblems easy by a good choice of the prox-function only if the function h is itself sufficiently easy.

Example 1 When $E = \mathbb{R}^n$, two classical setups are:

1. The Euclidean setup: $\|.\| = \|.\|_2 = \sqrt{\sum_{i=1}^n (x^i)^2}$ and $d(x) = \frac{1}{2}\|x - x_0\|_2^2$ with $x_0 \in Q$. In this case, the prox-center is x_0 and $V(x, z) = \frac{1}{2}\|x - z\|_2^2$.

2. When $Q = \Delta_n = \{x \in \mathbb{R}^n_+, \sum_{i=1}^n x^i = 1\}$, the l_1 setup: $\|.\| = \|.\|_1 = \sum_{i=1}^n \left|x^i\right|$ and $d(x) = \ln(n) + \sum_{i=1}^n x^i \ln(x^i)$ (entropy distance). In this case, the prox-center is $x_0 = \left(\frac{1}{n}, ..., \frac{1}{n}\right)^T$ and $V(x,z) = \sum_{i=1}^n x^i \ln\left(\frac{x^i}{z^i}\right)$.

For the analysis of our first-order methods, we denote by R, the quantity $\sqrt{d(x^*)}$ that represents in some sense the distance between the initial iterate x_0 (which is the minimizer of the prox-function) and the optimal solution x^*. As $d(x_0) = 0$ and $\langle \nabla d(x_0), x^* - x_0 \rangle \geq 0$, we have:

$$V(x^*, x_0) \leq d(x^*) = R^2.$$

2.3 Examples

Before developing different stochastic first-order methods, we present some examples of problems of the form 2.1 with stochastic oracle.

2.3.1 Lasso problem with stochastic gradient

The Lasso problem corresponds to problem (2.1) with $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, $h(x) = \lambda \|x\|_1$ with $\lambda > 0$ and $Q = \mathbb{R}^n$. When using the Euclidean setup, the sparsity promoter $h(x) = \lambda \|x\|_1$ can be considered as an easy convex function. Indeed for all $g \in \mathbb{R}^n$ and $\lambda, \beta \in \mathbb{R}^+_0$, we have:

$$\arg\min_{x \in \mathbb{R}^n} \{\langle g, x\rangle + \lambda \|x\|_1 + \frac{\beta}{2}\|x - z\|_2^2\} = \tau_{\frac{\lambda}{\beta}}(z - \frac{1}{\beta} g)$$

where $\tau_\alpha(x)^i = (\left|x^i\right| - \alpha)_+ \text{sgn}(x^i)$ is the shrinkage operator.
We are interested in situations where $\nabla f(x)$ is not computed exactly.

- One possible situation is when the computation of $\nabla f(x)$ is really affected by a stochastic noise and a bias. This is the case for example when instead of computing $\nabla f(x) = A^T Ax - A^T b$, we are only able to compute $G_{\delta,L}(x, \xi) = A^T A\overline{x} - A^T b + \xi$ where:
  1. ξ is a stochastic perturbation such that $E[\xi] = 0$ and $E[\|\xi\|_2^2] \leq \sigma^2$
  2. $\overline{x}$ is a shifted point of x such that $\|x - \overline{x}\|_2^2 \leq \frac{\delta}{\lambda_{\max}(A^T A)}$.
- Another situation is when the stochasticity is not present in the problem initially but we introduce it in order to reduce the computational cost of the first-order information. In the Lasso problem, introducing a randomization can be interesting for example when the number of row N of A is very large. In this case, denoting by a_i the ith row of A, the computation of the exact gradient $\nabla f(x) = \sum_{i=1}^N (x^T a_i - b_i) a_i$ can be very expensive ($O(nN)$ basics operations). It can be interesting to replace $\nabla f(x)$ by an unbiased estimate $G_{0,L}(x, \xi) = \frac{N}{M} \sum_{j=1}^M (x^T a_{\xi_j} - b_{\xi_j}) a_{\xi_j}$ where $\{\xi_1, ..., \xi_M\}$ is a subset of arrows uniformly chosen from $\{1, ..., N\}$. When M is chosen significantly smaller than N, the computation of this stochastic gradient is of course cheaper. However replacing the exact gradient by this stochastic estimate introduces a stochastic noise σ that depends on dissimilarities between different rows of A.

2.3.2 Smooth Expectation function

Let X be a random vector supported on $\Xi \subset \mathbb{R}^d$. Assume that f itself is defined by an expectation:

$$f(x) = E_{\eta \sim X}[F(x, \eta)] = \int_\Xi F(x, \eta) dP(\eta),$$

where $F(., \eta) \in F^{1,1}_{L(\eta)}(Q)$ for almost all $\eta \in \Xi \subset \mathbb{R}^d$. Then we have $\nabla f(x) = E_{\eta \sim X}[\nabla_1 F(x, \eta)]$ (see [17]) and $f \in F^{1,1}_L(Q)$ where $L = \int_\Xi L(\eta) dP(\eta)$ (assuming that $L(.)$ is integrable on

Ξ i.e. that $L < \infty$). However the computation of $\nabla f(x)$ i.e. of a multidimensional integral is to costly when the dimension d is high. Therefore it is typical to replace $\nabla f(x)$ by a stochastic gradient: we sample from the distribution of X, obtaining $\xi \in \Xi$ and compute $G_{0,L}(x,\xi) = \nabla_1 F(x,\xi)$. This stochastic gradient is unbiased (i.e. $\delta = 0$): $E_{\xi\sim X}[G_{\delta,L}(x,\xi)] = \nabla f(x)$ and the noise that we introduce can be characterized by

$$\sigma^2 = E_{\xi\sim X}[\|\nabla f(x) - G_{0,L}(x,\xi)\|^2] = \int_\Xi \left\| \int_\Xi (\nabla_1 F(x,\eta) - \nabla_1 F(x,\xi)) dP(\eta) \right\|^2 dP(\xi).$$

Of course, we can also add to f an easy convex function h, like a sparsity promoter $h(x) = \lambda \|x\|_1$.

2.3.3 Randomization of Quadratic Problem

We consider the situation where

1. $f(x) = l(x) + x^T A x$ with $l \in F^{1,1}_{L_l}(Q)$ and $A \succeq 0$
2. $h(x) = 0$
3. $Q = \Delta_n = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x^i = 1\}$.

For such a problem on the simplex, it is natural to use the l_1 setup (it does not means that we add to f the l_1 norm, $h(x) = \lambda \|x\|_1$, but only that we use $\|.\| = \|.\|_1$ and the entropy prox-function).

When the problem size is very large and when the computational cost of ∇h is not too expensive, the matrix vector product Ax becomes the dominant cost in the computation of $\nabla f(x)$. It could be very interesting to replace the costly matrix vector product by a randomized one. One possibility is to pick up from A the column i with probability x^i and to consider Ae_i (i.e the $i$the column of A) as the stochastic estimate of Ax. This randomization technique for matrix-vector multiplication on the unit simplex has been introduced recently in [4]. The obtained oracle is unbiased (i.e. $\delta = 0$) and introduces a noise of order $\|A\|_\infty$ that can be reasonable when $L_l >> \|A\|_\infty$.

3 Stochastic Primal Gradient Method

3.1 Scheme

In this method, we use only one sequence of coefficients $\{\beta_k\}_{k\geq 0}$. We assume that $\beta_k > L$ for all $k \geq 0$ and denote $\gamma_k = \frac{1}{\beta_k}$ (that can be interpreted as the stepsize).

Stochastic Primal Gradient Method (SPGM)

- **Initialization**
  Compute $x_0 = \arg\min_{x\in Q} d(x)$
- **Iteration** $k \geq 0$
  1. Let ξ_k be a realization of the random variable X_k
  2. Compute $G_{\delta,L}(x_k,\xi_k)$
  3. Compute $x_{k+1} = \arg\min_{x\in Q}[\langle G_{\delta,L}(x_k,\xi_k), x - x_k\rangle + h(x) + \beta_k V(x,x_k)]$
- **Approximate Solution**
  Compute $y_k = \frac{1}{\sum_{i=0}^{k-1}\gamma_i} \sum_{i=0}^{k-1} \gamma_i x_{i+1}$.

Remark 4 In the litterature, the stochasticity is typically assumed to enter the scheme via i.i.d. random variables. Here, we consider a more general situation where a random variable X is associated with each $x \in Q$. It means that:

1. The distribution of X_i depends only on the current iterate x_i, not on the history of the process $\xi_{[i]} = (\xi_0, ..., \xi_{i-1})$ that has led the scheme to the point x_i
2. The random variables $X_0, ..., X_k$ can have different distributions but must satisfy the uniform bounds 2.3,2.4 and 2.5 with the same σ and the same δ.

Of course, if we consider the particular case where all the random variables X have the same distribution, independently of x, we come back to the i.i.d. case. We will only use this i.i.d. assumption in the Section 7 and 8 in order to develop probabilities of large deviations.

The Primal Gradient Method is the most natural, classical first-order method.
In the deterministic smooth case, when the Euclidean setup is used and $h = 0$ we retrieve the classical gradient method (see [14]):

$$x_{k+1} = \arg\min_{x \in Q}\{f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{\beta_k}{2}\|x - x_k\|_2^2\} = \pi_Q(x_k - \frac{1}{\beta_k}\nabla f(x_k))$$

where π_Q denotes the Euclidean projection on Q.
If we choose all the coefficients β_i equal to the the Lispchitz constant of the gradient L, we obtain the famous convergence rate $O\left(\frac{LR^2}{k}\right)$ (which is however non-optimal for smooth convex problems).

This familly of schemes has also attracted a lot of attention in non-smooth convex optimization, it is simply the subgradient method ([18]) if we use the Euclidean setup and the Mirror Descent method ([10, 1]) with a general setup. With an increasing sequence of coefficients $\beta_i = \Theta\left(\frac{M\sqrt{i}}{R}\right)$, we obtain the optimal convergence rate $O\left(\frac{MR}{\sqrt{k}}\right)$ for deterministic non-smooth convex probem where M denotes the Lipschitz-constant of the function.

In stochastic non-smooth convex optimization, this scheme corresponds to the Stochastic Approximation (SA) method in the Euclidean case and to the Mirror Descent Stochastic Approximation (MDSA) method ([11]) in the general case. With the same kind of decreasing stepsizes γ_i (i.e. of increasing coefficients β_i) than in the deterministic case, these methods reach the unimprovable convergence rate $O\left(\frac{MR}{\sqrt{k}} + \frac{\sigma R}{\sqrt{k}}\right)$ where σ denotes the stochastic noise of the oracle.

In stochastic smooth convex optimization, this scheme has been considered recently by Lan in [7] under the name of Modified Mirror Descent SA method (MMDSA). He proposes to construct the approximate solution (i.e. the point for which we have the convergence rate) as $\frac{1}{\sum_{i=0}^{k-1}\gamma_i}\sum_{i=0}^{k-1}\gamma_i x_{i+1}$ (instead of $\frac{1}{\sum_{i=0}^{k}\gamma_i}\sum_{i=0}^{k}\gamma_i x_i$ for the usual MDSA method) and a constant stepsize policy but which is based on the oracle noise σ and on the performed number of iterations k. This method exhibits the rate of convergence $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$ which is optimal whith respect to the stochastic noise σ but not with respect to L, the Lipschitz-constant of the gradient.

In this section, we generalize the result of Lan in three directions:

- We consider the biased case, when the expectation of the stochastic gradient $G_{\delta,L}(x, \xi)$ is itself affected by a deterministic error δ. In this case, the convergence rate is $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$
- We propose a new stepsize policy that does not need anymore the knowledge of the performed number of iterations and gives the same convergence rate (up to a logarithmic factor)

- We consider the composite case when we add to f an easy convex function h (possibly non-smooth) i.e. that can be kept without modification in the auxiliary subproblems.

But first, let us start with the general convergence rate of this Stochastic Primal Gradient Method (SPGM):

3.2 General Convergence Rate

Theorem 1 For all $k \geq 0$, we have:

$$\phi(y_k) - \phi^* \leq$$

$$\frac{1}{\sum_{i=0}^{k-1} \gamma_i} \left(V(x^*, x_0) + \sum_{i=0}^{k-1} \frac{\gamma_i}{\beta_i - L} \left\| G_{\delta,L}(x_i) - g_{\delta,L}(x_i) \right\|_*^2 + \sum_{i=0}^{k-1} \gamma_i \langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - x_i \rangle \right) + \delta.$$

Proof. For simplicity, in all proofs of this paper, we denote $f_i = f_{\delta,L}(x_i)$, $F_i = F_{\delta,L}(x_i, \xi_i)$, $g_i = g_{\delta,L}(x_i)$ and $G_i = G_{\delta,L}(x_i, \xi_i)$.
Let $gh(x_{k+1}) \in \partial h(x_{k+1})$, from the definition of x_{k+1}, we have:

$$\langle \gamma_k G_k + \gamma_k gh(x_{k+1}) + \nabla d(x_{k+1}) - \nabla d(x_k), u - x_{k+1} \rangle \geq 0, \quad \forall u \in Q.$$

When rearranging terms, this inequality can be written as:

$$\begin{aligned} \gamma_k \langle G_k, x_k - u \rangle \quad \leq \quad & V(u, x_k) - V(u, x_{k+1}) + \gamma_k \langle G_k, x_k - x_{k+1} \rangle - V(x_{k+1}, x_k) \\ & + \gamma_k \langle gh(x_{k+1}), u - x_{k+1} \rangle. \end{aligned}$$

Denoting $d_k = \gamma_k \langle G_k, x_k - x_{k+1} \rangle - V(x_{k+1}, x_k)$, we obtain:

$$\begin{aligned} d_k \quad & \overset{(2.8)}{\leq} \quad \gamma_k \langle G_k, x_k - x_{k+1} \rangle - \frac{1}{2} \left\| x_k - x_{k+1} \right\|^2 \\ & = \quad \gamma_k [\langle g_k, x_k - x_{k+1} \rangle - \frac{L}{2} \left\| x_k - x_{k+1} \right\|^2] + \gamma_k [\langle G_k - g_k, x_k - x_{k+1} \rangle - \frac{\beta_k - L}{2} \left\| x_k - x_{k+1} \right\|^2] \\ & \overset{(2.2)}{\leq} \quad \gamma_k [f_k - f(x_{k+1}) + \delta] + \frac{\gamma_k}{\beta_k - L} \left\| G_k - g_k \right\|_*^2. \end{aligned}$$

where we use in the last inequality the fact that for all $g \in E^*, x \in E, \gamma > 0$:

$$\langle g, x \rangle - \frac{\zeta}{2} \left\| x \right\|^2 \leq \frac{1}{\zeta} \left\| g \right\|_*^2. \tag{3.1}$$

