On the linear convergence of the stochastic gradient method with constant step-size *

Volkan Cevher and Bằng Công Vũ

Laboratory for Information and Inference Systems (LIONS),
EPFL, Switzerland
{volkan.cevher,bang.vu}@epfl.ch

Abstract

The strong growth condition (SGC) is known to be a sufficient condition for linear convergence of the stochastic gradient method using a constant step-size γ (SGM-CS). In this paper, we provide a necessary condition, for the linear convergence of SGM-CS, that is weaker than SGC. Moreover, when this necessary is violated up to a additive perturbation σ, we show that both the projected stochastic gradient method using a constant step-size (PSGM-CS), under the restricted strong convexity assumption, and the proximal stochastic gradient method, under the strong convexity assumption, exhibit linear convergence to a noise dominated region, whose distance to the optimal solution is proportional to $\gamma\sigma$.

Keywords: Stochastic gradient, linear convergence, strong growth condition.

Mathematics Subject Classifications (2010): 47H05, 49M29, 49M27, 90C25

1 Introduction

In this paper, we consider the following stochastic convex optimization problem, which is widely studied in the literature; *cf.*, [3, 5, 2] for instances.

Problem 1.1 Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function with L-Lipschitz continuous gradient with an expectation form $f(x) = \mathbf{E}_\xi[K(x,\xi)]$. In the expectation, ξ is a random vector whose probability distribution P is supported on set a $\Omega \subset \mathbb{R}^m$, and $K\colon \mathbb{R}^d \times \Omega \to \mathbb{R}$ is convex function with respect to the variable x. Let $g\colon \mathbb{R}^d \to]-\infty, +\infty]$ be a proper lower semicontinuous convex function. Based on this setup, the problem we are interested in studying can be written as

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) + g(x), \tag{1.1}$$

under the following assumptions:

(i) It is possible to obtain independent and identically distributed (iid) samples $(\xi_t)_{t\in\mathbb{N}}$ of ξ.

* "This is a post-peer-review, pre-copyedit version of an article published in Optimization Letter. The final authenticated version is available online at: https://doi.org/10.1007/s11590-018-1331-1"

(ii) Given $(x_t, \xi_t) \in \mathbb{R}^d \times \Omega$, one can find a point $\nabla K(x_t, \xi_t)$ such that $\mathbf{E}_{\xi_t}[\nabla K(x_t, \xi_t)] = \nabla f(x_t)$. Here, the gradient $\nabla K(x, \xi)$ is taken with respect to x.

The proximal stochastic gradient method (*cf.*, [3, 5, 2], and the references therein) is an elementary method for solving Problem 1.1. This method is extremely simple and highly scalable since it only uses the proximity operator of g and an unbiased estimate of the gradient of f at each iteration. Hence, the method is popular in machine learning and signal processing applications.

In this paper, we focus our attention particularly to the case where g is the indicator of some nonempty, closed convex set C (*cf.*, [9, 13] and the references therein). Then, the proximal stochastic gradient method reduces to the projected stochastic gradient method (PSGM):

$$x_0 \in C \text{ and } (\forall t \in \mathbb{N})\ x_{t+1} = P_C(x_t - \gamma_t \nabla K(x_t, \xi_t)), \tag{1.2}$$

where $\gamma_t > 0$ is the step size. When C is the whole space, (1.2) is the stochastic gradient method (SGM).

While the computational cost of these stochastic methods is much cheaper than their deterministic counterparts, their slow convergence rate is problematic for obtaining high accuracy solutions. Indeed, even when f is strongly convex, PSGM only attains a sub-linear convergence rate in general.

To improve the convergence rate of PSGM, we can use variance reduction as proposed in [16]. When the objective f has a finite sum form ($f = n^{-1}\sum_{i=1}^n f_i$), this method computes the full gradient periodically. Hence, its per iteration cost is dimension dependent. For faster convergence, we can also use the stochastic averaged gradient algorithm (SAGA) in [10], which requires additional memory. Other modifications do exist to circumvent the convergence speed issue.

Surprisingly, SGM with constant step-size (SGM-CS) directly attains linear convergence when f is strongly convex and the strong growth condition (SGC) [6] is satisfied. When f has the finite sum structure, SGC can be written as follows with $B > 0$:

$$\max_{1 \le i \le n} \|\nabla f_i(x)\|^2 \le B\|\nabla f(x)\|^2. \tag{1.3}$$

Such conditions are also used in [14, 12] for the deterministic incremental gradient method and [4] for the incremental Newton method. Note that [6, 14, 12, 4] use above condition for $C = \mathbb{R}^d$.

