STOCHASTIC SUBGRADIENT METHOD FOR QUASI-CONVEX OPTIMIZATION PROBLEMS

YAOHUA HU, CARISA KWOK WAI YU, AND CHONG LI

ABSTRACT. In this paper, we propose a stochastic subgradient method to solve a nondifferentiable constrained quasi-convex optimization problem. A unit noisy (unbiased) quasi-subgradient, involving the stochastic noise in the quasi-subgradient, is employed in each iteration in place of the deterministic quasi-subgradient. Assuming the Hölder condition of order p, we investigate the convergence properties of the stochastic subgradient method by using the constant, diminishing and dynamic stepsize rules. The stochastic subgradient method has the attractive computational advantage that it avoids the difficulty of calculating the exact quasi-subgradient, while it shares the same convergence behavior as that of the exact subgradient method almost surely, which achieves a more precise tolerance than that of inexact subgradient method. We further apply the stochastic subgradient method to solve the Cobb-Douglas production efficiency problem. The numerical results verify our theoretical results and show the high efficiency of the stochastic subgradient method, especially for large-scale problems.

1. Introduction

Subgradient methods are popular iterative methods for solving nondifferentiable convex optimization problems. Following the pioneering works of Polyak [32] and Ermoliev [13], subgradient methods were further developed by Shor [34] and other researchers in the 1970s. Various properties of subgradient methods have been discovered over the last 40 years. In addition, many extensions and generalizations have been considered and numerous applications have been proposed; see [4, 5, 8, 31, 34] and references therein. Nowadays, because of the simple formulation and low storage requirement, subgradient methods remain important for solving nonsmooth and stochastic optimization problems, particularly for large-scale problems.

2010 *Mathematics Subject Classification.* Primary, 65K05, 90C26; Secondary, 49M37, 74S60.

Key words and phrases. Subgradient method; Stochastic approximate; Quasi-convex optimization; Cobb-Douglas production efficiency.

This work was supported in part by the National Natural Science Foundation of China (grant 11526144,11571308), Natural Science Foundation of SZU (grant 201544), Foundation for Distinguished Young Talents in Higher Education of Guangdong (grant 2015KQNCX145) and a grant from the Research Grants Council of Hong Kong (UGC/FDS14/P03/14).

It is worth noting that the exact subgradient could be difficult to compute in many applications. This is because of errors in measurements, uncertainty in data, or intractability in computation. In such situations, an alternative approach is to get a noisy estimate of the subgradient, which is usually possible and tractable. Adopting the noisy estimate as the true value, the resulting subgradient method is called the inexact subgradient method or stochastic subgradient method. The difference between variants of these two subgradient methods is due to the modes of noise: deterministic or stochastic.

In the context of deterministic optimization, the approximate subgradient method, where an ϵ-subgradient is employed, is widely studied in [1, 10, 23, 25, 34] for solving convex optimization problems. By applying the diminishing and nonvanishing stepsize rules, the convergence results are obtained in consideration of the convergence in objective values and the convergence to a neighborhood of the optimal solution set. In 2010, Nedić and Bertsekas [30] investigated the effect of the deterministic noise, including the noise in computation of subgradients and errors in computation of function values, on subgradient methods for convex optimization problems. They established the convergence to the optimal value within some tolerance, which is expressed in terms of noise and errors.

In the stochastic optimization literature, the stochastic subgradient method was pioneered by Ermoliev [13, 14, 15] and developed by Shor [34] and Bertsekas and Tsitsiklis [6]. Many convergence results of the stochastic subgradient method have been established. It was shown that its generated sequence could achieve the same convergence properties as that of the exact subgradient method almost surely, because the random steps help "average out" the statistical noise in subgradient evaluations. This property is significantly better than that of the inexact subgradient method, in which only the convergence to an approximate optimal value and to a neighborhood of the optimal solution set can be proved when the noise is not vanishing. Because of its cheap computation cost and exact convergence behavior, the stochastic subgradient method is extensively and effectively applied in many fields, such as the online and stochastic learning [12] and the large-scale feasibility problems arising in control [28]. One of the most important applications arises in the network applications, including in-network estimation, learning, signal processing and resource allocation. In particular, the researchers in [20, 26, 29, 33, 36] proposed the stochastic incremental subgradient methods to solve large-scale convex optimization problems in distributed networks and studied the convergence properties and the effect of stochastic errors on the stochastic incremental subgradient methods, including a cyclic and a (non-cyclic) Markov randomized incremental method, under the use of several types of stepsizes.