Therefore, we obtain:

$$\begin{aligned} \gamma_k \langle G_k, x_k - u \rangle \quad \leq \quad & V(u, x_k) - V(u, x_{k+1}) + \gamma_k [f_k - f(x_{k+1}) + \delta] + \frac{\gamma_k}{\beta_k - L} \left\| G_k - g_k \right\|_*^2 \\ & + \gamma_k \langle gh(x_{k+1}), u - x_{k+1} \rangle \\ \leq \quad & V(u, x_k) - V(u, x_{k+1}) + \gamma_k [f_k - f(x_{k+1}) + \delta] + \frac{\gamma_k}{\beta_k - L} \left\| G_k - g_k \right\|_*^2 \\ & + \gamma_k (h(u) - h(x_{k+1})). \end{aligned}$$

i.e:

$$\begin{aligned} \gamma_k [f(x_{k+1}) + h(x_{k+1})] \quad \leq \quad & V(u, x_k) - V(u, x_{k+1}) + \gamma_k [f_k + \langle g_k, u - x_k \rangle] + \gamma_k \langle G_k - g_k, u - x_k \rangle \\ & + \gamma_k \delta + \frac{\gamma_k}{\beta_k - L} \left\| G_k - g_k \right\|_*^2 + \gamma_k h(u) \\ \overset{(2.2)}{\leq} \quad & V(u, x_k) - V(u, x_{k+1}) + \gamma_k \phi(u) + \gamma_k [\langle G_k - g_k, u - x_k \rangle] \\ & + \gamma_k \delta + \frac{\gamma_k}{\beta_k - L} \left\| G_k - g_k \right\|_*^2. \end{aligned}$$

In particular, choosing $u = x^*$:

$$\begin{aligned}\gamma_k\phi(x_{k+1}) &\leq V(x^*,x_k) - V(x^*,x_{k+1}) + \gamma_k\phi^* + \gamma_k[\langle G_k - g_k, x^* - x_k\rangle] \\ &\quad +\gamma_k\delta + \frac{\gamma_k}{\beta_k - L}\|G_k - g_k\|_*^2.\end{aligned}$$

Summing these inequalities, we obtain:

$$\begin{aligned}\sum_{i=0}^{k-1}\gamma_i(\phi(x_{i+1}) - \phi^*) &\leq V(x^*,x_0) + \sum_{i=0}^{k-1}\gamma_i\langle G_i - g_i, x^* - x_i\rangle \\ &\quad + \sum_{i=0}^{k-1}\gamma_i\delta + \sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}\|G_i - g_i\|_*^2\end{aligned}$$

and therefore:

$$\begin{aligned}\phi(y_k) - \phi^* &\leq \frac{1}{\sum_{i=0}^{k-1}\gamma_i}\left(V(x^*,x_0) + \sum_{i=0}^{k-1}\gamma_i\langle G_i - g_i, x^* - x_i\rangle\right) \\ &\quad + \frac{1}{\sum_{i=0}^{k-1}\gamma_i}\sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}\|G_i - g_i\|_*^2 + \delta.\end{aligned}$$

□

Remark 5 We observe that the convergence rate does not depend on the difference $(F_{\delta,L}(x_i,\xi_i) - f_{\delta,L}(x_i))$. This is natural since the scheme itself does not use $F_{\delta,L}(x_i,\xi_i)$, the stochastic estimate of the function value. This property is shared by all methods considered in this paper.

Taking now the expectation with respect to the history of the random process $\xi_{[i]} = (\xi_0, ..., \xi_i)$, we obtain the following result:

Theorem 2 For all $k \geq 0$:

$$E_{\xi_1\sim X_1,\ldots\xi_k\sim X_k}[\phi(y_k) - \phi^*] \leq \frac{V(x^*,x_0)}{\sum_{i=0}^{k-1}\gamma_i} + \frac{1}{\sum_{i=0}^{k-1}\gamma_i}\sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}\sigma^2 + \delta.$$

Proof. As $E_{\xi_i\sim X_i}[G_i|\xi_{[i-1]}] = g_i$ and as x_i is a deterministic function of $(\xi_1, ..., \xi_{i-1})$, the expectation of $\langle G_i - g_i, x^* - x_i\rangle$, conditional on $\xi_{[i-1]} = (\xi_1, ..., \xi_{i-1})$, is zero. Therefore, we have $E_{\xi_1\sim X_1,\ldots,\xi_k\sim X_k}[\sum_{i=0}^{k-1}\gamma_i\langle G_i - g_i, x^* - x_i\rangle] = 0$. Furthermore, by assumption, $E_{\xi_i\sim X_i}[\|G_i - g_i\|_*^2|\xi_{[i-1]}] \leq \sigma^2$ and we obtain: $E_{\xi_1\sim X_1,\ldots,\xi_k\sim X_k}[\sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}\|G_i - g_i\|_*^2] \leq \sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}\sigma^2$. □

3.3 Choice of Stepsizes

3.3.1 Why do we need new stepsizes rules ?

In the deterministic smooth case (i.e. when the function f is smooth with a Lipschitz continuous gradient and the oracle is exact), the optimal stepsize (see [14]) is constant and equal to the inverse of the Lipschitz-constant of the gradient: $\gamma_i = \frac{1}{L}, \quad \forall i \geq 0$. If we keep this stepsizes rule in the stochastic case, we cannot apply Theorem 1 (that assumes $\gamma_i < \frac{1}{L}$) but with an easy modification in the proof of this theorem, we can obtain the following upper-bound:

$$\phi(y_k) - \phi^* \leq \frac{1}{\sum_{i=0}^{k-1}\gamma_i}\left(V(x^*,x_0) + \sum_{i=0}^{k-1}\gamma_i[\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x^* - x_{i+1}\rangle]\right) + \delta.$$

But as x_{i+1} depends on $G_{\delta,L}(x_i,\xi_i)$, we cannot say that $E[\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x^* - x_{i+1}\rangle|\xi_{[i-1]}] = 0$ but only that:

$$E[\langle G_{\delta,L}(x_i,\xi_i)-g_{\delta,L}(x_i), x^*-x_{i+1}\rangle|\xi_{[i-1]}] \leq \sqrt{E[\|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i)\|_*^2 |\xi_{[i-1]}]}D \leq \sigma D$$

where $D = \max_{x\in Q, y\in Q} \|x-y\|$ is the diameter of the feasible set. Therefore we obtain:

$$E[\phi(y_k) - \phi^*] \leq \frac{LR^2}{k} + \delta + \sigma D.$$

We see that with the classical stepsize policy, the effect of the stochastic noise does not decrease with the iterations. This is a behavior that we want to avoid, it would be preferable to obtain a method that could converge to the optimal value of our problem ϕ^* (or at least to $\phi^* + \delta$ in the biased case).

If we consider $\gamma_i = \frac{1}{CL}$ with $C > 1$, in this case we can apply the Theorems 1 and 2 and obtain:

$$E[\phi(y_k) - \phi^*] \leq \frac{CLR^2}{k} + \delta + \frac{\sigma^2}{(C-1)L}$$

but here also we obtain the same kind of behavior with a method that cannot decrease the stochastic noise effect when we increase the number of iterations. If we want to be able to converge to ϕ^* in the unbiased case or to $\phi^* + \delta$ in the biased case, a decresing stepsize policy must be used.

Remark 6 For non-smooth problems, the same kind of decreasing stepsize $\gamma_i = O\left(\frac{R}{M\sqrt{i}}\right)$ can be used both in the deterministic and the stochastic case. For smooth problem, the more aggressive constant stepsize $\gamma_i = O\left(\frac{1}{L}\right)$ (that leads to the improvement of the convergence rate in the deterministic case from $O\left(\frac{1}{\sqrt{k}}\right)$ to $O\left(\frac{1}{k}\right)$) is too large and not able to decrease the stochastic noise. In some sense, the gradient method is faster than the subgradient method but more sensible with respect to the stochastic error σ. When stochasticity is presents, we need to consider decreasing stepsize also in the smooth case (but decreasing only in term of σ not of L, i.e of the form $O\left(\frac{1}{L+\frac{\sigma}{R}\sqrt{i}}\right)$).

3.3.2 A new stepsize rule

By the complexity theory of first-order methods (see [10, 11, 7]), the best what we can expect in the stochastic case is a method that reduces the noise effect σ by a quantity $\Theta(\frac{\sigma R}{\sqrt{k}})$ after k iterations. This result gives us possibility to expect a better behavior for the SPGM that what we have obtained using the classical constant stepsize in the last section. In the same time, there is no hope to obtain a method with convergence rate $\Theta(\frac{LR^2+\sigma R}{k} + \delta)$. If we assume that the number of iterations N is known in advance, we can obtain the rate $\Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$ relatively easily. A constant stepsize (but that depends on the performed number of iterations) can be chosen. In [7], Lan has proposed the rule $\gamma_i = \min\left(\frac{1}{2L}, \sqrt{\frac{R^2}{2N\sigma^2}}\right), \quad \forall i \geq 0$ and obtained this desired rate of convergence. Another possible choice is $\gamma_i = \frac{1}{L+\frac{\sigma}{R}\sqrt{N}}$ that leads to

$$E[\phi(y_N) - \phi^*] \leq \frac{LR^2}{N} + \frac{2\sigma R}{\sqrt{N}} + \delta.$$

Remark 7 For a first-order method with convergence rate $O\left(\frac{LR^2}{k}\right)$ in the deterministic exact case, the effect of the deterministic bias δ cannot be better than an additional term δ (see [2]). Therefore, this convergence rate has an optimal dependance in δ and σ.

Remark 8 It is possible to obtain a better dependance in L using an accelerated method, like the Stochastic Fast Gradient Method (SFGM) (see section 6) with convergence rate $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}+k\delta\right)$ but, in this case, we pay this acceleration by a unavoidable worst dependance in δ.

However, the need of fixing in advance the number of iterations is not really a desirable property. Often in practice, we want to run a method for a given time and not for a given number of iterations. For this reason, it is interesting to develop a practical stepsizes rule which is not based on an a priori knowledge of the performed number of iterations and at the same times that keeps the convergence rate $\Theta(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}+\delta)$. This is not trivial. Indeed, contrarily to more sophisticated methods of the latter sections, there is few degrees of freedom in the SPGM: we have only one sequence of coefficients β_i $(=\frac{1}{\gamma_i})$. Consider the choice:

$$\gamma_i = \frac{L+\frac{\sigma}{2R}\sqrt{i+1}}{(L+\frac{\sigma}{R}\sqrt{i+1})^2}.$$

This stepsize decreases with rate $\Theta\left(\frac{1}{L+\frac{\sigma}{R}\sqrt{i}}\right)$ and we retrieve the optimal stepsize $\gamma_i=\frac{1}{L}$ in the deterministic case. We have for all $k\geq 1$:

$$\begin{aligned}
\sum_{i=0}^{k-1}\gamma_i &= \sum_{i=1}^{k}\frac{L+\frac{\sigma}{2R}\sqrt{i}}{(L+\frac{\sigma}{R}\sqrt{i})^2}\\
&\geq \int_1^{k+1}\frac{L+\frac{\sigma}{2R}\sqrt{x}}{(L+\frac{\sigma}{R}\sqrt{x})^2}dx = \left[\frac{x}{L+\frac{\sigma}{R}\sqrt{x}}\right]_1^{k+1} = \frac{Lk+\frac{\sigma}{R}(k+1-\sqrt{k+1})}{(L+\frac{\sigma}{R})(L+\frac{\sigma}{R}\sqrt{k+1})}\\
&\geq \frac{(2-\sqrt{2})\frac{\sigma}{R}k+Lk}{(L+\frac{\sigma}{R})(L+\frac{\sigma}{R}\sqrt{k+1})} \geq \frac{(2-\sqrt{2})k}{L+\frac{\sigma}{R}\sqrt{k+1}}
\end{aligned}$$

and therefore $\frac{1}{\sum_{i=0}^{k-1}\gamma_i}\leq\frac{L+\frac{\sigma}{R}\sqrt{k+1}}{(2-\sqrt{2})k}$. On the other hand, we have:

$$\frac{\gamma_i}{\beta_i-L} = \frac{(L+\frac{\sigma}{2R}\sqrt{i+1})^2}{(L+\frac{\sigma}{R}\sqrt{i+1})^2(\frac{\sigma^2}{R^2}(i+1)+\frac{3}{2}\frac{L\sigma}{R}\sqrt{i+1})}\leq\frac{1}{\frac{\sigma^2}{R^2}(i+1)+\frac{3}{2}\frac{L\sigma}{R}\sqrt{i+1}}\leq\frac{1}{\frac{\sigma^2}{R^2}(i+1)}.$$

and therefore $\sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i-L}\leq\frac{R^2}{\sigma^2}Har(k)$ where $Har(k)=\sum_{i=1}^{k}\frac{1}{i}\leq 1+\ln(k)$.

We obtain finally the convergence rate:

$$E[\phi(y_k)-\phi^*]\leq\frac{LR^2}{(2-\sqrt{2})k}(Har(k)+1)+\frac{\sigma\sqrt{k+1}R}{(2-\sqrt{2})k}(Har(k)+1)+\delta.$$

As $Har(k)\leq 1+\ln(k)$, we retrieve, up to a logarithmic factor, a rate of the form $\Theta\left(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}+\delta\right)$ but now using varying stepsizes that does not assume the knowledge of the performed number of iterations.

Remark 9 If we want to avoid the logarithmic factor in the convergence rate and if the set Q is bounded with diameter D, we can define the approximate solution as: $\overline{y}_N:\frac{1}{\sum_{i=N/2-1}^{N-1}\gamma_i}\sum_{i=N/2-1}^{N-1}\gamma_i x_{i+1}$ averaging only the last $\frac{N}{2}$ search points x_i (for simplicity, we assume here that N is even). In this case we obtain: $E[\phi(\overline{y}_N)-\phi^*]\leq\frac{2\sqrt{2}}{2-\sqrt{2}}\left(1+\frac{2}{N}+\ln(2)\right)\left(\frac{LD^2}{N}+\frac{\sigma D}{\sqrt{N}}\right)$. However this choice of averaging assumes the storage of all test points in memory, when N is not known a priori.

4 The machinery of estimate functions

The most recent and efficient first-order methods in deterministic smooth convex optimization are based on the machinery of estimate functions (see [14, 15, 16]).
The principle is to construct progressively:

1. A model $\Psi_k(x)$ of the function using typically all the previously accumulated first-order information,
2. A sequence of approximate solutions y_k (for which we obtain the convergence rate),

using two sequences of coefficients $\{\alpha_i\}_{i\geq 0}$ and $\{\beta_i\}_{i\geq 0}$, such that the two following inequalities are satisfied:

$$A_k\phi(y_k) \leq \Psi_k^* = \min_{x\in Q} \Psi_k(x) \text{ and } \Psi_k(x) \leq A_k\phi(x) + \beta_k d(x), \quad \forall x \in Q$$

where $A_k = \sum_{i=0}^k \alpha_i$ and $d(x)$ is the prox-function chosen in the setup.
The convergence rate depends directly on the two sequences of coefficients. Indeed, $A_k\phi(y_k) \leq \Psi_k^* \leq \Psi_k(x^*) \leq A_k\phi^* + \beta_k d(x^*)$, and therefore: $\phi(y_k) - \phi^* \leq \frac{\beta_k d(x^*)}{A_k}$.

Remark 10 In the deterministic case, β_k is often chosen equal to L, at least when this Lipschitz-constant of the gradient is known. We will see that this constant coefficient policy is not anymore the best choice in the stochastic case.

Remark 11 The fact that the model $\Psi_k(x)$ is based on all the previously accumulated first-order information during the k first steps of the scheme does not mean that we have to store all these datas in memory (like what is needed for classical bundle methods in non-smooth optimization). We have typically only to store and update a weighted sum of the accumulated gradients.

The methods based on this principle typically update different sequences of iterates:

- a sequence x_k, where we compute the first-order information,
- the sequence $v_k = \arg\min_{x\in Q} \Psi_k(x)$ of minimizers of the estimate functions $\Psi_k(x)$,
- a sequence of approximate solutions y_k, for which we obtain the convergence rate,
- sometimes, one or more additional sequences often obtained using gradient steps.