In this work, we prove that SGC is also a necessary condition for the linear convergence of SGM-CS with step-size γ. When SGC is violated up to an additive perturbation σ and f is restricted strongly convex, we show that PSGM-CS exhibits linear convergence to a noise dominated region, whose distance to the optimal solution is proportional to $\gamma\sigma$. To our knowledge, this result is new. We also derive similar results to the proximal stochastic gradient method.

The paper is organized as follows. We first recall some basic notations in convex analysis in [1] below. Section 2 then presents our main results with a necessary and sufficient condition for the linear convergence of SGM with constant step-size. We also extend these results to the PSGM and the proximal stochastic gradient method. Section 3 studies the necessary condition in the context of the linear convergence of randomized Kaczmarz algorithm. We conclude in Section 4.

Notations. Given a non empty closed convex set C, the projection of x onto C id denoted by $P_C x$. The indicator of C is denoted by ι_C. The proximity operator of a proper lower semicontinuous

convex function g is denoted by prox_g. We denote $\operatorname{dom}(g)$ the effective domain of g. The subdifferential of g at p is defined by $\partial g(p) = \big\{u \in \mathbb{R}^d \mid (\forall x \in \mathbb{R}^d)\ g(x) - g(p) \geq \langle x - p \mid u\rangle\big\}$. When ∂g is a singleton, g is a differentiable function and it is denoted by $\nabla g(p)$. The identity operator is denoted by Id. A single-valued operator $B\colon \mathbb{R}^d \to \mathbb{R}^d$ is β-cocoercive, for some $\beta \in]0, +\infty[$, if

$$(\forall x \in \mathbb{R}^d)(y \in \mathbb{R}^d)\ \langle x - y \mid Bx - By\rangle \geq \beta\|Bx - By\|^2.$$

Given an i.i.d sequence $(\xi_t)_{t\in\mathbb{N}}$, we denote $\mathbf{E}_{\xi_t}[x]$ is the conditional expectation of x with respect to the history $\xi_{[t-1]} = \{\xi_0, \xi_1, \ldots, \xi_{t-1}\}$.

2 Main results

Let us first recall the proximal stochastic gradient algorithm which was proposed for solving Problem 1.1. Let $x_0 \in \mathbb{R}^d$ and $(\xi_t)_{t\in\mathbb{N}}$ be an iid sequence, and let $\gamma_t > 0$. We iterate as follows

$$(\forall t \in \mathbb{N}) \quad x_{t+1} = \operatorname{prox}_{\gamma_t g}(x_t - \gamma_t \nabla K(x_t, \xi_t)). \tag{2.1}$$

Let us define the stochastic gradient mapping, $\mathcal{G}(x_t, \xi_t) = \gamma_t^{-1}(x_t - x_{t+1})$. By the definition of the proximity operator, there exists $q_{t+1} \in \partial g(x_{t+1})$ such that

$$\mathcal{G}(x_t, \xi_t) = q_{t+1} + \nabla K(x_t, \xi_t). \tag{2.2}$$

Our main result can be now stated.

Theorem 2.1 *Suppose that the solution set $\mathcal{S}$ is non-empty, and conditioned on $\xi_{[t-1]} = \{\xi_0, \xi_1 \ldots, \xi_{t-1}\}$:*

$$(\forall t \in \mathbb{N}) \quad \mathbf{E}_{\xi_t}[\|x_{t+1} - x^*\|^2] \leq \omega\|x_t - x^*\|^2 + \gamma_t^2\sigma^2, \tag{2.3}$$

for some constant $\omega \in]0, 1[$, constant $\sigma \in \mathbb{R}$ and $x^* \in \mathcal{S}$. Then, the following holds.