Most papers in the literature of subgradient methods focus on the category of convex optimization problems. Recently, much attention have been drawn beyond the convex category. One of the most important types is quasi-convex optimization problems, which have many important applications in various areas, such as economics, engineering, management science and various applied sciences; see [3, 9, 17, 35] and references therein. However, the study of using subgradient methods to solve quasi-convex optimization problems is limited. Kiwiel [22] studied convergence properties of the exact subgradient method for solving quasi-convex optimization problems under the use of the diminishing stepsize rule. By extending this work and further using the constant stepsize rule, Hu, Yang and Sim [19] proposed a generic inexact subgradient method to solve quasi-convex optimization problems, and studied the influence of the deterministic noise by describing convergence results in both objective values and iterates and finite convergence to the approximate optimality.

Note in [19, 22] that the quasi-subgradient used in subgradient methods is a normal vector to a strict sublevel set of the objective function at the current iterate. The exact quasi-subgradient is quite difficult to calculate, because the strict sublevel set of a quasi-convex function is difficult to approach exactly. Thus, the stochastic approximate is an alternative approach to get a noisy estimate of the subgradient and make the subgradient method more implementable. However, the research on the stochastic subgradient method for solving quasi-convex optimization problems is still in its infancy. Motivated by practical and theoretical reasons, in this paper, we focus on a stochastic subgradient method for solving the constrained quasi-convex optimization problems

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in X, \end{array} \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is quasi-convex and continuous, and X is nonempty, closed and convex. We denote its optimal solution set and optimal value by X^* and f_*, respectively.

Inspired by the idea in [19, 29] and references therein, we explore convergence properties of the stochastic subgradient method by using the constant, diminishing and dynamic stepsize rules, where a unit noisy (unbiased) quasi-subgradient (see Definition 2.2) is adopted in each iteration. Note in [19] that the epigraph of a convex function is convex; while only the sublevel set of a quasi-convex function is convex. Lacking the convexity assumed in [29], the quasi-convex function is more difficult to deal with. The main technical challenge of the convergence analysis of the stochastic subgradient method is to establish a proper basic inequality, which is a key tool in the literature of subgradient methods, in terms of expectation. To this end, we assume the Hölder condition holds, as in [19], and employ the property of the unit noisy quasi-subgradient. Our convergence results show that the stochastic subgradient method shares the same convergence behavior as that of the exact subgradient method (see [22, Theorem 1]) almost surely, which achieves a better tolerance than that of inexact subgradient method in [19]. Another contribution of our paper is to introduce the dynamic stepsize rule in quasi-convex optimization. To the best of our knowledge, this is the first attempt to apply the dynamic stepsize rule in the subgradient method for solving quasi-convex optimization problems.

In addition, the fractional programming is considered as an application of the quasi-convex model. One example is the Cobb-Douglas production efficiency problem [7]. By applying the stochastic subgradient method, we conduct some numerical experiments on this problem. The numerical results can verify our theoretical analysis and show that the stochastic subgradient type method is highly efficient for the Cobb-Douglas production efficiency problem, even for large-scale problems.

The paper is organized as follows. In section 2, we present the notation and preliminary results used in this paper. In section 3, we investigate convergence properties of the stochastic subgradient method by using the constant, diminishing and dynamic stepsize rules. Its application to the Cobb-Douglas production efficiency problem is demonstrated in section 4.