The easiest way to implement the idea of sequence of estimate functions is the Dual Gradient Method (DGM) introduced by Nesterov in [16]. In this method, x_k is exactly equal to v_k and the approximate solutions y_k are constructed using gradient steps from the points x_k. However the rate of convergence of this method is $O\left(\frac{LR^2}{k}\right)$, not better than using the classical gradient method. A more sophisticated implementation of this machinery leads to the Fast Gradient Method (FGM), developed by Nesterov in different versions [14, 15, 16] and that can reach the optimal convergence rate for deterministic smooth convex problems $O\left(\frac{LR^2}{k^2}\right)$.

In this paper, we generalize the concept of estimate functions sequence assuming now that the model of the function $\Psi_k(x)$ is constructed using stochastic first-order information (possibly with bias) and the sequences $\{y_k\}_{k\geq 0}$ and $\{\Psi_k(x)\}_{k\geq 0}$ satisfies the two inequalities:

$$A_k\phi(y_k) \leq \Psi_k^* + E_k \text{ and } \Psi_k(x) \leq A_k\phi(y_k) + \beta_k d(x) + \overline{E}_k(x), \quad \forall x \in Q$$

where E_k and $\overline{E}_k(x)$ represent random errors coming from the stochastic noise σ and the bias δ.
With this notion of stochastic estimate functions, we obtain the convergence rate:

$$\phi(y_k) - \phi^* \leq \frac{\beta_k d(x^*)}{A_k} + \frac{E_k + \overline{E}_k(x^*)}{A_k}$$

and therefore:

$$E[\phi(y_k) - \phi^*] \leq \frac{\beta_k d(x^*)}{A_k} + \frac{E[E_k + \overline{E}_k(x^*)]}{A_k}$$

since the coefficients $\{\alpha_i\}$ and $\{\beta_i\}$ are deterministic (they will be based on the noise level σ but not on the realizations of the random variables $X_1, ..., X_k$).

Using this framework, we will develop in Section 5 a stochastic dual gradient method with convergence rate $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$ and in Section 6, a stochastic fast gradient method with convergence rate $O\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}} + k\delta\right)$.

Remark 12 In the deterministic case, the model $\Psi_k(x)$ is typically chosen of the form: $\Psi_k(x) = \beta_k d(x) + \sum_{i=0}^k \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + h(x)]$. In the stochastic case, we will simply modify this model using the stochastic first-order information instead of the exact one: $\Psi_k(x) = \beta_k d(x) + \sum_{i=0}^k \alpha_i [F_{\delta,L}(x_i, \xi_i) + \langle G_{\delta,L}(x_i, \xi_i), x - x_i \rangle + h(x)]$.

Remark 13 Compared to classical gradient method, the methods based on this principle of estimate functions are more sophisticated and often less intuitive.

However, they provides us typically with more degrees of freedom (multiple sequences of coefficients, multiple sequences of iterates), that make these methods more flexible to new situations (we will see that adaptation of the DGM and FGM to the stochastic case is in some sense easier than for the PGM) and well-suited for acceleration (cfr the optimal rate of the Fast gradient method.)

5 Stochastic Dual Gradient Method

5.1 Scheme

In this method we use two sequences of coefficients:

$$\{\alpha_k\}_{k \geq 0} \text{ with } \alpha_0 \in]0, 1] \text{ and } \{\beta_k\}_{k \geq 0} \text{ with } \beta_{k+1} \geq \beta_k > L \quad \forall k \geq 0.$$

Furthermore the two sequences must satisfy the coupling condition:

$$\beta_k \geq \alpha_{k+1} \beta_{k+1}, \quad \forall k \geq 0. \tag{5.1}$$

We define also $A_k = \sum_{i=0}^k \alpha_i$.

Stochastic Dual Gradient Method (SDGM)

- **Initialization**
  1. Compute $x_0 = \arg\min_{x \in Q} d(x)$
  2. Let ξ_0 be a realization of the random variable X_0
  3. Compute $G_{\delta,L}(x_0, \xi_0)$
  4. Compute

$$w_0 = \arg\min_{x \in Q} \{\beta_0 d(x) + \alpha_0 \langle G_{\delta,L}(x_0, \xi_0), x - x_0 \rangle + \alpha_0 h(x)\} \tag{5.2}$$

- **Iteration** $k \geq 0$
  1. Compute

$$x_{k+1} = \arg\min_{x \in Q} \{\beta_k d(x) + \sum_{i=0}^k \alpha_i \langle G_{\delta,L}(x_i, \xi_i), x - x_i \rangle + A_k h(x)\} \tag{5.3}$$

2. Let ξ_{k+1} be a realization of the random variable X_{k+1}
3. Compute $G_{\delta,L}(x_{k+1}, \xi_{k+1})$
4. Compute

$$w_{k+1} = \arg\min_{x \in Q} \{\beta_{k+1} V(x, x_{k+1}) + \langle G_{\delta,L}(x_{k+1}), x - x_{k+1} \rangle + h(x)\} \quad (5.4)$$

- **Approximate Solution** $y_k = \frac{1}{\sum_{i=0}^k \alpha_i} \sum_{i=0}^k \alpha_i w_i$.

The dual gradient method has been introduced in [16] by Nesterov in the deterministic (composite) case and using the Euclidean setup.
We generalize this method in two directions:

- We generalize the method to the non-Euclidean setting, using auxiliary subproblems based only on the prox-function $d(x)$
- We adapt the method to the stochastic case (possible with bias). We will see that the classical choice $\beta_i = L$ is not anymore a good idea when stochasticity is present and we propose an increasing policy for the sequence $\{\beta_i\}$ that leads to a convergence rate of the form $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$ (or $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$ when the oracle is biased).

But first, we start with the general convergence rate of this stochastic dual gradient method:

5.2 General Convergence rate

Denote by $\Psi_k(x) = \beta_k d(x) + \sum_{i=0}^k \alpha_i [F_{\delta,L}(x_i, \xi_i) + \langle G_{\delta,L}(x_i, \xi_i), x - x_i \rangle + h(x)]$, our model of the objective function, $\Psi_k^* = \min_{x \in Q} \Psi_k(x)$ its minimal value on the feasible set and $\xi_{[k]} = (\xi_0, ..., \xi_k)$ the history of the random process after k iterations.

Let us show that the two sequences $\{y_k\}_{k \geq 0}$ and $\{\Psi_k(x)\}_{k \geq 0}$ define a sequence of estimate functions.

Lemma 1 For all $k \geq 0$, we have:

1.
$$A_k \phi(y_k) \leq \Psi_k^* + E_k \quad (5.5)$$
where $E_k = \sum_{i=0}^k \alpha_i \delta + \sum_{i=0}^k \alpha_i [f_{\delta,L}(x_i) - F_{\delta,L}(x_i, \xi_i)] + \sum_{i=0}^k \frac{\alpha_i}{\beta_i - L} \left\| G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i) \right\|_*^2$

2.
$$\Psi_k(x) \leq A_k \phi(x) + \beta_k d(x) + \overline{E}_k(x), \quad \forall x \in Q \quad (5.6)$$
where $\overline{E}_k(x) = \sum_{i=0}^k \alpha_i [F_{\delta,L}(x_i, \xi_i) - f_{\delta,L}(x_i) + \langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x - x_i \rangle]$

Proof. 1. First, we will show by recurrence that the inequality: $\sum_{i=0}^k \alpha_i \phi(w_i) \leq \Psi_k^* + E_k$ is satisfied for all $k \geq 0$.

- For $k = 0$, we have:

$$\begin{aligned} \phi(w_0) &\overset{(2.2)}{\leq} f_0 + \langle g_0, w_0 - x_0 \rangle + \frac{L}{2} \left\| w_0 - x_0 \right\|^2 + \delta + h(w_0) \\ &= F_0 + \langle G_0, w_0 - x_0 \rangle + \frac{\beta_0}{2} \left\| w_0 - x_0 \right\|^2 + \delta + h(w_0) \\ &\quad + (f_0 - F_0) + \langle g_0 - G_0, w_0 - x_0 \rangle - \frac{\beta_0 - L}{2} \left\| w_0 - x_0 \right\|^2 \end{aligned}$$

and we obtain, using 2.6, 3.1 and the fact $0 < \alpha_0 \leq 1$:

$$\begin{aligned}\alpha_0\phi(w_0) &\leq \alpha_0[F_0 + \langle G_0, w_0 - x_0\rangle + h(w_0)] + \beta_0 d(w_0) \\ &\quad + \frac{\alpha_0}{\beta_0 - L}\|G_0 - g_0\|_*^2 + \alpha_0(f_0 - F_0) \\ &\overset{(5.2)}{=} \Psi_0^* + \frac{\alpha_0}{\beta_0 - L}\|G_0 - g_0\|_*^2 + \alpha_0(f_0 - F_0).\end{aligned}$$

- Now assume that this inequality is satisfied for $k \geq 0$ i.e. that we have:

$$\sum_{i=0}^{k}\alpha_i\phi(w_i) \leq \Psi_k^* + E_k.$$

Then as $\beta_{k+1} \geq \beta_k$ and by definition of V we have:

$$\begin{aligned}\Psi_{k+1}^* &= \min_{x\in Q}\{\beta_{k+1}d(x) + \sum_{i=0}^{k+1}\alpha_i[F_i + \langle G_i, x - x_i\rangle + h(x)]\} \\ &\geq \min_{x\in Q}\{\beta_k V(x, x_{k+1}) + \beta_k d(x_{k+1}) + \beta_k\langle\nabla d(x_{k+1}), x - x_{k+1}\rangle \\ &\quad + \sum_{i=0}^{k+1}\alpha_i[F_i + \langle G_i, x - x_i\rangle]\} + A_{k+1}h(x).\end{aligned}$$

Let $gh(x_{k+1}) \in \partial h(x_{k+1})$, by optimality condition defining x_{k+1}:

$$\langle\beta_k\nabla d(x_{k+1}) + \sum_{i=0}^{k}\alpha_i G_i + A_k gh(x_{k+1}), x - x_{k+1}\rangle \geq 0, \quad \forall x \in Q$$

and therefore:

$$\begin{aligned}\Psi_{k+1}^* &\geq \beta_k d(x_{k+1}) + \sum_{i=0}^{k}\alpha_i[F_i + \langle G_i, x_{k+1} - x_i\rangle] \\ &\quad + \min_{x\in Q}\{\beta_k V(x, x_{k+1}) + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle]\} \\ &\quad + A_{k+1}h(x) + A_k\langle gh(x_{k+1}), x_{k+1} - x\rangle \\ &\geq \beta_k d(x_{k+1}) + \sum_{i=0}^{k}\alpha_i[F_i + \langle G_i, x_{k+1} - x_i\rangle] \\ &\quad + \min_{x\in Q}\{\beta_k V(x, x_{k+1}) + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle]\} \\ &\quad + A_k h(x_{k+1}) + \alpha_{k+1}h(x) \\ &= \Psi_k^* + \alpha_{k+1}\min_{x\in Q}\{F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x) + \frac{\beta_k}{\alpha_{k+1}}V(x, x_{k+1})\}.\end{aligned}$$

As $\frac{\beta_k}{\alpha_{k+1}} \geq \beta_{k+1}$, by definition of w_{k+1} and as $V(w_{k+1}, x_{k+1}) \geq \frac{1}{2}\|x_{k+1} - w_{k+1}\|^2$,

we obtain:

$$\begin{aligned}
\Psi^*_{k+1} &\overset{(5.1)}{\geq} \Psi^*_k + \alpha_{k+1}\min_{x\in Q}\{F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x) + \beta_{k+1}V(x, x_{k+1})\} \\
&\overset{(5.4)}{=} \Psi^*_k + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, w_{k+1} - x_{k+1}\rangle + h(w_{k+1}) + \beta_{k+1}V(w_{k+1}, x_{k+1})] \\
&\overset{(2.8)}{\geq} \Psi^*_k + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, w_{k+1} - x_{k+1}\rangle + h(w_{k+1}) + \frac{\beta_{k+1}}{2}\|x_{k+1} - w_{k+1}\|^2] \\
&= \Psi^*_k + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, w_{k+1} - x_{k+1}\rangle + h(w_{k+1}) + \frac{L}{2}\|w_{k+1} - x_{k+1}\|^2] \\
&\quad +\alpha_{k+1}[F_{k+1} - f_{k+1} + \langle G_{k+1} - g_{k+1}, w_{k+1} - x_{k+1}\rangle + \frac{\beta_{k+1} - L}{2}\|w_{k+1} - x_{k+1}\|^2] \\
&\overset{(2.2),(3.1)}{\geq} \Psi^*_k + \alpha_{k+1}(f(w_{k+1}) - \delta + h(w_{k+1})) \\
&\quad +\alpha_{k+1}[F_{k+1} - f_{k+1}] - \frac{\alpha_{k+1}}{\beta_{k+1} - L}\|G_{k+1} - g_{k+1}\|^2_* \\
&\geq \sum_{i=0}^{k+1}\alpha_i(f(w_i) + h(w_i)) - E_k - \alpha_{k+1}\delta \\
&\quad +\alpha_{k+1}[F_{k+1} - f_{k+1}] - \frac{\alpha_{k+1}}{\beta_{k+1} - L}\|G_{k+1} - g_{k+1}\|^2_*
\end{aligned}$$

and therefore: $\sum_{i=0}^{k+1}\alpha_i\phi(w_i) \leq \Psi^*_{k+1} + E_{k+1}$ where $E_{k+1} = E_k + \alpha_{k+1}\delta + \alpha_{k+1}[f_{k+1} - F_{k+1}] + \frac{\alpha_{k+1}}{\beta_{k+1}-L}\|G_{k+1} - g_{k+1}\|^2_*$.

We have proved that $\sum_{i=0}^{k}\alpha_i\phi(w_i) \leq \Psi^*_k + E_k$ and using the definition of $y_k = \frac{1}{\sum_{i=0}^{k}\alpha_i}\sum_{i=0}^{k}\alpha_i w_i$, $A_k = \sum_{i=0}^{k}\alpha_i$ and the convexity of ϕ, we obtain now: $A_k\phi(y_k) \leq \Psi^*_k + E_k$ for all $k \geq 0$.

2. On the other hand, for all $x \in Q$, we have also:

$$\begin{aligned}
\Psi_k(x) &\overset{(2.2)}{\leq} \beta_k d(x) + \sum_{i=0}^{k}\alpha_i(f(x) + h(x)) + \sum_{i=0}^{k}\alpha_i[F_i - f_i + \langle G_i - g_i, x - x_i\rangle] \\
&= \beta_k d(x) + A_k\phi(x) + \overline{E}_k(x).
\end{aligned}$$

□

As we have proved that we are in the framework of estimate functions, we can now obtain directly the convergence rate for the SDGM:

Theorem 3 For all $k \geq 0$, we have:

$$\phi(y_k) - \phi^* \leq \frac{\beta_k d(x^*)}{A_k} + \delta$$
$$+\frac{1}{A_k}\left(\sum_{i=0}^{k}\frac{\alpha_i}{\beta_i - L}\|G_{\delta,L}(x_i) - g_{\delta,L}(x_i)\|^2_* + \sum_{i=0}^{k}\alpha_i\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - x_i\rangle\right).$$

Taking now the expectation with respect to the random process history $\xi_{[k]}$, we obtain the following result:

Theorem 4 For all $k \geq 0$:

$$E_{\xi_0\sim X_0,\ldots\xi_k\sim X_k}[\phi(\hat{x}_k) - \phi^*] \leq \frac{\beta_k d(x^*)}{A_k} + \frac{1}{A_k}\sum_{i=0}^{k}\frac{\alpha_i}{\beta_i - L}\sigma^2 + \delta.$$

Proof. Completely similar to the proof of theorem 2. □

5.3 Choice of the Coefficients

In the deterministic smooth case, the coefficients of the dual gradient method developed in [16] are chosen constant: $\beta_i = L$ and $\alpha_i = 1$ for all $i \geq 0$.