(i) *We have*

$$\mathbf{E}_{\xi_t}[\|\mathcal{G}(x_t, \xi_t)\|^2] \leq \frac{1}{1-\omega}\|\mathbf{E}_{\xi_t}[\mathcal{G}(x_t, \xi_t)]\|^2 + \sigma^2. \tag{2.4}$$

(ii) *If $g \equiv c$ is a constant function, then $q_t \equiv 0$ and*

$$\mathbf{E}_{\xi_t}[\|\nabla K(x_t, \xi_t)\|^2] \leq \frac{1}{1-\omega}\|\nabla f(x_t)\|^2 + \sigma^2. \tag{2.5}$$

Proof. (i): We have $(\forall t \in \mathbb{N})$ $x_{t+1} = x_t - \gamma_t\mathcal{G}(x_t, \xi_t)$. Hence, we have

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \|x_t - x^* - \gamma_t\mathcal{G}(x_t, \xi_t)\|^2\\ &= \|x_t - x^*\|^2 - 2\gamma_t\langle x_t - x^* \mid \mathcal{G}(x_t, \xi_t)\rangle + \gamma_t^2\|\mathcal{G}(x_t, \xi_t)\|^2.\end{aligned}$$

Since x_t depends on the history $\xi_{[t-1]}$, and independent of ξ_t, taking conditional expectation with respect to $\xi_{[t-1]}$, we obtain

$$\mathbf{E}_{\xi_t}[\|x_{t+1} - x^*\|^2] = \|x_t - x^*\|^2 - 2\gamma_t\langle x_t - x^* \mid \mathbf{E}_{\xi_t}[\mathcal{G}(x_t, \xi_t)]\rangle + \gamma_t^2\mathbf{E}_{\xi_t}[\|\mathcal{G}(x_t, \xi_t)\|^2]. \tag{2.6}$$

Now, using (2.3), we derive from (2.6) that

$$\gamma_t^2 \mathbf{E}_{\xi_t}[\|\mathcal{G}(x_t,\xi_t)\|^2] \leq (\omega-1)\|x_t - x^*\|^2 + 2\gamma_t \langle x_t - x^* \mid \mathbf{E}_{\xi_t}[\mathcal{G}(x_t,\xi_t)]\rangle + \gamma_t^2\sigma^2 \tag{2.7}$$

Note that, by Cauchy-Schwarz inequality,

$$\begin{aligned} 2\gamma_t \langle x_t - x^* \mid \mathbf{E}_{\xi_t}[\mathcal{G}(x_t,\xi_t)]\rangle &\leq 2\gamma_t \|x_t - x^*\| \|\mathbf{E}_{\xi_t}[\mathcal{G}(x_t,\xi_t)]\| \\ &\leq (1-\omega)\|x_t - x^*\|^2 + \frac{\gamma_t^2}{1-\omega}\|\mathbf{E}_{\xi_t}[\mathcal{G}(x_t,\xi_t)]\|^2, \end{aligned} \tag{2.8}$$

which implies that

$$2\gamma_t \langle x_t - x^* \mid \mathbf{E}_{\xi_t}[\mathcal{G}(x_t,\xi_t)]\rangle + (\omega-1)\|x_t - x^*\|^2 \leq \frac{\gamma_t^2}{1-\omega}\|\mathbf{E}_{\xi_t}[\mathcal{G}(x_t,\xi_t)]\|^2. \tag{2.9}$$

Therefore, it follows from (2.7) that

$$\gamma_t^2 \mathbf{E}_{\xi_t}[\|\mathcal{G}(x_t,\xi_t)\|^2] \leq \frac{\gamma_t^2}{1-\omega}\|\mathbf{E}_{\xi_t}[\mathcal{G}(x_t,\xi_t)]\|^2 + \gamma_t^2\sigma^2, \tag{2.10}$$

which proves (2.4).

(ii). Since g is a constant function, for all x, $\partial g(x) = \{0\}$, hence $q_{t+1} = 0$ and $\mathcal{G}(x_t,\xi_t) = \nabla K(x_t,\xi_t)$. □

Remark 2.2 In the remainder of this paper, if $\sigma^2 > 0$, (2.5) is called the weak growth condition (WGC) of f; and if $\sigma^2 = 0$, (2.5) is called the growth condition (GC) of f. The growth condition is much weaker than the strong growth condition. We have

$$(SGC) \Longrightarrow (GC) \Longrightarrow (WGC). \tag{2.11}$$

Remark 2.3 Our necessary condition (2.4) remains valid for non-convex, non-smooth f. It also holds in the context of solving monotone inclusions [5] where ∇f is replaced by any cocoercive operator B and ∂g is replaced by any maximally monotone operator A (see [1] for definitions), and $\nabla K(x_t,\xi_t)$ is replaced by any stochastic estimate $r(x_t,\xi_t)$ of Bx_t as in [5]. More precisely, let us consider the following iteration

$$x_{t+1} = (\mathrm{Id} + \gamma_t A)^{-1}(x_t - \gamma_t r(x_t,\xi_t)), \tag{2.12}$$

aiming at solving the following monotone inclusion

$$\text{find } x^* \in \mathbb{R}^d \text{ such that } \ 0 \in Ax^* + Bx^*. \tag{2.13}$$

Suppose that the solution set $\mathcal{S}_1$ of (2.13) is non-empty, and (2.3) is satisfied for some $x^* \in \mathcal{S}_1$. Then (2.4) holds.