2. Notation and preliminary results

Let us consider the n-dimensional Euclidean space $\mathbb{R}^n$. A vector is viewed as a column one, and the inner product of two vectors $x, y \in \mathbb{R}^n$ is denoted by $\langle x, y\rangle$. We use $\|x\|$ to denote the standard Euclidean norm, i.e., $\|x\| = \sqrt{\langle x, x\rangle}$. For a set $Z \subseteq \mathbb{R}^n$, we denote the closure, boundary and relative interior of Z by $\text{cl}Z$, $\text{bd}Z$ and $\text{ri}Z$, respectively. For $x \in \mathbb{R}^n$ and $Z \subseteq \mathbb{R}^n$, $\text{dist}(x, Z)$ and $P_Z(x)$ denote the Euclidean distance of x from Z and the projection of x onto Z, respectively, i.e.,

$$\text{dist}(x, Z) := \inf_{z\in Z} \|x - z\| \quad \text{and} \quad P_Z(x) := \arg\min_{z\in Z} \|x - z\|.$$

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\{\phi_k : \Omega \to \mathbb{R}\}$ a sequence of functions on the probability space. The limit inferior of $\{\phi_k\}$ is an extended real valued function defined by

$$\left(\liminf_{k\to\infty} \phi_k\right)(\omega) = \sup_{k\geq 0} \inf_{n\geq k} \phi_n(\omega) \quad \text{for any } \omega \in \Omega.$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be quasi-convex if for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0, 1]$, the following inequality holds

$$f((1 - \alpha)x + \alpha y) \leq \max\{f(x), f(y)\}.$$

For each $\alpha \in \mathbb{R}$, we denote the level sets of f by

$$\text{lev}_{<\alpha} f := \{x \in \mathbb{R}^n : f(x) < \alpha\} \quad \text{and} \quad \text{lev}_{\leq\alpha} f := \{x \in \mathbb{R}^n : f(x) \leq \alpha\}.$$

It is well-known that f is quasi-convex if and only if $\text{lev}_{<\alpha} f$ (and/or $\text{lev}_{\leq\alpha} f$) is convex for all $\alpha \in \mathbb{R}$. Throughout this paper, we assume that

- $f : \mathbb{R}^n \to \mathbb{R}$ is quasi-convex and continuous.

The subdifferential of a quasi-convex function plays an important role in quasi-convex optimization. Several different types of subdifferentials of quasi-convex functions were introduced in the literature, see, e.g., [2, 16, 19, 22, 27]. The earliest one is the Greenberg-Pierskalla subdifferential proposed in [16], which is defined by a quasi-conjugate function based on the quasi-convexity structure. Recently, to meet much boarder class of applications, Kiwiel [22] and Hu, Yang and Sim [19] introduced a quasi-subdifferential, which is a normal cone to the strict sublevel set of the quasi-convex function, and applied such a subgradient in their proposed subgradient methods. Here, we recall the definition of quasi-subdifferential as follows.

Definition 2.1. The quasi-subdifferential of f at x is defined by

$$\partial f(x) = \left\{ g : \langle g, y - x \rangle \leq 0, \forall y \in \text{lev}_{<f(x)} f \right\}.$$

Any vector $g \in \partial f(x)$ is called a quasi-subgradient of f at x.

The relationships between the quasi-subdifferential and convex subdifferential, the quasi-subdifferential and the Greenberg-Pierskalla subdifferential were described in [19].

Definition 2.2. Let $x \in \mathbb{R}^n$ and $\tilde{g}(x) \in \mathbb{R}^n$ be a random vector. $\tilde{g}(x)$ is called

(a) a noisy (unbiased) quasi-subgradient of f at x if $\mathbf{E}\tilde{g}(x) \in \partial f(x)$, that is,

$$\mathbf{E}\langle \tilde{g}(x), y - x \rangle \leq 0 \quad \text{for each } y \in \text{lev}_{<f(x)} f.$$

(b) a unit noisy quasi-subgradient of f at x if it is a noisy quasi-subgradient of f at x and $\|\tilde{g}(x)\| = 1$.

The Hölder condition of order p is used to describe some properties of the quasi-subgradient in [24], and to investigate convergence properties of the inexact subgradient method in [19]. It is a critical assumption for the study of convergence analysis in quasi-convex optimization. It is worth noting that the Hölder condition of order 1 is equivalent to the bounded subgradient assumption, assumed in [23, 29, 30], whenever f is convex.