If we keep these coefficients in the stochastic case, we cannot apply Theorem 3 (that assumes $\beta_i > L$) but with an easy modification in the proof of this theorem, we can obtain the following upper-bound:

$$\phi(y_k) - \phi^* \leq \frac{LR^2}{k} + \delta + \frac{1}{k}\sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - w_i \rangle.$$

As w_i depends on $G_{\delta,L}(x_i)$, we cannot say that $E[\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - w_i \rangle | \xi_{[i-1]}] = 0$ but only $E[\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - w_i \rangle | \xi_{[i-1]}] \leq \sqrt{E[\|G_{\delta,L}(x_i) - g_{\delta,L}(x_i)\|_*^2 | \xi_{[i-1]}]} D \leq \sigma D$ where $D = \max_{x \in Q, y \in Q} \|x - y\|$ is the diameter of the feasible set. Therefore we have:

$$E[\phi(y_k) - \phi^*] \leq \frac{LR^2}{k} + \delta + D\sigma.$$

We see that with the classical choice of the coefficients, the effect of the stochastic noise does not decrease with the iterations.
If we consider $\beta_i = CL$ with $C > 1$, in this case we can apply the Theorems 3 and 4 and obtain $E[\phi(y_k) - \phi^*] \leq \frac{CLR^2}{k} + \delta + \frac{\sigma^2}{(C-1)L}$ but here also we obtain the same kind of behavior with a method that cannot decrease the stochastic noise effect when we increase the number of iterations. If we want to be able to converge to ϕ^* in the unbiased case or to $\phi^* + \delta$ in the biased case, an increasing sequence of coefficients β_i must be used.

On the other hand, often in practice, we want to run a method for a given time and not for a given number of iterations. For this reason, it is interesting to develop a practical stepsizes rule for the stochastic dual gradient method which is not based on a a priori knowledge of the performed number of iterations and at the same times that can reach the convergence rate $\Theta(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta)$.
Consider the choice $\alpha_i = a$ with $0 < a \leq 1$ and $\beta_i = L + b\frac{\sigma}{R}(i+1)^c$.

We have:

- $$\frac{\beta_k R^2}{A_k} = \frac{LR^2}{a(k+1)} + \frac{b\sigma R}{a(k+1)^{1-c}}$$
- $$\frac{1}{A_k}\sum_{i=0}^{k} \frac{\alpha_i}{\beta_i - L}\sigma^2 = \frac{\sigma R}{(k+1)b}\sum_{i=1}^{k+1} i^{-c} \leq \frac{\sigma R}{(k+1)b}\int_0^{k+1} x^{-c}dx \leq \frac{\sigma R}{b(k+1)^c}.$$

We obtain therefore using the theorem 4 :

$$E[\phi(y_k) - \phi^*] \leq \frac{LR^2}{a(k+1)} + \frac{b\sigma R}{a(k+1)^{1-c}} + \frac{\sigma R}{b(k+1)^c}.$$

Optimizing the rate of convergence of the term depending on σ, we choose $c = \frac{1}{2}$. The optimal choice for c is clearly $1/2$ for which we obtain a convergence rate of the form $\Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$. For the choice of a and b, we need to ensure the condition (5.1) i.e. $(L + b\frac{\sigma}{R}(k+1)^{1/2}) \geq a(L + \frac{\sigma}{R}(k+2)^{1/2})$ for all $k \geq 0$. A sufficient condition is $a \leq \sqrt{\frac{k+1}{k+2}}$ for all $k \geq 0$ and we obtain the condition $a \leq \frac{1}{\sqrt{2}}$. We take $\alpha_i = a = \frac{1}{\sqrt{2}}$, for all $i \geq 0$ and therefore:

$$E[\phi(y_k) - \phi^*] \leq \frac{\sqrt{2}LR^2}{(k+1)} + \frac{\sqrt{2}b\sigma R}{\sqrt{k+1}} + \frac{\sigma R}{b\sqrt{k+1}} + \delta.$$

The optimal choice for b is $2^{-1/4}$ and we obtain:

Theorem 5 If the sequences $\{\alpha_i\}_{i\geq 0}$ and $\{\beta_i\}_{i\geq 0}$ are chosen for all $i \geq 0$ as $\alpha_i = \frac{1}{\sqrt{2}}$ and $\beta_i = L + \frac{\sigma}{2^{1/4}R}(i+1)^{1/2}$ then the sequence generated by the SDGM satisfies:

$$E[\phi(y_k) - \phi^*] \leq \frac{\sqrt{2}LR^2}{(k+1)} + \frac{2^{5/4}\sigma R}{\sqrt{k+1}} + \delta = \Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right).$$

6 Stochastic Fast Gradient Method

6.1 Scheme

In this method we use also two sequences of coefficients:

$$\{\alpha_i\}_{i\geq 0} \text{ with } \alpha_0 \in]0,1] \text{ and } \{\beta_i\}_{i\geq 0} \text{ with } \beta_{k+1} \geq \beta_k > L, \quad \forall k \geq 0.$$

But now the two sequences must satisfy another coupling condition:

$$\alpha_k^2 \beta_k \leq (\sum_{i=0}^{k} \alpha_i)\beta_{k-1}, \quad \forall k \geq 1. \tag{6.1}$$

We define also $A_k = \sum_{i=0}^{k} \alpha_i$ and $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$.
SFGM (Stochastic Fast Gradient Method):

- **Initialization**
  1. Compute $x_0 = \arg\min_{x\in Q} d(x)$
  2. Let ξ_0 be a realization of the random variable X_0
  3. Compute $G_{\delta,L}(x_0, \xi_0)$
  4. Compute

$$y_0 = \arg\min_{x\in Q}\{\beta_0 d(x) + \alpha_0\langle G_{\delta,L}(x_0,\xi_0), x - x_0\rangle + h(x)\} \tag{6.2}$$

- **Iteration** $k \geq 0$
  1. Compute

$$z_k = \arg\min_{x\in Q}\{\beta_k d(x) + \sum_{i=0}^{k} \alpha_i\langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle + A_k h(x)\} \tag{6.3}$$

  2.

$$x_{k+1} = \tau_k z_k + (1-\tau_k)y_k \tag{6.4}$$

  3. Let ξ_{k+1} be a realization of the random variable X_{k+1}
  4. Compute $G_{\delta,L}(x_{k+1}, \xi_{k+1})$
  5. Compute

$$\hat{x}_{k+1} = \arg\min_{x\in Q}\{\beta_k V(x, z_k) + \alpha_{k+1}\langle G_{\delta,L}(x_{k+1},\xi_{k+1}), x - z_k\rangle + \alpha_{k+1}h(x)\} \tag{6.5}$$

  6. Let

$$y_{k+1} = \tau_k \hat{x}_{k+1} + (1-\tau_k)y_k. \tag{6.6}$$

This method is a generalization to the stochastic case of one of the newest variants of the famous fast gradient methods or Nesterov optimal methods for smooth convex optimization (the methods that reach the optimal convergence rate $O\left(\frac{LR^2}{k^2}\right)$ in the deterministic case). This variant has been introduced in [15] by Nesterov. It is based on the machinery of estimates functions (providing a more flexible method) and can be used easily with a non-Euclidean setup since it is based only on subproblems in terms of the prox-function $d(x)$.

In this work, we adapt the fast gradient method to the stochastic case, develop a new practical policy for the sequences $\{\alpha_i\}$ and $\{\beta_i\}$ and prove that with this choice the method can reach the unimprovale rate of convergence $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$ in unbiased stochastic smooth convex optimization.

The optimal rate $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$ has been obtained for the first time by Lan in [7] using an accelerated version of the Mirror Descent SA method with fixed stepsize based on the performed number of iterations. However, our method based on the estimates sequence principle does not assume the a priori knowledge of the number of iterations and does not assume the boundnesses of the feasible set. Furthermore, our analysis consider also the composite case when we add to f an easy convex function $h(x)$ and the situation when the oracle is not only stochastic but also affected by a bias δ. We obtain a convergence rate of the form $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}+k\delta\right)$. There is a phenomenon of errors accumulation with rate $\Theta(k\delta)$. It has been established in [2] that this is in fact unavoidable for any fast first-order method that reach the optimal dependance with respect to L in the convergence rate (i.e. $O\left(\frac{LR^2}{k^2}\right)$).

6.2 General convergence rate

Denote by $\Psi_k(x)=\beta_k d(x)+\sum_{i=0}^k \alpha_i[F_{\delta,L}(x_i,\xi_i)+\langle G_{\delta,L}(x_i,\xi_i),x-x_i\rangle+h(x)]$, our model of the objective function, $\Psi_k^*=\min_{x\in Q}\Psi_k(x)$ its minimal value on the feasible set and $\xi_{[k]}=(\xi_0,...,\xi_k)$ the history of the random process after k iterations.

Let us show that $\{y_k\}_{k\geq 0}$ and $\{\Psi_k(x)\}_{k\geq 0}$ define a sequence of estimate functions.

Lemma 2 For all $k\geq 0$, we have:

1. $$A_k\phi(y_k)\leq \Psi_k^*+E_k \tag{6.7}$$
   where $E_k=\sum_{i=0}^k A_i\delta+\sum_{i=0}^k \frac{A_i}{(\beta_i-L)}\left\|G(x_i,\xi_i)-g_{\delta,L}(x_i)\right\|_*^2+\sum_{i=0}^k\alpha_i(f_{\delta,L}(x_i)-F_{\delta,L}(x_i,\xi_i))+\sum_{i=1}^k A_{i-1}\langle g_{\delta,L}(x_i)-G_{\delta,L}(x_i,\xi_i),x_i-y_{i-1}\rangle$.
2. $$\Psi_k(x)\leq A_k\phi(x)+\beta_k d(x)+\overline{E}_k(x),\quad \forall x\in Q \tag{6.8}$$
   where $\overline{E}_k(x)=\sum_{i=0}^k\alpha_i[F_{\delta,L}(x_i,\xi_i)-f_{\delta,L}(x_i)+\langle G_{\delta,L}(x_i,\xi_i)-g_{\delta,L}(x_i),x-x_i\rangle]$

Proof. 1. First, we want to prove by recurrence that the inequality $A_k\phi(y_k)\leq\Psi_k^*+E_k$ is satisfied for all $k\geq 0$.

- It is true for $k = 0$. Indeed:

$$\begin{aligned}
\Psi_0^* &\overset{(6.2)}{=} \beta_0 d(y_0) + \alpha_0[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0)] \\
&\overset{(2.6)}{\geq} \frac{\beta_0}{2}\|y_0 - x_0\|^2 + \alpha_0[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0)] \\
&\geq \alpha_0[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0) + \frac{\beta_0}{2}\|y_0 - x_0\|^2] \\
&= \alpha_0[f_0 + \langle g_0, y_0 - x_0\rangle + h(y_0) + \frac{L}{2}\|y_0 - x_0\|^2] \\
&\quad + \alpha_0[F_0 - f_0 + \langle G_0 - g_0, y_0 - x_0\rangle + \frac{\beta_0 - L}{2}\|y_0 - x_0\|^2] \\
&\overset{(2.2),(3.1)}{\geq} \alpha_0[f(y_0) + h(y_0) - \delta] + \alpha_0[F_0 - f_0] - \frac{\alpha_0}{\beta_0 - L}\|G_0 - g_0\|_*^2 .
\end{aligned}$$

- Assume that it is true for $k \geq 0$ i.e that we have $A_k\phi(y_k) \leq \Psi_k^* + E_k$.
  Let $gh(z_k) \in \partial h(z_k)$, by the optimality condition of the problem defining z_k:

$$\langle \beta_k \nabla d(z_k) + \sum_{i=0}^{k} G_i + A_k gh(z_k), x - z_k\rangle \geq 0, \quad \forall x \in Q.$$

  Therefore as $\beta_{k+1} \geq \beta_k$:

$$\begin{aligned}
\Psi_{k+1}(x) &= \beta_{k+1} d(x) + \sum_{i=0}^{k+1} \alpha_i[F_i + \langle G_i, x - x_i\rangle] + A_{k+1}h(x) \\
&\geq \beta_k V(x, z_k) + \beta_k d(z_k) + \beta_k\langle \nabla d(z_k), x - z_k\rangle \\
&\quad + \sum_{i=0}^{k+1} \alpha_i[F_i + \langle G_i, x - x_i\rangle] + A_{k+1}h(x) \\
&\geq \beta_k V(x, z_k) + \beta_k d(z_k) + \sum_{i=0}^{k} \alpha_i[F_i + \langle G_i, z_k - x_i\rangle] \\
&\quad + A_{k+1}h(x) + \langle A_k gh(z_k), z_k - x\rangle + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle].
\end{aligned}$$

  But as $A_{k+1}h(x) + \langle A_k gh(z_k), z_k - x\rangle \geq A_k h(z_k) + \alpha_{k+1}h(x)$, we have:

$$\begin{aligned}
\Psi_{k+1}(x) &\geq \beta_k d(z_k) + \sum_{i=0}^{k} \alpha_i[F_i + \langle G_i, z_k - x_i\rangle + h(z_k)] \\
&\quad + \beta_k V(x, z_k) + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)] \\
&\overset{(6.3)}{=} \Psi_k^* + \beta_k V(x, z_k) + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)].
\end{aligned}$$

On the other hand:

$$\begin{aligned}
& \Psi_k^* + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)] \\
\geq\ & A_k\phi(y_k) - E_k + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)] \\
\overset{(2.2)}{\geq}\ & A_k[f_{k+1} + \langle g_{k+1}, y_k - x_{k+1}\rangle] - E_k + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle] \\
& + A_k h(y_k) + \alpha_{k+1}h(x) \\
=\ & A_{k+1}F_{k+1} + \langle G_{k+1}, A_k(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1})\rangle - E_k \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
& + A_k h(y_k) + \alpha_{k+1}h(x) \\
\overset{(6.4)}{=}\ & A_{k+1}F_{k+1} + \alpha_{k+1}\langle G_{k+1}, x - z_k\rangle - E_k \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
& + A_k h(y_k) + \alpha_{k+1}h(x).
\end{aligned}$$

We obtain:

$$\begin{aligned}
\Psi_{k+1}^* \quad \geq\ & A_{k+1}F_{k+1} + \min_{x\in Q}\{\beta_k V(x, z_k) + \alpha_{k+1}\langle G_{k+1}, x - z_k\rangle + \alpha_{k+1}h(x)\} - E_k \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] + A_k h(y_k) \\
\overset{(6.5)}{=}\ & A_{k+1}F_{k+1} + \beta_k V(\hat{x}_{k+1}, z_k) + \alpha_{k+1}\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle + \alpha_{k+1}h(\hat{x}_{k+1}) - E_k \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] + A_k h(y_k) \\
\overset{(2.8)}{\geq}\ & A_{k+1}[F_{k+1} + \tau_k\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle + \frac{\beta_k}{2A_{k+1}}\left\|\hat{x}_{k+1} - z_k\right\|^2] - E_k \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
& + A_{k+1}[\tau_k h(\hat{x}_{k+1}) + (1 - \tau_k)h(y_k)] \\
\overset{(6.1),(6.6)}{\geq}\ & A_{k+1}[F_{k+1} + \tau_k\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle + \frac{\beta_{k+1}\tau_k^2}{2}\left\|\hat{x}_{k+1} - z_k\right\|^2] \\
& - E_k + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
& + A_{k+1}h(y_{k+1}) \\
\overset{(6.4),(6.6)}{\geq}\ & A_{k+1}[F_{k+1} + \langle G_{k+1}, y_{k+1} - x_{k+1}\rangle + \frac{\beta_{k+1}}{2}\left\|y_{k+1} - x_{k+1}\right\|^2] \\
& - E_k + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
& + A_{k+1}h(y_{k+1})
\end{aligned}$$

and therefore:

$$\begin{aligned}
\Psi_{k+1}^* \quad =\ & A_{k+1}[f_{k+1} + \langle g_{k+1}, y_{k+1} - x_{k+1}\rangle + \frac{L}{2}\left\|y_{k+1} - x_{k+1}\right\|^2] \\
& - E_k + \alpha_{k+1}[F_{k+1} - f_{k+1}] + A_k\langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle \\
& + A_{k+1}[\langle G_{k+1} - g_{k+1}, y_{k+1} - x_{k+1}\rangle + \frac{\beta_{k+1} - L}{2}\left\|y_{k+1} - x_{k+1}\right\|^2] \\
& + A_{k+1}h(y_{k+1}) \\
\overset{(2.2),(3.1)}{\geq}\ & A_{k+1}(f(y_{k+1}) + h(y_{k+1}) - \delta) - E_k + \alpha_{k+1}[F_{k+1} - f_{k+1}] \\
& + A_k\langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle - \frac{A_{k+1}}{\beta_{k+1} - L}\left\|G_{k+1} - g_{k+1}\right\|_*^2.
\end{aligned}$$

The inequality is therefore also satisfied for $k + 1$ and we have proved our recurrence.