In the next theorem, we show that (2.5) is also a sufficient condition for linear convergence (with $\sigma = 0$) of the stochastic gradient method for the class of restricted strongly convex function f. Restricted strong convexity is much weaker than strong convexity, some examples and properties of restricted strongly convex functions can be found in [17]. Note that if f is a strongly convex function, A is a linear mapping, then the composite function $f \circ A$ is restricted strongly convex.

Theorem 2.4 *Suppose that $g = \iota_C$ for some non-empty closed convex set C in $\mathbb{R}^d$ such that the set $\mathcal{S}$ of solutions is non-empty, and that f is μ-restricted strongly convex on C in the sense that $(\forall x \in C)\ f(x) - f(P_{\mathcal{S}}x) \geq 0.5\mu\|x - P_{\mathcal{S}}x\|^2$. Suppose that $f^\star = \inf_{x\in\mathbb{R}^d} f(x) \in \mathbb{R}$, and the following weak growth condition is satisfied:*

$$\mathbf{E}_{\xi_t}[\|\nabla K(x_t, \xi_t)\|^2] \leq M\|\nabla f(x_t)\|^2 + \sigma^2 \tag{2.14}$$

for some positive constant M such that $\mu < 4LM$, and $\sigma \in \mathbb{R}$. Let us define $\gamma_t = \gamma < 1/(LM)$, and set $\rho = \gamma\mu(1 - \gamma LM) \in]0, 1[$. Then, it holds that

$$\mathbf{E}_{\xi_t}[\|x_{t+1} - \overline{x}_{t+1}\|^2] \leq (1 - \rho)\|x_t - \overline{x}_t\|^2 + \gamma^2\sigma_1^2, \tag{2.15}$$

where $\overline{x}_t$ is the projection of x_t onto the set of solutions $\mathcal{S}$ and $\sigma_1^2 = \sigma^2 + 2LM(\min_{x\in C} f(x) - f^\star)$.

Proof. Since $\mathcal{S} \subset C$ and $\overline{x}_t \in C$, we have

$$\begin{aligned}\|x_{t+1} - \overline{x}_{t+1}\|^2 &\leq \|x_{t+1} - \overline{x}_t\|^2 \\ &= \|P_C(x_t - \gamma\nabla K(x_t, \xi_t)) - P_C\overline{x}_t\|^2 \\ &\leq \|x_t - \overline{x}_t - \gamma\nabla K(x_t, \xi_t)\|^2,\end{aligned} \tag{2.16}$$

where the last inequality follows from the non-expansiveness of P_C. Hence, we obtain,

$$\|x_{t+1} - \overline{x}_{t+1}\|^2 \leq \|x_t - \overline{x}_t\|^2 - 2\gamma\,\langle x_t - \overline{x}_t \mid \nabla K(x_t, \xi_t)\rangle + \gamma^2\|\nabla K(x_t, \xi_t)\|^2. \tag{2.17}$$

Since x_t depends on the history $\xi_{[t-1]}$, and independent of ξ_t, taking conditional expectation with respect to $\xi_{[t-1]}$, and using the condition (2.14), we obtain

$$\mathbf{E}_{\xi_t}[\|x_{t+1} - \overline{x}_{t+1}\|^2] \leq \|x_t - \overline{x}_t\|^2 - 2\gamma\,\langle x_t - \overline{x}_t \mid \nabla f(x_t)\rangle + \gamma^2 M\|\nabla f(x_t)\|^2 + \gamma^2\sigma^2. \tag{2.18}$$

Using the L-Lipschitz continuous of ∇f, it follows that

$$\|\nabla f(x_t)\|^2 \leq 2L(f(x_t) - f^\star) = 2L(f(x_t) - f(\overline{x}_t)) + 2L(f(\overline{x}_t) - f^\star). \tag{2.19}$$