Definition 2.3. Let $p > 0$ and $L > 0$. $f : \mathbb{R}^n \to \mathbb{R}$ is said to satisfy the Hölder condition of order p with modulus L on $\mathbb{R}^n$ if

$$f(x) - f_* \leq L \text{dist}^p(x, X^*) \quad \text{for each } x \in \mathbb{R}^n.$$

The following lemma describes an important property of a quasi-convex function, which satisfies the Hölder condition. This property locally relates the quasi-subgradient with objective values, which is the key to establish the basic inequality in convergence analysis.

Lemma 2.4. *Let $p > 0$ and $L > 0$. Let $x \in X \setminus X^*$, and $\tilde{g}(x)$ be a unit noisy quasi-subgradient of f at x. Suppose that f satisfies the Hölder condition of order p with modules L on $\mathbb{R}^n$. Then it holds for any $x^* \in X^*$ that*

$$\mathbf{E}\langle \tilde{g}(x), x - x^* \rangle \geq \left(\frac{f(x) - f_*}{L} \right)^{\frac{1}{p}}.$$

Proof. By the blanket assumption that f is quasi-convex and continuous, it follows that its level set $\text{lev}_{<f(x)}f$ is convex and open. Given $x^* \in X^*$, we define

$$r := \inf \left\{ \|y - x^*\| : y \in \text{bd}\left(\text{lev}_{<f(x)}f\right) \right\}.$$

One has by the Hölder condition that

$$f(y) - f_* \leq L\text{dist}^p(y, X^*) \quad \text{for each } y \in \mathbb{R}^n.$$

This, by taking the infimun over $\text{bd}\left(\text{lev}_{<f(x)}f\right)$, implies that

$$f(x) - f_* \leq L \inf \left\{ \text{dist}^p(y, X^*) : y \in \text{bd}\left(\text{lev}_{<f(x)}f\right) \right\} \leq Lr^p. \tag{2.1}$$

Let $\delta \in (0, 1)$. Since $\|\tilde{g}(x)\| = 1$, we obtain that $x^* + \delta r\tilde{g}(x) \in \text{lev}_{<f(x)}f$, and then it follows from Definition 2.2 that

$$\mathbf{E}\langle \tilde{g}(x), x^* + \delta r\tilde{g}(x) - x\rangle \leq 0,$$

that is, $\mathbf{E}\langle \tilde{g}(x), x - x^*\rangle \geq \delta r$. Then one has that $\mathbf{E}\langle \tilde{g}(x), x - x^*\rangle \geq r$, since $\delta \in (0, 1)$ is arbitrary. Hence, by (2.1), we obtain that

$$\mathbf{E}\langle \tilde{g}(x), x - x^*\rangle \geq \left(\frac{f(x) - f_*}{L} \right)^{\frac{1}{p}}.$$

The proof is complete. □

We end this section by recalling the supermartingale convergence theorem in [6, page 148], which is useful in the study of convergence properties of the stochastic subgradient method for quasi-convex optimization problems, so as to make the paper more self-contained.

Lemma 2.5. *Let $\{Y_k\}$, $\{Z_k\}$ and $\{W_k\}$ be three sequences of nonnegative random variables, and let $\{\mathcal{F}_k\}$ be a sequence of sets of random variables such that $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for each k. Suppose that the following conditions are satisfied for each k:*

(a) *Y_k, Z_k and W_k are functions of the random variables in $\mathcal{F}_k$;*
(b) *$\mathbf{E}\{Y_{k+1} \mid \mathcal{F}_k\} \leq Y_k - Z_k + W_k$;*
(c) *$\sum_{k=0}^{\infty} W_k < \infty$.*

Then $\sum_{k=0}^{\infty} Z_k < \infty$, and the sequence $\{Y_k\}$ converges to a nonnegative random variable Y, almost surely.