2. Now let us prove that (6.8) is satisfied for all $x \in Q$ and $k \geq 0$. Indeed:

$$\begin{aligned}\Psi_k(x) &= \beta_k d(x) + \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, x - x_i \rangle] + \sum_{i=0}^{k} \alpha_i [F_i - f_i] + \sum_{i=0}^{k} \alpha_i [\langle G_i - g_i, x - x_i \rangle] \\ &\quad + A_k h(x) \\ &\overset{(2.2)}{\leq} \beta_k d(x) + A_k (f(x) + h(x)) + \sum_{i=0}^{k} \alpha_i [F_i - f_i] + \sum_{i=0}^{k} \alpha_i [\langle G_i - g, x - x_i \rangle].\end{aligned}$$

□

As we have proved that we are in the framework of estimate functions, we can now obtain directly the convergence rate for the SFGM:

Theorem 6 For all $k \geq 0$, we have:

$$\phi(y_k) - \phi^* \leq \frac{1}{A_k} \left(\beta_k d(x^*) + \sum_{i=0}^{k} A_i \delta + \sum_{i=0}^{k} \frac{A_i}{\beta_i - L} \left\| G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i) \right\|_*^2 \right)$$

$$+ \frac{1}{A_k} \left(\sum_{i=1}^{k} A_{i-1} \langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), y_{i-1} - x_i \rangle + \sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x^* - x_i \rangle \right).$$

Taking the expectation with respect to $\xi_{[k]}$, the history of the random process, we obtain the following result:

Theorem 7 For all $k \geq 0$, we have:

$$E_{\xi_1 \sim X_1, \ldots \xi_k \sim X_k} [\phi(y_k) - \phi^*] \leq \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^{k} A_i \delta}{A_k} + \frac{1}{A_k} \sum_{i=0}^{k} \frac{A_i}{\beta_i - L} \sigma^2.$$

Proof. Same proof that for the Theorem 2 but now using the fact that y_{i-1} is also a deterministic function of $\xi_{[i-1]}$. □

6.3 Choice of the Coefficients

In the deterministic smooth case, the coefficients of the fast gradient method are chosen as $\beta_i = L$ and $\alpha_i = \frac{i+1}{2}$ for all $i \geq 0$.

If we keep these coefficients in the stochastic case, we cannot apply the Theorem 6 (that assumes $\beta_i > L$) but with an easy modification in the proof of this theorem, we can simply replace the term: $\sum_{i=0}^{k} \frac{A_i}{\beta_i - L} \left\| G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i) \right\|_*^2$ by $\sum_{i=0}^{k} A_i \langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x_i - y_i \rangle$ in the upper-bound given by this theorem.

But as y_i depends on $G_{\delta,L}(x_i)$, we cannot say that $E[\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - y_i \rangle | \xi_{[i-1]}] = 0$ but only: $E[\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - y_i \rangle | \xi_{[i-1]}] \leq \sqrt{E[\left\| G_{\delta,L}(x_i) - g_{\delta,L}(x_i) \right\|_*^2 | \xi_{[i-1]}]} D \leq \sigma D$ where $D = \max_{x \in Q, y \in Q} \|x - y\|$ is the diameter of the feasible set. Therefore we have:

$$E[\phi(\hat{x}_k) - \phi^*] \leq \frac{4Ld(x^*)}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta + \frac{1}{3}(k+3)D\sigma.$$

We see that with the classical choice of the coefficients, the effect of the stochastic noise σ does not decrease with the iterations like what we want to obtain. But in fact, it does not even stay constant like what we have obtained for the SPGM and SDGM with classical coefficients. Here the situation is even worse, the effect of the noise is increasing with the number of iterations, there is a phenomenon of error accumulation. This higher sensitivity

of the fast gradient method with respect to the noise has been already observed in [2, 19] when the error is deterministic. In [2], it has been established that it is an intrinsic property of any fast first-order method with optimal convergence rate $\Theta\left(\frac{LR^2}{k^2}\right)$. In our case, it means that a dependence in the bias δ of the form $\Theta(k\delta)$ is unavoidable. However, concerning the stochastic noise σ, the situation is better, we can modify the sequence of coefficients β_i in order to avoid this increasing dependence in σ in the convergence rate. If we consider $\beta_i = CL$ with $C > 1$, we can apply Theorems 6, 7 and obtain:

$$E[\phi(\hat{y}_k) - \phi^*] \leq \frac{4CLR^2}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta + \frac{1}{3(C-1)L}(k+3)\sigma^2.$$

But we obtain the same kind of bad behavior with an accumulation of errors both for the stochastic part σ and the deterministic bias δ.

In this subsection, we want to develop a practical stepsizes rule for the stochastic fast gradient method which is not based on a a priori knowledge of the performed number of iterations and at the same times that can reach the convergence rate $\Theta(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}} + k\delta)$. Consider the choice $\alpha_i = \frac{i+1}{a}$ and $\beta_i = L + b\frac{\sigma}{R}(i+2)^c$. Then we have $A_k = \sum_{i=0}^k \alpha_i = \frac{1}{2a}(k+1)(k+2)$ and the condition 6.1 becomes: $\frac{(k+1)^2}{a^2}(L + \frac{\sigma}{R}b(k+2)^c) \leq \frac{(k+1)(k+2)}{2a}(L + \frac{\sigma}{R}b(k+1)^c)$. A sufficient condition is to have:

1. $\frac{(k+1)^2}{a^2} \leq \frac{(k+1)(k+2)}{2a}$ for all $k \geq 0$ i.e. $a \geq 2$
2. $\frac{(k+1)^2}{a^2}\frac{\sigma}{R}b(k+2)^c \leq \frac{(k+1)(k+2)}{2a}\frac{\sigma}{R}b(k+1)^c$ for all $k \geq 0$ i.e. $a \geq 2^c$.

Assuming that $c \geq 1$, we choose $a = 2^c$. Then the condition $\alpha_k^2\beta_k \leq A_k\beta_{k-1}$ is satisfied, independently of the precise choice of b and c. With the choice of the sequences $\alpha_i = \frac{i+1}{2^c}$ and $\beta_i = L + \frac{\sigma}{R}b(i+2)^c$, we obtain $\frac{\beta_k R^2}{A_k} = \frac{2^{c+1}(L+\frac{\sigma}{R}b(k+2)^c)R^2}{(k+1)(k+2)}$, $\frac{\sum_{i=0}^k A_i}{A_k} = \frac{1}{3}(k+3)\delta$ and

$$\begin{aligned}\frac{1}{A_k}\sum_{i=0}^k \frac{A_i}{\beta_i - L}\sigma^2 &= \frac{\sigma R}{(k+1)(k+2)b}\sum_{i=0}^k \frac{(i+1)}{(i+2)^{c-1}} \\ &\leq \frac{\sigma R}{(k+1)(k+2)b}\int_1 k+1(x+2)^{2-c}dx \leq \frac{\sigma R}{b(3-c)}\frac{(k+3)^{3-c}}{(k+1)(k+2)}.\end{aligned}$$

The bound given by Theorem 7 becomes as follows;

$$E(\phi(y_k) - \phi^*] \leq \frac{2^{c+1}LR^2}{(k+1)(k+2)} + \frac{2^{c+1}b\sigma R}{(k+1)(k+2)} + \frac{\sigma R}{b(3-c)}\frac{(k+3)^{3-c}}{(k+1)(k+2)}$$

Now if we choose $c = 3/2$, the two terms depending on b and c are of order $\Theta(\frac{\sigma R}{k^{1/2}})$ and we obtain:

$$E_{\xi_1 \sim X_1, \ldots \xi_k \sim X_k}[\phi(y_k) - \phi^*] \leq \frac{2^{5/2}LR^2}{(k+1)(k+2)} + \frac{(2^{5/2}b + \frac{2}{3b})(k+3)^{3/2}\sigma R}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta.$$

The optimal choice of b is $\frac{1}{2^{3/4}\sqrt{3}}$ and we obtain in this case the final result:

Theorem 8 If the sequences $\{\alpha_i\}_{i\geq 0}$ and $\{\beta_i\}_{i\geq 0}$ are chosen in the following way: $\alpha_i = \frac{i+1}{2\sqrt{2}}$ and $\beta_i = L + \frac{\sigma}{2^{3/4}\sqrt{3}R}(i+2)^{3/2}$ for all $i \geq 0$ then the sequence generated by the SFGM satisfies:

$$\begin{aligned}E[\phi(y_k) - \phi^*] &\leq \frac{2^{5/2}LR^2}{(k+1)(k+2)} + \frac{2^{11/4}(k+3)^{3/2}\sigma R}{\sqrt{3}(k+1)(k+2)} + \frac{1}{3}(k+3)\delta \\ &= \Theta\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}} + k\delta\right).\end{aligned}$$

Remark 14 Due to the higher sensitivity of the FGM with respect to the stochastic noise σ, we need to increase the sequences of coefficients β_i at a fast rate $\Theta(L + \frac{\sigma}{R} i^{3/2})$ in order to decrase the stochastic noise at an optimal rate $O\left(\frac{\sigma R}{\sqrt{k}}\right)$. For the DGM which is more robust with respect to the errors, the increase of β_i can be limited to the rate $\Theta(L + \frac{\sigma}{R} i^{1/2})$.

7 Probability of large deviation

In the previous sections, we have obtained for different stochastic first-order methods, an upper bound on the expected value of the non-optimality gap $\phi(y_k) - \phi^*$. Now we want also to obtain an upper bound on the probability of large deviation for the same gap. The approach presented in this section is strongly linked with has been done in [11] for the mirror descent SA method in the non-smooth stochastic case.
In this section, we need the following assumption:
Assumption H7

1. For all $x \in E$, the random variables X have the same distribution such that $X_0, ..., X_k$ can be seen as *i.i.d.* random variables.
2. The stochastic approximate gradient $G_{\delta,L}(x, \xi)$ satisfies the condition $E_{\xi \sim X}\left[\exp\left(\frac{\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*^2}{\sigma^2}\right)\right] \leq \exp(1), \quad \forall x \in Q$. Due to the Jensen inequality, this assumption is stronger that the assumption that we have done previously: $E_{\xi \sim X}[\|G_{\delta,L}(x, \xi) - g_{\delta,L}(x)\|_*^2] \leq \sigma^2, \quad \forall x \in Q$.
3. The set Q is bounded with diameter $D = \max_{x \in Q, y \in Q} \|x - y\|$.

First of all, we establish two lemmas that will be useful in order to derive probability of large deviations for different first-order methods.

Lemma 3 Let $\xi_0, ..., \xi_k$ be a sequence of realizations of the i.i.d. random variables $X_0, ..., X_k$ and let $\Delta_i = \Delta_i(\xi_{[i]})$ be a deterministic function of $\xi_{[i]}$ such that for all $i \geq 0$:

$$E[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right) |\xi_{[i-1]}] \leq \exp(1)$$

and $c_0, ..., c_k$ is a sequence of positive coefficients. Then we have for any $k \geq 0$ and any $\Omega \geq 0$:

$$Prob\left(\sum_{i=0}^k c_i \Delta_i^2 \geq (1 + \Omega) \sum_{i=0}^k c_i \sigma^2\right) \leq \exp(-\Omega).$$

Proof. Using the convexity of the exponent and the linearity of the expectation, we obtain:

$$\begin{aligned} E\left[\exp\left(\frac{\sum_{i=0}^k c_i \Delta_i^2}{\sum_{i=0}^k c_i \sigma^2}\right)\right] &\leq \frac{\sum_{i=0}^k c_i \sigma^2 E\left[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right)\right]}{\sum_{i=0}^k c_i \sigma^2} \\ &= \frac{\sum_{i=0}^k c_i \sigma^2 E_{\xi_0 \sim X_0, ..., \xi_i \sim X_i}\left[E_{\xi_i \sim X_i}\left[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right) |\xi_{[i-1]}\right]\right]}{\sum_{i=0}^k c_i \sigma^2} \\ &\leq \exp(1). \end{aligned}$$

Therefore by the Markov inequality, for any $\tilde{\Omega} > 0$ we obtain: $\text{Prob}\left(\exp\left(\frac{\sum_{i=0}^k c_i \Delta_i^2}{\sum_{i=0}^k c_i \sigma^2}\right) \geq \tilde{\Omega}\right) \leq \frac{\exp(1)}{\tilde{\Omega}}$. Equivalently for any $\Omega \in \mathbb{R}$, we obtain: $\text{Prob}\left(\exp\left(\frac{\sum_{i=0}^k c_i \Delta_i^2}{\sum_{i=0}^k c_i \sigma^2}\right) \geq \exp(1 + \Omega)\right) \leq \exp(-\Omega)$. □

Lemma 4 Let $\xi_0, ...\xi_0$ be a sequence of realizations of the i.i.d. random variables $X_0, ..., X_k$ and let Γ_k and η_k be deterministic functions of $\xi_{[k]}$ such that:

1. $E[\Gamma_i|\xi_{[i-1]}] = 0$
2. $|\Gamma_i| \leq c_i\eta_i$ where c_i is a positive deterministic constant
3. $E[\exp\left(\frac{\eta_i^2}{\sigma^2}\right)|\xi_{[i-1]}] \leq \exp(1)$.

Then $Prob\left(\sum_{i=0}^k \Gamma_i \geq \sqrt{3}\sqrt{\Omega}\sigma\sqrt{\sum_{i=0}^k c_i^2}\right) \leq \exp(-\Omega)$ for all $k \geq 0$ and all $\Omega \geq 0$.

Proof. This result is a particular case of Lemma 2 in [8]. □

Now we are able using these two lemmas to establish easily probability of large deviation for the SDGM and the SFGM.