Moreover, using the convexity of f, we also have

$$\langle \overline{x}_t - x_t \mid \nabla f(x_t)\rangle \leq f(\overline{x}_t) - f(x_t). \tag{2.20}$$

Inserting (2.19) and (2.20) into (2.18), we get

$$\begin{aligned}\mathbf{E}_{\xi_t}[\|x_{t+1} - \overline{x}_{t+1}\|^2] &\leq \|x_t - \overline{x}_t\|^2 + 2\gamma(f(\overline{x}_t) - f(x_t)) + \gamma^2 2LM(f(x_t) - f(\overline{x}_t)) + \gamma^2\sigma_1^2 \\ &= \|x_t - \overline{x}_t\|^2 - 2\gamma(1 - \gamma LM)(f(x_t) - f(\overline{x}_t)) + \gamma^2\sigma_1^2 \\ &\leq \|x_t - \overline{x}_t\|^2 - \gamma\mu(1 - \gamma LM)\|x_t - \overline{x}_t\|^2 + \gamma^2\sigma_1^2 \\ &= (1 - \rho)\|x_t - \overline{x}_t\|^2 + \gamma^2\sigma_1^2,\end{aligned} \tag{2.21}$$

where the last inequality follows from the μ-restricted strongly convex of f, which proves the desired result. □

Remark 2.5 If f is μ-restricted strongly convex, we can find $\mu > 0$ such that $\mu < 4LM$. When $C = \mathbb{R}^d$, $\sigma = 0$ and $\mu \leq 4LM$, the optimal choice of γ is $1/(2LM)$.

Example 2.6 Suppose that $K(\cdot,\xi)$ is a differentiable function with L_ξ-Lipschitz gradient such that $L_0 = \sup_{\xi\in\Omega} L_\xi < +\infty$. If f is μ-restricted strongly convex and $(\forall t \in \mathbb{N})\ \mathbf{E}_{\xi_t}[\|\nabla K(\overline{x}_t,\xi_t)\|^2] \leq \beta^2 < +\infty$ almost surely, for some positive constant β, then

$$(\forall t \in \mathbb{N})\ \mathbf{E}_{\xi_t}[\|\nabla K(x_t,\xi_t)\|^2] \leq (4L_0/\mu)\|\nabla f(x_t)\|^2 + 2\mathbf{E}_{\xi_t}[\|\nabla K(\overline{x}_t,\xi_t)\|^2], \tag{2.22}$$

where $\overline{x}_t$ is the projection of x_t onto the set of minimizers $\mathcal{S}$. Hence, the condition (2.14) is satisfied with $M = 4L_0/\mu$ and $\sigma^2 = 2\beta^2$.

Proof. Indeed, using the cococercivity of $\nabla K(\cdot,\xi)$, we have

$$\begin{aligned}
\mathbf{E}_{\xi_t}[\|\nabla K(x_t,\xi_t)\|^2] &\leq 2\mathbf{E}_{\xi_t}[\|\nabla K(x_t,\xi_t) - \nabla K(\overline{x}_t,\xi_t)\|^2] + 2\mathbf{E}_{\xi_t}[\|\nabla K(\overline{x}_t,\xi_t)\|^2] \\
&\leq 2L_0\mathbf{E}_{\xi_t}[\langle x_t - \overline{x}_t \mid \nabla K(x_t,\xi_t) - \nabla K(\overline{x}_t,\xi_t)\rangle] + 2\mathbf{E}_{\xi_t}[\|\nabla K(\overline{x}_t,\xi_t)\|^2] \\
&\leq 2L_0 \langle x_t - \overline{x}_t \mid \nabla f(x_t) - \nabla f(\overline{x}_t)\rangle + 2\mathbf{E}_{\xi_t}[\|\nabla K(\overline{x}_t,\xi_t)\|^2] \\
&\leq 2L_0 \langle x_t - \overline{x}_t \mid \nabla f(x_t)\rangle + 2\mathbf{E}_{\xi_t}[\|\nabla K(\overline{x}_t,\xi_t)\|^2].
\end{aligned} \tag{2.23}$$