3. Stochastic subgradient method and convergence analysis

The aims of this section are to propose a stochastic subgradient method to solve the quasi-convex optimization problem (1.1), and to investigate its convergence properties by using different types of stepsize rules. The stochastic subgradient method is formally stated by the following algorithm.

Algorithm 3.1. Select an initial point $x_0 \in \mathbb{R}^n$ and a sequence of stepsizes $\{v_k\} \subseteq (0, +\infty)$. Having x_k, we calculate a unit noisy (unbiased) quasi-subgradient $\tilde{g}(x_k)$ and update x_{k+1} by

$$x_{k+1} = P_X(x_k - v_k\tilde{g}(x_k)). \tag{3.1}$$

For quasi-convex optimization problems, the difference between the stochastic subgradient method and the subgradient methods proposed in [19] and [22] is that Kiwiel [22] studied an exact subgradient method, a deterministic noise is considered in the inexact subgradient method [19], while a stochastic noisy subgradient is employed in the stochastic subgradient method.

The stepsize rule has a critical effect on the convergence behavior and computational performance of subgradient methods. In this paper, we consider the following typical stepsize rules.

(a) *Constant stepsize rule:*

$$v_k = v(>0) \quad \text{for each } k.$$

(b) *Diminishing stepsize rule:*

$$v_k > 0, \quad \lim_{k\to\infty} v_k = 0, \quad \sum_{k=0}^{\infty} v_k = \infty. \tag{3.2}$$

(c) *Dynamic stepsize rule:*

$$v_k = \gamma_k \left(\frac{f(x_k) - f_*}{L}\right)^{\frac{1}{p}} \quad \text{for each } k, \tag{3.3}$$

where $0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2$.

Throughout this section, to investigate convergence properties of the stochastic subgradient method (Algorithm 3.1), we make the following assumption:

- f satisfies the Hölder condition of order p with modulus L on $\mathbb{R}^n$.

We now start the convergence analysis by providing the following basic inequality, which shows a significant property of a stochastic subgradient iteration.

Lemma 3.1. *Let $\{x_k\}$ be a sequence generated by Algorithm 3.1. Fix some $n \in \mathbb{N}$, and let $\mathcal{F}_n := \{x_0, x_1, \ldots, x_n\}$. If $x_n \notin X^*$, then it holds for any $x^* \in X^*$ that*

$$\mathbf{E}\left\{\|x_{n+1} - x^*\|^2 \mid \mathcal{F}_n\right\} \le \|x_n - x^*\|^2 - 2v_n \left(\frac{f(x_n) - f_*}{L}\right)^{\frac{1}{p}} + v_n^2. \tag{3.4}$$

Proof. By (3.1) and the nonexpansive property of projection operator, for any $x^* \in X^*$, we have that

$$\begin{aligned} \|x_{n+1} - x^*\|^2 &\le \|x_n - v_n\tilde{g}(x_n) - x^*\|^2 \\ &= \|x_n - x^*\|^2 - 2v_n \langle \tilde{g}(x_n), x_n - x^* \rangle + v_n^2. \end{aligned}$$

By taking the conditional expectation with respect to $\mathcal{F}_n$, it follows that

$$\begin{aligned} \mathbf{E}\left\{\|x_{n+1} - x^*\|^2 \mid \mathcal{F}_n\right\} &\le \|x_n - x^*\|^2 - 2v_n\mathbf{E}\left\{\langle \tilde{g}(x_n), x_n - x^* \rangle \mid \mathcal{F}_n\right\} + v_n^2 \\ &\le \|x_n - x^*\|^2 - 2v_n \left(\tfrac{f(x_n) - f_*}{L}\right)^{\frac{1}{p}} + v_n^2, \end{aligned}$$

where the last inequality follows from Lemma 2.4. The proof is complete. □

By virtue of Lemma 3.1, we will establish in Theorems 3.2, 3.4 and 3.6 the convergence results of Algorithm 3.1 for stepsize rules (a), (b) and (c), respectively. In particular, we prove in Theorem 3.6 that any sequence generated by Algorithm 3.1 with the dynamic stepsize rule converges to an optimal solution of (1.1) almost surely, which is the first attempt to apply the dynamic stepsize rule in the subgradient method for quasi-convex optimization, to the best of our knowledge.