7.1 Probability of large deviation for SDGM

In the SDGM, the non-optimality gap $\phi(y_k) - \phi^*$ can be bounded by the sum of three terms (see Theorem 3) :

1. $H_1(k) = \frac{1}{A_k}\beta_k d(x^*) + \delta$
2. $H_2(k, \xi_{[k]}) = \frac{1}{A_k}\sum_{i=0}^k \frac{\alpha_i}{\beta_i - L}\left\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\right\|_*^2$
3. $H_3(k, \xi_{[k]}) = \frac{1}{A_k}\sum_{i=0}^k \alpha_i\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - x_i\rangle$.

The first term is deterministic but the two others are random. Therefore in order to obtain a probability of large deviation for $\phi(y_k) - \phi^*$, a natural approach is to obtain probability of large deviation for $H_2(k, \xi_{[k]})$ and $H_3(k, \xi_{[k]})$ separatly.

For $H_2(k, \xi_{[k]})$, using the lemma 3 with $\Delta_i = \left\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\right\|_*$ and $c_i = \frac{\alpha_i}{A_k(\beta_i - L)}$, we obtain that for any $k \geq 0$ and for any $\Omega \geq 0$:

$$Prob\left(H_2(k, \xi_{[k]}) \geq \frac{1+\Omega}{A_k}\sum_{i=0}^k \frac{\alpha_i}{\beta_i - L}\sigma^2\right) \leq \exp(-\Omega).$$

For $H_3(k, \xi_{[k]})$, using the lemma 4 with $\Gamma_i = \frac{\alpha_i}{A_k}\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle$, $\eta_i = \left\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\right\|_*$ and $c_i = \frac{\alpha_i D}{A_k}$, we obtain that for any $k \geq 0$ and for any $\Omega \geq 0$:

$$Prob\left(H_3(k, \xi_{[k]}) \geq \frac{\sqrt{3}\sqrt{\Omega}D\sigma}{A_k}\sqrt{\sum_{i=0}^k \alpha_i^2}\right) \leq \exp(-\Omega).$$

In conclusion, we obtain the following probability of large deviation for the SDGM:

Theorem 9 If the assumption H7 is satisfied, then for all $k \geq 0$ and all $\Omega \geq 0$:

$$Prob\left(\phi(y_k) - \phi^* \geq \frac{\beta_k d(x^*)}{A_k} + \delta + \frac{(1+\Omega)}{A_k}\sum_{i=0}^k \frac{\alpha_i}{\beta_i - L}\sigma^2 + \frac{\sqrt{3\Omega}D\sigma}{A_k}\sqrt{\sum_{i=0}^k \alpha_i^2}\right)$$
$$\leq 2\exp(-\Omega).$$

Using in particular the optimal coefficients policy $\alpha_i = \frac{1}{\sqrt{2}}$ and $\beta_i = L + \frac{\sigma}{2^{1/4}R}(i+1)^{1/2}$ for all $i \geq 0$, we obtain that for all $k \geq 0$ and all $\Omega \geq 0$:

$$Prob\left(\phi(y_k) - \phi^* \geq \Gamma_0(k) + \Gamma_1(k) + \Gamma_2(k) + \Gamma_3(k)\right) \leq 2\exp(-\Omega)$$

where $\Gamma_0(k) = \frac{\sqrt{2}LR^2}{k+1}$, $\Gamma_1(k) = \delta$, $\Gamma_2(k) = \frac{2^{5/4}\sigma R}{\sqrt{k+1}}$ and $\Gamma_3(k) = \frac{2^{1/4}\Omega\sigma R}{\sqrt{k+1}} + \frac{\sqrt{3\Omega}D\sigma}{\sqrt{k}}$.

Remark 15 By theorem 5, we have $E[\phi(y_k) - \phi^*] \leq \Gamma_0(k) + \Gamma_1(k) + \Gamma_2(k)$ and $\Gamma_3(k)$ represents therefore the deviation from the expected non-optimality gap.

Therefore a sufficient condtion for ensuring $Prob(\phi(y_k) - \phi^* \geq \epsilon) \leq 1 - \gamma$ with $0 < \gamma < 1$, is to perform

$$k = \max\left(\frac{8LR^2}{\epsilon}, \frac{142\sigma^2R^2}{\epsilon^2}, \frac{36\sigma^2R^2}{\epsilon^2}\ln^2\left(\frac{2}{1-\gamma}\right), \frac{75\sigma^2D^2}{\epsilon^2}\ln\left(\frac{2}{1-\gamma}\right)\right)$$

iterations with $\delta \leq \frac{\epsilon}{5}$.

Remark 16 Exactly the same kind of analysis can be done for SPGM using Theorem 1 and Lemma 3 and 4. For this method, the probability of large deviation is given by :

$$Prob\left(\phi(y_k) - \phi^* \geq \frac{V(x^*, x_0)}{\sum_{i=0}^{k-1}\gamma_i} + \delta + \frac{(1+\Omega)}{\sum_{i=0}^{k-1}\gamma_i}\sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}\sigma^2 + \frac{\sqrt{3}\sqrt{\Omega}D\sigma}{\sum_{i=0}^{k-1}\gamma_i}\sqrt{\sum_{i=0}^{k-1}\gamma_i^2}\right)$$
$$\leq 2\exp(-\Omega)$$

for all $k \geq 0$ and all $\Omega \geq 0$.

7.2 Probability of large deviation for the SFGM

In Theorem 6, we have obtained that for the SFGM, the gap $\phi(y_k) - \phi^*$ can be bounded by the sum of four quantities:

1. $I_1(k) = \frac{1}{A_k}\left(\beta_k d(x^*) + \sum_{i=0}^k A_i\delta\right)$
2. $I_2(k, \xi_{[k]}) = \frac{1}{A_k}\sum_{i=0}^k \frac{A_i}{\beta_i - L}\left\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i, \xi_i)\right\|_*^2$
3. $I_3(k, \xi_{[k]}) = \frac{1}{A_k}\sum_{i=1}^k A_{i-1}\langle G_{\delta,L}(x_i, \xi) - g_{\delta,L}(x_i), y_{i-1} - x_i\rangle = \frac{1}{A_k}\sum_{i=1}^k \alpha_{i-1}\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), y_{i-1} - z_{i-1}\rangle$
4. $I_4(k, \xi_{[k]}) = \frac{1}{A_k}\sum_{i=0}^k \alpha_i\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle$.

The first term $I_1(k)$ is deterministic but the three others are random.

For $I_2(k, \xi_{[k]})$, we use Lemma 3 with $\Delta_i = \left\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\right\|_*$ and $c_i = \frac{A_i}{A_k(\beta_i - L)}$, we obtain:

$$Prob\left(I_2(k, \xi_{[k]}) \geq \frac{1+\Omega}{A_k}\sum_{i=0}^k \frac{A_i}{\beta_i - L}\sigma^2\right) \leq \exp(-\Omega)$$

for any $k \geq 0$ and for any $\Omega \geq 0$.

For $I_3(k, \xi_{[k]})$, using Lemma 4 (starting however the sum at $i = 1$ instead of $i = 0$) with $\Gamma_i = \frac{\alpha_{i-1}}{A_k}\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), y_{i-1} - z_{i-1}\rangle$, $\eta_i = \left\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\right\|_*$ and $c_i = \frac{\alpha_{i-1}D}{A_k}$, we obtain:

$$Prob\left(I_3(k, \xi_{[k]}) \geq \frac{\sqrt{3}\sqrt{\Omega}D\sigma}{A_k}\sqrt{\sum_{i=1}^k \alpha_{i-1}^2}\right) \leq \exp(-\Omega)$$

for any $k \geq 1$ and for any $\Omega \geq 0$.

For $I_4(k,\xi_{[k]})$, using Lemma 4 with $\Gamma_i = \frac{\alpha_i}{A_k}\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle$, $\eta_i = \|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i)\|_*$ and $c_i = \frac{\alpha_i D}{A_k}$, we obtain:

$$Prob\left(I_4(k,\xi_{[k]}) \geq \frac{\sqrt{3}\sqrt{\Omega}D\sigma}{A_k}\sqrt{\sum_{i=0}^{k}\alpha_i^2}\right) \leq \exp(-\Omega)$$

for any $k \geq 0$ and for any $\Omega \geq 0$.

In conclusion, we obtain the following probability of large deviation for the SFGM:

Theorem 10 Assume that assumption H6 is satisfied, then for all $k \geq 0$ and all $\Omega \geq 0$:

$$Prob\left(\phi(y_k) - \phi^* \geq \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^{k} A_i}{A_k}\delta + \frac{(1+\Omega)}{A_k}\sum_{i=0}^{k}\frac{A_i}{\beta_i - L}\sigma^2 + \frac{2\sqrt{3\Omega}D\sigma}{A_k}\sqrt{\sum_{i=0}^{k}\alpha_i^2}\right)$$

$$\leq 3\exp(-\Omega).$$

Using in particular the optimal coefficients policy i.e. $\alpha_i = \frac{i+1}{2\sqrt{2}}$ and $\beta_i = L + \frac{\sigma}{2^{3/4}\sqrt{3}R}(i+2)^{3/2}$ for all $i \geq 0$, we obtain:

$$\text{Prob } (\phi(y_k) - \phi^* > \Lambda_0(k) + \Lambda_1(k) + \Lambda_2(k) + \Lambda_3(k)) \leq 3\exp(-\Omega)$$

where $\Lambda_0(k) = \frac{2^{5/2}LR^2}{(k+1)(k+2)}$, $\Lambda_1(k) = \frac{k+3}{3}\delta$, $\Lambda_2(k) = \frac{2^{11/4}(k+3)^{3/2}\sigma R}{\sqrt{3}(k+1)(k+2)}$ and $\Lambda_3(k) = \frac{2^{7/4}\Omega\sigma R}{\sqrt{3}}\frac{(k+3)^{3/2}}{(k+1)(k+2)} + \frac{2\sqrt{\Omega}\sigma D}{\sqrt{3}}\sqrt{\frac{2k+3}{(k+1)(k+2)}}$.

Remark 17 By Theorem 8, we have $E[\phi(y_k) - \phi^*] \leq \Lambda_0(k) + \Lambda_1(k) + \Lambda_2(k)$ and $\Lambda_3(k)$ represents therefore the deviation from the expected non-optimality gap.

Therefore a sufficient condtion for ensuring $Prob(\phi(y_k) - \phi^* \geq \epsilon) \leq 1 - \gamma$ with $0 < \gamma < 1$, is to perform

$$k = \max\left(6\sqrt{\frac{LR^2}{\epsilon}}, \frac{671\sigma^2R^2}{\epsilon^2}, \frac{168\sigma^2R^2}{\epsilon^2}\ln^2\left(\frac{2}{1-\gamma}\right), \frac{17\sigma^2D^2}{\epsilon^2}\ln\left(\frac{2}{1-\gamma}\right)\right)$$

with $\delta \leq \frac{\epsilon}{5}$.

8 Postoptimization: Accuracy certificate

In this section, we do the following assumption:

Assumption H7

1. For all $x \in E$, the random variables X have the same distribution such that $X_0, ..., X_k$ can be seen as *i.i.d.* random variables.
2. $E_{\xi\sim X}\{\exp\left(\frac{|F_{\delta,L}(x,\xi) - f_{\delta,L}(x)|^2}{\sigma_F^2}\right)\} \leq \exp(1)$
3. $E_{\xi\sim X}\{\exp\left(\frac{\|G_{\delta,L}(x,\xi) - f_{\delta,L}(x)\|_*^2}{\sigma_G^2}\right)\} \leq \exp(1)$
4. We have a zero-order oracle for the function h that can compute $h(x)$ for all $x \in Q$.
5. The set Q is bounded with diameter $D = \max_{x\in Q, y\in Q}\|x - y\|$.

After running k iterations of one of the stochastic first-order methods, we obtain a feasible point $y_k \in Q$ for the optimization problem 2.1.
We have obtained in the previous sections, theoretical guarantee for the expected non optimality gap $\phi(y_k) - \phi^*$ and for the probability of large deviations of this gap from his expected value. However, we could be also interested to estimate the actual value of $\phi(y_k) - \phi^*$ since in practice the quality of y_k can be better that what is guaranted by worst-case oriented theoretical bounds. If we want to estimate $\phi(y_k) - \phi^*$:

1. We need to compute $\phi(y_k) = f(y_k) + h(y_k)$ or at least a stochastic estimate of $\phi(y_k)$
2. We need to compute a lower bound on ϕ^* or at least a random number Φ^* which is on average (and with small probability of large deviation) a lower bound on ϕ^*.

In the deterministic case:

1. We can compute $\phi(y_k) = f(y_k) + h(y_k)$ using the exact oracle
2. we can obtain a lower bound on ϕ^*, minimizing on Q, the sum of h with the linearization of f at y_k:

$$\phi^* \geq \min_{x \in Q} \{f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + h(x)\}.$$

In the stochastic case, $f(y_k)$ and $\nabla f(y_k)$ are typically unavailable (or to costly to compute) and we will try to use accurate estimates of these quantities using our stochastic oracle. We proceed as follow:

1. We generate N independent samples $\eta_1, ..., \eta_N$ from the random variable Y_k
2. We compute $F_{\delta,L}(y_k, \eta_1), ..., F_{\delta,L}(y_k, \eta_N)$ and $G_{\delta,L}(y_k, \eta_1), ..., G_{\delta,L}(y_k, \eta_N)$ using the stochastic oracle
3. In order to reduce the noise, we construct better estimates of $f(y_k)$ and $\nabla f(y_k)$ using averaging:

$$F_{\delta,L}(y_k, \eta_1, ..., \eta_n) = \frac{1}{N} \sum_{i=1}^{N} F_{\delta,L}(y_k, \eta_i) \text{ and } G_{\delta,L}(y_k, \eta_1, ..., \eta_n) = \frac{1}{N} \sum_{i=1}^{N} G_{\delta,L}(y_k, \eta_i).$$

Now we can obtain:

- **A good random estimate of** $\phi(y_k)$: $F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + h(y_k)$
  Indeed, we have:

  $$\phi(y_k) - \delta \leq E_{\eta_1 \sim Y_k, ... \eta_N \sim Y_k}[F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + h(y_k)] = f_{\delta,L}(y_k) + h(y_k) \leq \phi(y_k)$$

  and if we increase the number of samples N, we decrease the probability of deviation of $F_{\delta,L}(y_k, \eta_1, ..., \eta_n)$ from his expected value $f_{\delta,L}(y_k)$: $Prob\left(|F_{\delta,L}(y_k, \eta_1, ..., \eta_N) - f_{\delta,L}(y_k)| \geq K\right) \leq \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}K}{\sqrt{2}\sigma_F} - 1\right)^2\right)$ (using the Theorem 2.1 (ii) in [5]) and therefore:

  $$Prob\left(|F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + h(y_k) - \phi(y_k)| \geq K + \delta\right) \leq \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}K}{\sqrt{2}\sigma_F} - 1\right)^2\right).$$

- **An approximate lower bound for** ϕ^***:**

  $$\Phi^* = \min_{x \in Q} \{F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + \langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N), x - y_k \rangle + h(x)\}$$

  which on average, provides us with a lower bound on ϕ^*. The probability of deviation of Φ^* from being really a lower bound on ϕ^* decreases with the size of the sample. Indeed, we have:

Theorem 11 For all $\beta \geq 0$:

$$Prob(\phi^* \geq \Phi^* - \beta)$$
$$\geq 1 - \max\left(\exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F} - 1\right)^2\right), \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}D\sigma_G} - \sqrt{\kappa}\right)^2\right)\right)$$

where κ is the constant of regularity of $(E, \|.\|)$ (see [5]).