Suppose that f is μ-restricted strongly convex. We have $f(x_t) - f(\overline{x}_t) \geq 0.5\mu\|x_t - \overline{x}_t\|^2$ and $f(\overline{x}_t) - f(x_t) \geq \langle x_t - \overline{x}_t \mid -\nabla f(x_t)\rangle$. Adding them, we get $\langle x_t - \overline{x}_t \mid \nabla f(x_t)\rangle \geq 0.5\mu\|x_t - \overline{x}_t\|^2$. Therefore, $\|x_t - \overline{x}_t\| \leq (2/\mu)\|\nabla f(x_t)\|$. We have

$$2L_0 \langle x_t - \overline{x}_t \mid \nabla f(x_t)\rangle \leq 2L_0\|x_t - \overline{x}_t\|\|\nabla f(x_t)\| \leq 4L_0\mu^{-1}\|\nabla f(x_t)\|^2. \tag{2.24}$$

Inserting this into (2.23), we get the result. □

Example 2.7 Since $\mathbf{E}_{\xi_t}[\langle \nabla K(x_t,\xi_t) - \nabla f(x_t) \mid \nabla f(x_t)\rangle] = 0$, we have

$$\mathbf{E}_{\xi_t}[\|\nabla K(x_t,\xi_t)\|^2] = \|\nabla f(x_t)\|^2 + \mathbf{E}_{\xi_t}[\|\nabla K(x_t,\xi_t) - \nabla f(x_t)\|^2]. \tag{2.25}$$

Therefore, under the standard condition $\mathbf{E}_{\xi_t}[\|\nabla K(x_t,\xi_t) - \nabla f(x_t)\|^2] \leq \sigma^2$, the condition (2.14) is satisfied.

In the case when $(\forall t \in \mathbb{N})\ \partial g(x_t) = \{Q\}$, then $q_{t+1} = -\nabla f(x^*)$. In this case, the necessary condition, with $\sigma = 0$, becomes $\mathbf{E}_{\xi_t}[\|\nabla K(x_t,\xi_t) - \nabla f(x^*)\|^2] \leq M\|\nabla f(x_t) - \nabla f(x^*)\|^2$. Whenever, this condition is satisfied and f is strongly convex, we can prove that the linear convergence of the proximal stochastic gradient method is obtained. However, the following result shows that (2.14) is also a sufficient for linear convergence to a noise dominated region of the proximal stochastic gradient method.

Proposition 2.8 *Suppose that f is μ-strongly convex, and the weak growth condition* (2.14) *is satisfied. Set $\sigma_1^2 = 2(1+2M)\|\nabla f(x^*)\|^2 + 2\sigma^2$, where x^* is the optimal solution. Let $\gamma_t = \gamma$ be chosen such that $\rho = \gamma\mu(1-2\gamma LM) \in]0,1[$. Then, for iteration* (2.1)*, we have*

$$\mathbf{E}_{\xi_t}[\|x_{t+1} - x^*\|^2] \leq (1-\rho)\|x_t - x^*\|^2 + \gamma^2\sigma_1^2. \tag{2.26}$$

Proof. Since (2.14) is satisfied. Then

$$\begin{aligned}
\mathbf{E}_{\xi_t}[\|\nabla K(x_t,\xi_t) - \nabla f(x^*)\|^2] &\leq 2M\|\nabla f(x_t)\|^2 + 2\|\nabla f(x^*)\|^2 + 2\sigma^2 \\
&\leq 4M\|\nabla f(x_t) - \nabla f(x^*)\|^2 + 2(1+2M)\|\nabla f(x^*)\|^2 + 2\sigma^2.
\end{aligned} \tag{2.27}$$

Since $\text{prox}_{\gamma g}$ is non-expansive and $x^* = \text{prox}_{\gamma g}(x^* - \gamma \nabla f(x^*))$, we have

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &\leq \|x_t - x^* - \gamma(\nabla K(x_t, \xi_t) - \nabla f(x^*))\|^2 \\ &= \|x_t - x^*\|^2 - 2\gamma \langle x_t - x^* \mid \nabla K(x_t, \xi_t) - \nabla f(x^*)\rangle + \gamma^2 \|\nabla K(x_t, \xi_t) - \nabla f(x^*)\|^2.\end{aligned}$$

Taking conditional expectation both sides and using (2.27), we get

$$\begin{aligned}\mathbf{E}_{\xi_t}[\|x_{t+1} - x^*\|^2] &\leq \|x_t - x^*\|^2 - 2\gamma \langle x_t - x^* \mid \nabla f(x_t) - \nabla f(x^*)\rangle \\ &\quad + \gamma^2 4M \|\nabla f(x_t) - \nabla f(x^*)\|^2 + \gamma^2 \sigma_1^2 \\ &\leq \|x_t - x^*\|^2 - (2\gamma - \gamma^2 4LM) \langle x_t - x^* \mid \nabla f(x_t) - \nabla f(x^*)\rangle + \gamma^2 \sigma_1^2 \\ &\leq (1 - \rho)\|x_t - x^*\|^2 + \gamma^2 \sigma_1^2, \end{aligned} \tag{2.28}$$

where the first inequality follows from the cocoercivity of ∇f, and the last equality follows from the strong convexity of f. □