Proof. Applying the Theorem 2.1 (ii) in [5] to $F_{\delta,L}(y_k, \eta_1, ..., \eta_N)$, we obtain for all $\beta \geq 0$:

$$Prob\left(|F_{\delta,L}(y_k, \eta_1, ..., \eta_N) - f_{\delta,L}(y_k)| \geq \frac{\beta}{2}\right) \leq \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F} - 1\right)^2\right).$$

Applying the same theorem to $G_{\delta,L}(y_k, \eta_1, ..., \eta_N)$, we obtain for all $\beta \geq 0$:

$$Prob\left(\|G_{\delta,L}(y_k, \eta_1, ..., \eta_N) - g_{\delta,L}(y_k)\|_* \geq \frac{\beta}{2}\right)$$
$$\leq \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_G} - \sqrt{\kappa}\right)^2\right).$$

Now as

$$f(x) \geq f_{\delta,L}(y_k) + \langle g_{\delta,L}(y_k), x - y_k\rangle, \quad \forall x \in Q$$

we have:

$$\begin{aligned}
& Prob\left(\exists x \in Q : \phi(x) \leq F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + \langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N), x - y_k\rangle + h(x) - \beta\right) \\
= \;& Prob\left(\exists x \in Q : f(x) \leq F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + \langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N), x - y_k\rangle - \beta\right) \\
\leq \;& Prob\left(\exists x \in Q : F_{\delta,L}(y_k, \eta_1, ..., \eta_N) - f_{\delta,L}(y_k) + \langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N) - g_{\delta,L}(y_k), x - y_k\rangle \geq \beta\right) \\
\leq \;& \max(Prob\left(F_{\delta,L}(y_k, \eta_1, ..., \eta_N) - f_{\delta,L}(y_k) \geq \frac{\beta}{2}\right), \\
& Prob\left(\|G_{\delta,L}(y_k, \eta_1, ..., \eta_N) - g_{\delta,L}(y_k)\|_* \geq \frac{\beta}{2D}\right)) \\
\leq \;& \max\left(\exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F} - 1\right)^2\right), \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}D\sigma_G} - \sqrt{\kappa}\right)^2\right)\right),
\end{aligned}$$

and therefore:

$$Prob\left(\forall x \in Q : \phi(x) \geq F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + \langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N), x - y_k\rangle + h(x) - \beta\right)$$
$$\geq 1 - \max\left(\exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F} - 1\right)^2\right), \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}D\sigma_G} - \sqrt{\kappa}\right)^2\right)\right).$$

In particular, we have:

$$\begin{aligned}
& Prob\left(\phi^* \geq \Phi^* - \beta\right) \\
= \;& Prob\left(\min_{x\in Q} \phi(x) \geq \min_{x\in Q}\{F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + \langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N), x - y_k\rangle + h(x)\} - \beta\right) \\
\geq \;& 1 - \max\left(\exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F} - 1\right)^2\right), \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}D\sigma_G} - \sqrt{\kappa}\right)^2\right)\right).
\end{aligned}$$

□

Remark 18 When $2 \leq p \leq +\infty$, the constant of regularity of $(\mathbb{R}^n, \|.\|_p)$ satisfies:

$$\kappa \leq \min(p-1, 2\ln(n)).$$

The regularity constants of various other normed spaces can be found in [5].

9 Numerical Experiments: Quadratic Problem with Stochastic noise

In this section, we want to test the methods developed in this paper (and to compare it with existing methods) on convex quadratic problems over the simplex:

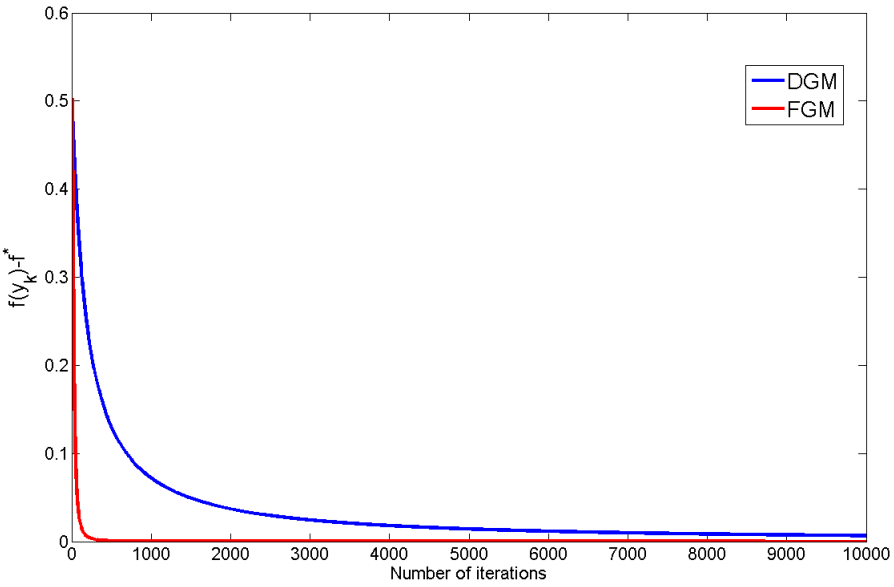
$$f^* = \min_{x \in \Delta_n} f(x) = \frac{1}{2} x^T A x \tag{9.1}$$

where $A \succeq 0$ and the l_1 setup is used.

Remark 19 As SPGM and SDGM share the same theoretical behavior and as the numerical results obtained using both methods are comparable, we do not consider in this section SPGM but only the methods that are really new in the stochastic context i.e. SDGM and SFGM.

In the exact case i.e. when the exact gradient $\nabla f(x) = Ax$ is available, the Fast Gradient Method (used with exact gradients and constant coefficients $\beta_i = L = \|A\|_\infty$) ouperforms significantly the Dual Gradient Method (used with exact gradient and constant coefficients $\beta_i = L = \|A\|_\infty$). Performing 10 000 iterations, we obtain for $f(y_k) - f^*$:

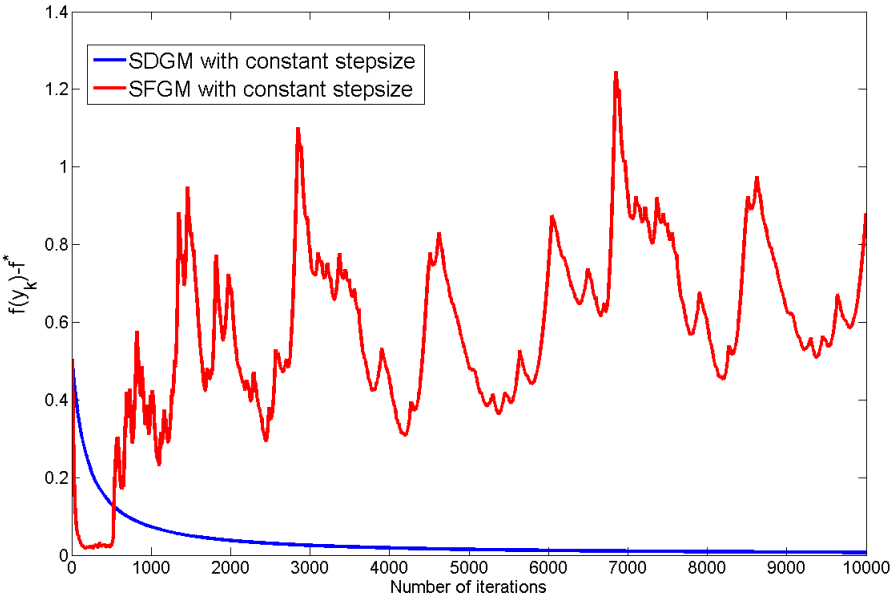
Num. Iter.	10	100	1000	10000
DGM	0.478796	0.329690	0.0720594	0.0066759
FGM	0.427691	0.0233784	3.6576e-4	8.3417e-6



This result is completely expected by the theory, the FGM exhibits a convergence rate of the form $\Theta\left(\frac{LR^2}{k^2}\right)$, significantly better than $\Theta\left(\frac{LR^2}{k}\right)$ for the DGM.

Now assume that we have only access to a stochastic gradient $G_{\delta,L}(x,\xi) = Ax + \xi$ where ξ is a stochastic noise (with normal distribution) such that $E[\xi] = 0$ and $E[\|\xi\|_*^2] \leq \sigma^2$. We consider first a reasonable noise level $\sigma = 1$. We can try to apply the SDGM and the SFGM with constant coefficients $\beta_i = L$ like what we do in the exact case. This choice is not recommended by the theory since the SDGM exhibits in this case a rate $\Theta\left(\frac{LR^2}{k} + \sigma D\right)$ and the SFGM $\Theta\left(\frac{LR^2}{k^2} + kD\sigma\right)$. Performing 10000 iterations, we obtain:

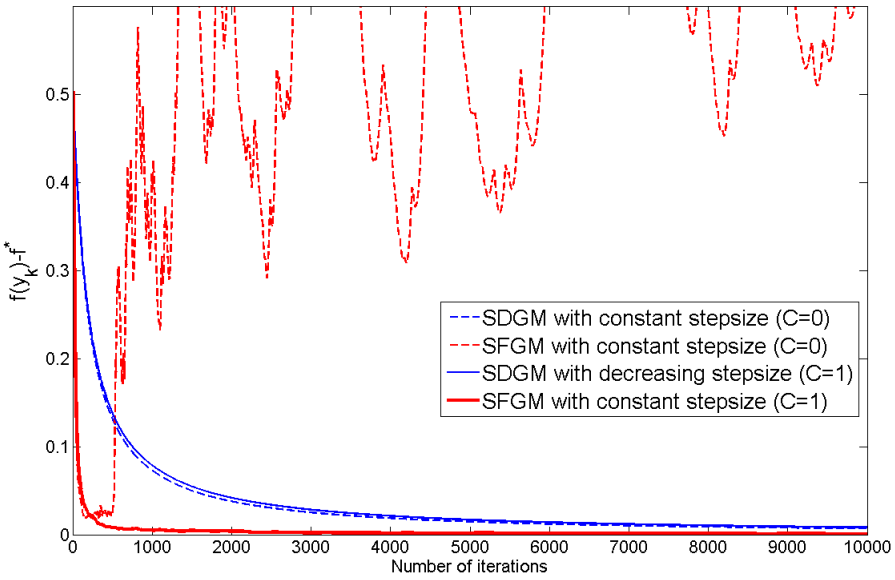
Num. Iter.	10	100	1000	10000
SDGM (C=0)	0.481479	0.335265	0.0728529	0.00698266
SFGM (C=0)	0.428563	0.0385569	0.399419	0.881574



SDGM exhibits here a slow but convergent behavior. However, we see that SFGM is unstable and suffers from accumulation of errors. This bad behavior of the SFGM when used with constant stepsize ($\gamma_i = \frac{1}{L}$) and a stochastic oracle has been predicted by the theory. The SDGM is slow but more robust to the errors, the method is still convergent even with this aggressive constant stepsize policy.

In order to avoid this sensitivity to the stochastic noise σ, we use now the decreasing stepsize policies developed in this paper i.e. the increasing sequence of coefficients: $\beta_i = L + \frac{C\sigma}{2^{1/4}R}(i+1)^{1/2}$ for the SDGM and $\beta_i = L + \frac{C\sigma}{2^{3/4}\sqrt{3}R}(i+2)^{3/2}$ for the SFGM. When $C = 0$, we retrieve the constant stepsize policy and $C = 1$ corresponds to the theoretical optimal choice.
With the theoretical optimal choice C=1, we obtain:

Num. Iter.	10	100	1000	10000
SDGM (C=1)	0.481753	0.339013	0.0786420	0.00822247
SFGM (C=1)	0.431472	0.0531080	0.00491995	7.851197e-4



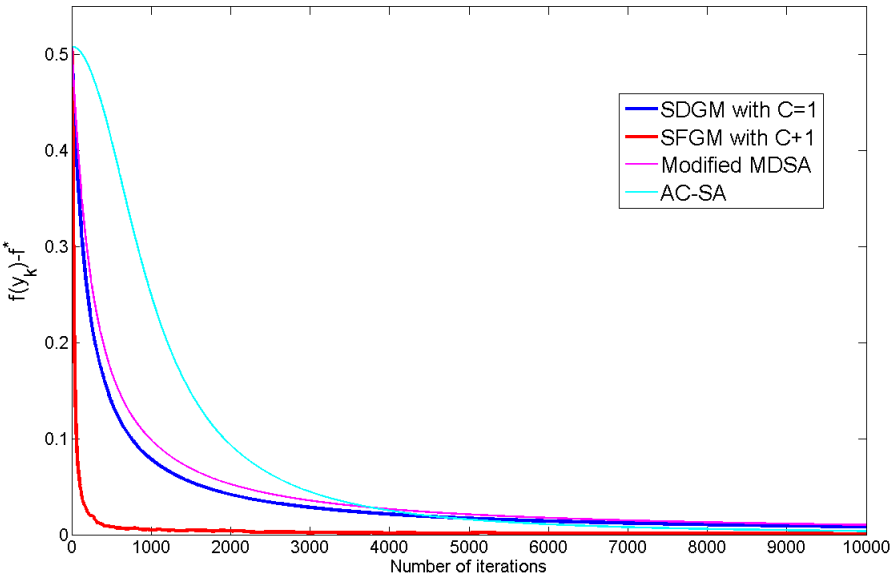
The SFGM retrieves his good behavior, the method is significantly faster than the SDGM and can decrease now the effect of the oracle noise (instead of increasing it with constant stepsizes). We see here clearly the importance of using decreasing stepsizes in the stochastic case (at least for the fast-gradient method). For the SDGM, for this level of noise, a decreasing sequence of stepsize seems not necessary and slow down a little bit the convergence.

We can also compare our methods (SDGM and SFGM) with the methods developed by Lan in [7]:

- The Modified Mirror Descent SA (MMDSA) method with convergence rate $\Theta\left(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}\right)$ (like what we obtain for the SDGM when used with C=1)
- The Accelerated SA (AC-SA) method with convergence rate $\Theta\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$ (like what we obtain for the SFGM when used with C=1).