Remark 2.9 When $\sigma_1 > 0$, (2.26) implies that we get linear converge to a noise dominated region proportional to $\gamma\sigma_1$. In the case, $g = 0$ and $g = \iota_C$, this kind of convergence result can be found in [8] and [7], respectively. For the case of the stochastic proximal point algorithm, it is presented in [11].

Remark 2.10 The proposition above remains valid for (2.12). Here ∇f and ∂g are replaced by a cocoercive, strongly monotone operator B and a maximally monotone operator A, respectively; and $\nabla K(x_t, \xi_t)$ is replaced by unbiased estimate $r(x_t, \xi_t)$ of Bx_t as in [5].

Remark 2.11 Under the same conditions as in Proposition 2.8, we see that in the case when γ_t is not constant, $\gamma_t = \mathcal{O}(1/(1+t))$, then there exists $t_0 \in \mathbb{N}$ such that

$$(\forall t \geq t_0) \quad \mathbf{E}[\|x_t - x^*\|^2] = \mathcal{O}(1/t), \tag{2.29}$$

where the expectation is taken over the whole history. This convergence rate is known in [5].

3 Special instances of the necessary condition

We have already proved that the growth condition

$$\mathbf{E}_{\xi_t}[\|\nabla K(x_t, \xi_t)\|^2] \leq M \|\nabla f(x_t)\|^2, \tag{3.1}$$

is the necessary and sufficient condition for linear convergence of the stochastic gradient method for the class of convex differentiable function with gradient Lipschitz and restricted strongly convex. We study this necessary condition to establish the linear convergence of randomized Kaczmarz algorithm [15] and of the stochastic gradient method as in [6].

Let $(a_i)_{1 \leq i \leq m}$ be sequence of colum vectors, with norm 1, in $\mathbb{R}^d$ and $b \in \mathbb{R}^m$ with $(m \geq d)$. Set $(\forall i \in \{1, \ldots, m\})$ $C_i = \big\{x \in \mathbb{R}^d \mid \langle a_i \mid x\rangle = b_i\big\}$. Let A be a matrix with rows $(a_i^T)_{1 \leq i \leq m}$. Let us consider the problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) = \frac{1}{2m} \sum_{i=1}^{m} \|x - P_{C_i} x\|^2, \tag{3.2}$$

under the assumptions that $\varnothing \neq \cap_{i=1}^m C_i$ and A is a full rank matrix. Set $f_i = 0.5\|x - P_{C_i}x\|^2$. Let i_k be chosen uniformly at random in $\{1, \ldots, m\}$. Then

$$\begin{aligned}\mathbf{E}_{i_k}[\|\nabla f_{i_k}(x)\|^2] &= \frac{1}{m}\sum_{i=1}^m \|x - P_{C_i}x\|^2 \\ &= \frac{1}{m}\sum_{i=1}^m |\langle a_i \mid x\rangle - b_i|^2 \\ &= \frac{1}{m}\|Ax - b\|^2. \end{aligned} \tag{3.3}$$

Let us define $A^\dagger = (A^TA)^{-1}A^T$. Then $\|A^\dagger(Ax-b)\| \leq \|(A^TA)^{-1}\|\|A^T(Ax-b)\|$. Let $x^* \in \cap_{i=1}^m C_i$ be $x^* = A^\dagger b$. Then $\|A^\dagger(Ax-b)\| = \|x - A^\dagger b\| = \|x - x^*\| \geq \|A\|^{-1}\|A(x-x^*)\| = \|A\|^{-1}\|Ax-b\|$. Therefore, upon setting $M = m\|A\|^2\|(A^TA)^{-1}\|^2$, we have

$$\mathbf{E}_{i_k}[\|\nabla f_{i_k}(x)\|^2] = \frac{1}{m}\|Ax-b\|^2 \leq \frac{\|A\|^2\|(A^TA)^{-1}\|^2}{m}\|A^T(Ax-b)\|^2 = M\|\nabla f(x)\|^2, \tag{3.4}$$

which shows that the necessary condition (2.5) is satisfied with $\sigma = 0$. Furthermore, since the objective function is restricted strongly convex, in view of above theorem, the stochastic gradient method converges linearly which was also known in [15] with $\gamma = 1$. Further connections to the randomized Kaczmarz algorithm can be found in [8] where the case $\cap_{i=1}^m C_i = \varnothing$ is investigated. In this work, they show that the stochastic gradient method converges linearly to a noise dominated region proportional to $\gamma\sigma$ with $\sigma = 2\mathbf{E}_{i_k}[\|\nabla f_{i_k}(x^*)\|^2]$.

In the general case of f_i. The condition (3.1) is satisfied when

$$(\forall i \in \{1, \ldots, n\})(\forall x \in C) \quad \|\nabla f_i(x)\|^2 \leq M\|\nabla f(x)\|^2. \tag{3.5}$$

4 Conclusions

The strong growth condition is used in [14] where the incremental gradient method converges with a sufficiently small constant step size and in [12] where incremental gradient method converges linearly with a sufficiently small constant step size. Furthermore, and it is also recently used in [4] for linear convergence of the incremental Newton method, and in [6] for linear convergence of the stochastic gradient method. All the existing work agrees that the strong growth condition is very strong, it requires at least the vanishing of stochastic gradient at optimal solution. Unfortunately, our work shows that it is necessary to achieve linear convergence.

Acknowledgments. The authors would like to thank Yen-Huan-Li, Ahmet Alacaoglu for useful discussions. We thank the referees for their suggestions and correction which helped to improve the first version of the manuscript. The work of B. Cong Vu and V. Cevher was supported by European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 725594 - time-data).

References

[1] H. H. Bauschke, P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011).

[2] P. L. Combettes and J.-C. Pesquet, Stochastic approximations and perturbations in forward-backward splitting for monotone operators, *Pure Appl. Funct. Anal.*, vol. 1, pp. 13-37, 2016.

[3] J. C. Duchi and Y. Singer, Efficient online and batch learning using forward backward splitting, *J. Mach. Learn. Res.,* vol. 10, pp. 2899–2934, 2009.

[4] M. Gürbüzbalaban, A. Ozdaglar, P. Parrilo, A globally convergent incremental Newton method, *Math. Program.*, vol. 151, pp. 283-313, 2015.

[5] L. Rosasco, S. Villa, and B. C. Vũ, Stochastic Forward-Backward Splitting for Monotone Inclusions, J. Optim. Theory Appl., vol.169, pp. 388-406, 2016.

[6] M. Schmidt and N. Le Roux, Fast convergence of stochastic Gradient descent under a strong growth condition, 2013, https://arxiv.org/pdf/1308.6370.pdf

[7] A. Nedić and D. Bertsekas, Convergence rate of incremental subgradient algorithms, chapter in Stochastic Optimization: Algorithms and Applications, S. Uryasev and P. Pardalos edts., 263-304, 2000.

[8] D. Needell, N. Srebro and R. Ward, Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm, *Math. Program.,* vol. 155, pp. 549-573, 2016.

[9] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, vol. 19, pp. 1574-1609, 2008.

[10] N. Le Roux, M. Schmidt, and F. Bach, A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. *Adv. Neural Inf. Process. Syst.*, pp. 2663–2671, 2012.

[11] E. Ryu and S. Boyd, Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent, http://web.stanford.edu/ eryu/, 2016.

[12] P. Tseng, An incremental gradient (-projection) method with momentum term and adaptive stepsize rule, *SIAM J. Optim.*, vol. 8, pp. 506-531, 1998.

[13] O. Shamir and T. Zhang, Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes, *ICML*, 2013.

[14] M. V. Solodov, Incremental gradient algorithms with stepsizes bounded away from zero, *Comput. Optim. Appl.*, vol. 11, pp. 23-35, 1998.

[15] T. Strohmer and R. Vershynin, A randomized Kaczmarz algorithm with exponential convergence, *J. Fourier Anal. Appl.*, vol. 15, pp. 262-278, 2009.

[16] L. Xiao and T. Zhang, A proximal stochastic gradient method with progressive variance reduction, *SIAM J. Optim.*, vol., pp. 2057-2075, 2014.

[17] H. Zhang and L. Cheng, Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization, *Optim. Lett.*, vol. 9, pp. 961–979, 2015.