An important property of the methods developed by Lan is the fact that they are based on the a priori knowledge of the performed number of iterations N. The goal of these methods is to reach a good accuracy after N iterations, not for intermediate $0 < k < N$. Performing 10000 iterations of our two methods and the two methods developed by Lan, we obtain:

Num. Iter.	10	100	1000	10000
SDGM (C=1)	0.481753	0.339013	0.0786420	0.00822247
SFGM (C=1)	0.431472	0.0531080	0.00491995	7.851197e-4
MMDSA	0.491474	0.376019	0.0986267	0.0100789
AC-SA	0.508434	0.503937	0.249861	0.00365878

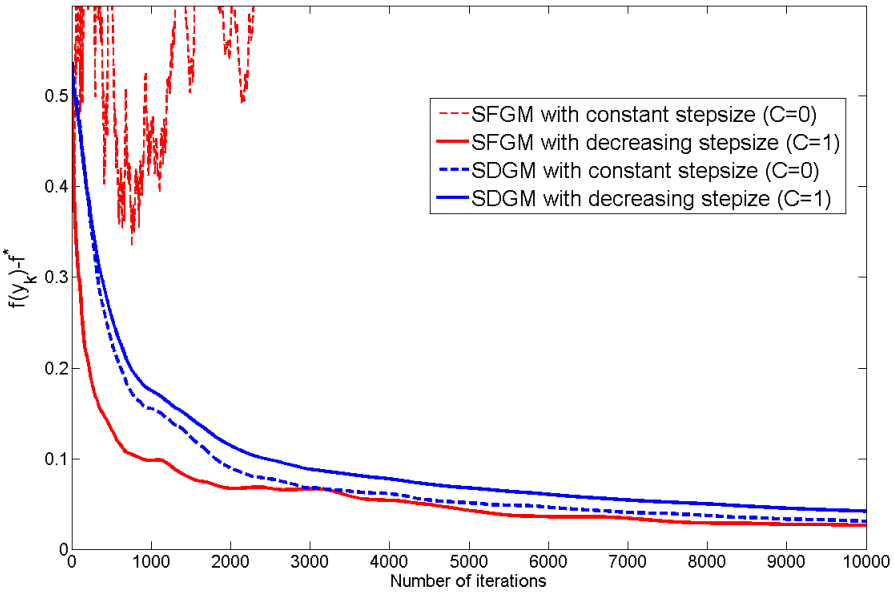


For the gradient-type methods (i.e. the SDGM and the MMDSA method), the two methods exhibits the same kind of behavior with however a faster convergence for our SDGM. For the fast-gradient-type methods (i.e. the SFGM and the AC-SA method), the AC-SA is only efficient if we perform really N iterations, not for an intermediate number of iterations whereas the SFGM is fast everywhere. We see here clearly the advantage of methods that are not based on a fixed number of iterations.

In conclusion, when the stochastic noise is reasonable (here 1 % of the Lipschitz-constant of the gradient), the SFGM with decreasing stepsize seems to be the method of choice. This method is fast (compare to SDGM and MMDSA method), is not sensitive to the oracle error (compare to SFGM with constant stepsize) and is flexible, does not need to perform exactly an a priori fixed number of iterations (which is the case for the AC-SA method).

We consider now the situation when the noise σ is significantly more important: $\sigma = 10$. First, we compare the SDGM with the SFGM, both using constant or decreasing stepsizes:

Num. Iter.	10	100	1000	10000
SDGM (C=0)	0.526044	0.467577	0.155076	0.030702
SDGM (C=1)	0.523209	0.463160	0.1751157	0.0419122
SFGM (C=0)	0.48812	0.5741252	0.446167	0.975812
SFGM (C=1)	0.462503	0.292540	0.097984	0.026385

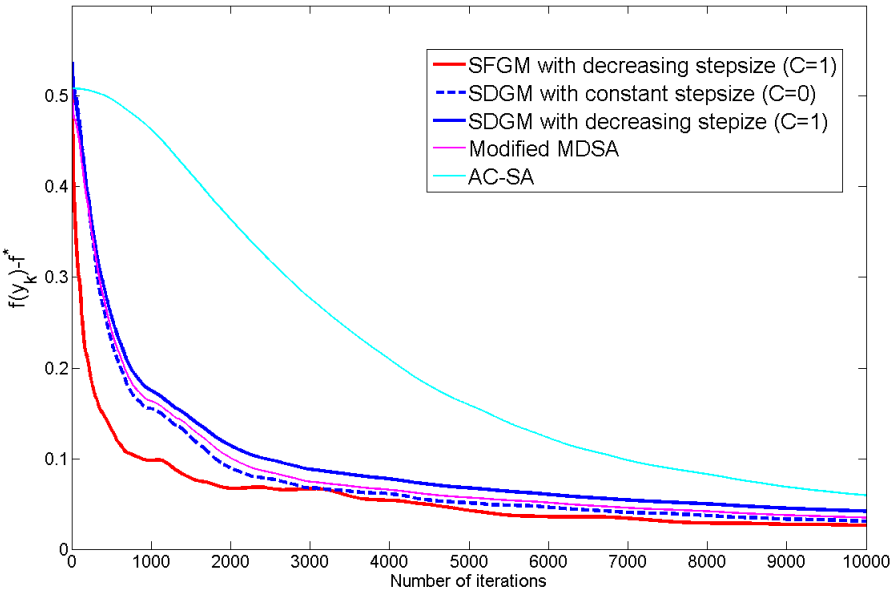


We observe that:

- The SFGM must be used with decreasing stepsizes in order to avoid a bad accumulation errors. This phenomenon has been already observed for $\sigma = 1$.
- The SFGM with decreasing stepsize is a little bit faster than the SDGM with decreasing stepsize. However the advantage of the SFGM is significantly reduced compare to the case $\sigma = 1$. This is natural, when the noise is large, the advantage of a convergence rate $O\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}}\right)$ over $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$ becomes negligible, the dominant term in the convergence rate becomes quickly the bad term coming from the noise.
- The SDGM can be used with constant stepsize and this more aggressive choice gives a faster convergence. It seems that the robustness of the SDGM (more important than expected by the theory) is sufficient in order to avoid a decreasing stepsize even when $\sigma = 10$. The worst-case oriented decreasing stepsize policy seems to slow down the method unnecessarily on this numerical example.

Now we can compare also our methods with the methods developed by Lan on this noisy example:

Num. Iter.	10	100	1000	10000
SDGM (C=0)	0.526044	0.467577	0.155076	0.030702
SDGM (C=1)	0.523209	0.463160	0.1751157	0.0419122
SFGM (C=0)	0.48812	0.5741252	0.446167	0.975812
SFGM (C=1)	0.462503	0.292540	0.097984	0.026385
MMDSA	0.494363	0.4463241	0.1633532	0.034726
AC-SA	0.508496	0.508166	0.4631923	0.0593871



We observe that:

- The AC-SA method performs badly on this example. This method is very slow at the begining (the method being designed only to reach a good accuracy after the fixed number of iterations N) and even ater the N iterations, the obtained solution is not so accurate. The SFGM with decreasing stepsize that share the same convergence rate $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$ is clearly a better choice.
- The MMDSA method of Lan performs well on this noisy example but the best choice for a gradient type method seems to be the SDGM with aggressive constant stepsize.

References

[1] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* , **Vol. 31, No. 3**, 167-175 (2009).

[2] O. Devolder, F. Glineur and Yu. Nesterov. First-order Methods of Smooth Convex Optimization with Inexact Oracle. *submitted to Mathematical programming, Serie A*, (2011).

[3] C. Hu, J.T. Kwok and W. Pan.. Accelerated Gradient Methods for Stochastic Optimization and Online Learning. *Neural Information Processing Systems (NIPS), Vancouver, Canada*, (2009).

[4] A. Juditsky, K. Karzan and A. Nemirovski. l_1 minimization via randomized first order algorithms. *submitted to Mathematical programming, Serie A*, (2010).

[5] A. Juditsky and A. Nemirovski. Large Deviations of Vector-Valued Martingales in 2-smooth Normed Spaces. *Submitted to the Annals of Probability.*, (2008).

[6] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, I: General purpose methods *To appear in: S. Sra, S. Nowozin, S. Wright, Eds., Optimization for Machine Learning, The MIT Press*, (2011).

[7] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming Serie A, Online First* (2010)

[8] G. Lan, A. Nemirovski and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming Serie A, Online First* (2011)

[9] Q. Lin, X. Chen and J. Pena. A Smoothing Stochastic Gradient Method for Composite Optimization. *Manuscript: arXiv:1008.5204v2*, (2010).

[10] A. Nemirovski and D. Yudin.Problem complexity and method efficiency in optimization. *John Wiley* (1983)

[11] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *Siam Journal of Optimization*, **Vol. 19, No. 4**, 1574-1609 (2009).

[12] Yu. Nesterov. A method for unconstrained convex minimization with the rate of convergence of $O(\frac{1}{k^2})$, *Doklady AN SSSR*, **269**, 543-547 (1983).

[13] Yu. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex function, *Èkonom. i. Mat. Metody (In Russian)*, **24**, 509-517 (1988).

[14] Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers (2004)

[15] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming, Serie A*, **103**, 127-152 (2005).

[16] Yu. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper*, **76**, (2007)

[17] A. Shapiro, D. Dentcheva and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM Series on Optimization, (2009)

[18] N.Z. Shor. Minimization Methods for Non-Differentiable Functions. *Springer Series in Computational Mathematics. Springer-Verlag* (1985).

[19] M. Schmidt, N. Le Roux and F. Bach. Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization. INRIA, Preprint, (2011).

Recent titles

CORE Discussion Papers

2011/27. David DE LA CROIX and Axel GOSSERIES. The natalist bias of pollution control.
2011/28. Olivier DURAND-LASSERVE, Axel PIERRU and Yves SMEERS. Effects of the uncertainty about global economic recovery on energy transition and CO_2 price.
2011/29. Ana MAULEON, Elena MOLIS, Vincent J. VANNETELBOSCH and Wouter VERGOTE. Absolutely stable roommate problems.
2011/30. Nicolas GILLIS and François GLINEUR. Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization.
2011/31. Nguyen Thang DAO and Julio DAVILA. Implementing steady state efficiency in overlapping generations economies with environmental externalities.
2011/32. Paul BELLEFLAMME, Thomas LAMBERT and Armin SCHWIENBACHER. Crowdfunding: tapping the right crowd.
2011/33. Pierre PESTIEAU and Gregory PONTHIERE. Optimal fertility along the lifecycle.
2011/34. Joachim GAHUNGU and Yves SMEERS. Optimal time to invest when the price processes are geometric Brownian motions. A tentative based on smooth fit.
2011/35. Joachim GAHUNGU and Yves SMEERS. Sufficient and necessary conditions for perpetual multi-assets exchange options.
2011/36. Miguel A.G. BELMONTE, Gary KOOP and Dimitris KOROBILIS. Hierarchical shrinkage in time-varying parameter models.
2011/37. Quentin BOTTON, Bernard FORTZ, Luis GOUVEIA and Michael POSS. Benders decomposition for the hop-constrained survivable network design problem.
2011/38. J. Peter NEARY and Joe THARAKAN. International trade with endogenous mode of competition in general equilibrium.
2011/39. Jean-François CAULIER, Ana MAULEON, Jose J. SEMPERE-MONERRIS and Vincent VANNETELBOSCH. Stable and efficient coalitional networks.
2011/40. Pierre M. PICARD and Tim WORRALL. Sustainable migration policies.
2011/41. Sébastien VAN BELLEGEM. Locally stationary volatility modelling.
2011/42. Dimitri PAOLINI, Pasquale PISTONE, Giuseppe PULINA and Martin ZAGLER. Tax treaties and the allocation of taxing rights with developing countries.
2011/43. Marc FLEURBAEY and Erik SCHOKKAERT. Behavioral fair social choice.
2011/44. Joachim GAHUNGU and Yves SMEERS. A real options model for electricity capacity expansion.
2011/45. Marie-Louise LEROUX and Pierre PESTIEAU. Social security and family support.
2011/46. Chiara CANTA. Efficiency, access and the mixed delivery of health care services.
2011/47. Jean J. GABSZEWICZ, Salome GVETADZE and Skerdilajda ZANAJ. Migrations, public goods and taxes.
2011/48. Jean J. GABSZEWICZ and Joana RESENDE. Credence goods and product differentiation.
2011/49. Jean J. GABSZEWICZ, Tanguy VAN YPERSELE and Skerdilajda ZANAJ. Does the seller of a house facing a large number of buyers always decrease its price when its first offer is rejected?
2011/50. Mathieu VAN VYVE. Linear prices for non-convex electricity markets: models and algorithms.
2011/51. Parkash CHANDER and Henry TULKENS. The Kyoto *Protocol*, the Copenhagen *Accord*, the Cancun *Agreements*, and beyond: An economic and game theoretical exploration and interpretation.
2011/52. Fabian Y.R.P. BOCART and Christian HAFNER. Econometric analysis of volatile art markets.
2011/53. Philippe DE DONDER and Pierre PESTIEAU. Private, social and self insurance for long-term care: a political economy analysis.
2011/54. Filippo L. CALCIANO. Oligopolistic competition with general complementarities.
2011/55. Luc BAUWENS, Arnaud DUFAYS and Bruno DE BACKER. Estimating and forecasting structural breaks in financial time series.
2011/56. Pau OLIVELLA and Fred SCHROYEN. Multidimensional screening in a monopolistic insurance market.

Recent titles

CORE Discussion Papers - continued

2011/57. Knud J. MUNK. Optimal taxation in the presence of a congested public good and an application to transport policy.
2011/58. Luc BAUWENS, Christian HAFNER and Sébastien LAURENT. Volatility models.
2011/59. Pierre PESTIEAU and Grégory PONTHIERE. Childbearing age, family allowances and social security.
2011/60. Julio DÁVILA. Optimal population and education.
2011/61. Luc BAUWENS and Dimitris KOROBILIS. Bayesian methods.
2011/62. Florian MAYNERIS. A new perspective on the firm size-growth relationship: shape of profits, investment and heterogeneous credit constraints.
2011/63. Florian MAYNERIS and Sandra PONCET. Entry on difficult export markets by Chinese domestic firms: the role of foreign export spillovers.
2011/64. Florian MAYNERIS and Sandra PONCET. French firms at the conquest of Asian markets: the role of export spillovers.
2011/65. Jean J. GABSZEWICZ and Ornella TAROLA. Migration, wage differentials and fiscal competition.
2011/66. Robin BOADWAY and Pierre PESTIEAU. Indirect taxes for redistribution: Should necessity goods be favored?
2011/67. Hylke VANDENBUSSCHE, Francesco DI COMITE, Laura ROVEGNO and Christian VIEGELAHN. Moving up the quality ladder? EU-China trade dynamics in clothing.
2011/68. Mathieu LEFEBVRE, Pierre PESTIEAU and Grégory PONTHIERE. Measuring poverty without the mortality paradox.
2011/69. Per J. AGRELL and Adel HATAMI-MARBINI. Frontier-based performance analysis models for supply chain management; state of the art and research directions.
2011/70. Olivier DEVOLDER. Stochastic first order methods in smooth convex optimization.

Books

J. HINDRIKS (ed.) (2008), *Au-delà de Copernic: de la confusion au consensus* ? Brussels, Academic and Scientific Publishers.
J-M. HURIOT and J-F. THISSE (eds) (2009), *Economics of cities.* Cambridge, Cambridge University Press.
P. BELLEFLAMME and M. PEITZ (eds) (2010), *Industrial organization: markets and strategies*. Cambridge University Press.
M. JUNGER, Th. LIEBLING, D. NADDEF, G. NEMHAUSER, W. PULLEYBLANK, G. REINELT, G. RINALDI and L. WOLSEY (eds) (2010), *50 years of integer programming, 1958-2008: from the early years to the state-of-the-art*. Berlin Springer.
G. DURANTON, Ph. MARTIN, Th. MAYER and F. MAYNERIS (eds) (2010), *The economics of clusters – Lessons from the French experience*. Oxford University Press.
J. HINDRIKS and I. VAN DE CLOOT (eds) (2011), *Notre pension en heritage*. Itinera Institute.
M. FLEURBAEY and F. MANIQUET (eds) (2011), *A theory of fairness and social welfare*. Cambridge University Press.
V. GINSBURGH and S. WEBER (eds) (2011), *How many languages make sense? The economics of linguistic diversity*. Princeton University Press.

CORE Lecture Series

D. BIENSTOCK (2001), Potential function methods for approximately solving linear programming problems: theory and practice.
R. AMIR (2002), Supermodularity and complementarity in economics.
R. WEISMANTEL (2006), Lectures on mixed nonlinear programming.