Theorem 3.2. *Let $\{x_k\}$ be a sequence generated by Algorithm 3.1 with the constant stepsize rule. Then*

$$\liminf_{k\to\infty} f(x_k) \le f_* + L\left(\frac{v}{2}\right)^p \quad \textit{almost surely},$$

that is,

$$P\left(\left\{\omega\in\Omega : \liminf_{k\to\infty} f(x_k)(\omega) \le f_* + L\left(\frac{v}{2}\right)^p\right\}\right) = 1.$$

Proof. Given $\delta > 0$, we consider the feasible level set X_δ defined by

$$X_\delta := X \cap \mathrm{lev}_{<f_*+L(\frac{v}{2}+\delta)^p} f,$$

and let $y_\delta \in X$ be such that $f(y_\delta) = f_* + L\delta^p$ (by the continuity of f). Note that $y_\delta \in X_\delta$ by construction. Define a new process $\{\hat{x}_k\}$ by $\hat{x}_0 = x_0$ and

$$\hat{x}_{k+1} := \begin{cases} P_X\left(\hat{x}_k - v_k\tilde{g}(\hat{x}_k)\right), & \text{if } \hat{x}_k \notin X_\delta, \\ y_\delta, & \text{otherwise.}\end{cases}$$

Thus the process $\{\hat{x}_k\}$ is identical to $\{x_k\}$, except that once $\hat{x}_k$ enters the feasible level set X_δ and then the process terminates with $\hat{x}_k = y_\delta \in X_\delta$. Assume that $\hat{x}_k \notin X_\delta$ for any k and let $\hat{\mathcal{F}}_k := \{\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_k\}$. It says that $f(\hat{x}_k) \ge f_* + L\left(\frac{v}{2}+\delta\right)^p$, and then it follows from Lemma 3.1 (with v in place of v_k) that the following relation holds for any $x^* \in X^*$ and any k

$$\begin{aligned} \mathbf{E}\left\{\|\hat{x}_{k+1} - x^*\|^2 \mid \hat{\mathcal{F}}_k\right\} &\le \|\hat{x}_k - x^*\|^2 - 2v\left(\tfrac{f(\hat{x}_k)-f_*}{L}\right)^{\frac{1}{p}} + v^2 \\ &\le \|\hat{x}_k - x^*\|^2 - 2v\delta. \end{aligned}$$

Then by Lemma 2.5, it follows that $\sum_{k=0}^{\infty} 2v\delta < \infty$ almost surely, which is impossible. Hence $\hat{x}_k \notin X_\delta$ only occurs finitely many times, and $\hat{x}_k \in X_\delta$ for large k. Consequently, in the original process, it holds that

$$\liminf_{k\to\infty} f(x_k) \le f_* + L\left(\frac{v}{2}+\delta\right)^p \quad \text{almost surely.}$$

Since $\delta > 0$ is arbitrary, by letting $\delta \to 0$, we arrive at the conclusion. □

Remark 3.3. Theorem 3.2 shows the convergence of Algorithm 3.1 to the optimal value within some tolerance given in terms of the constant stepsize, that is,

$$T_{\mathrm{Sto}}(v) = L\left(\frac{v}{2}\right)^p.$$

Recall from [19, Theorem 3.1] that the tolerance away from the optimal value for the inexact subgradient method by using the constant stepsize rule is presented in terms of noise (R), errors (ϵ) and the constant stepsize, i.e.,

$$T_{\text{Inexact}}(v, R, \epsilon) = L\left(Rd + \frac{v}{2}(1+R)^2\right)^p + \epsilon.$$

By contrast, the stochastic subgradient method achieves a better tolerance than that of the inexact subgradient method. In particular, $T_{\text{Sto}}(v) = T_{\text{Inexact}}(v, 0, 0)$, which is the tolerance of the exact subgradient method by using the same constant stepsize. This shows the advantage of using randomization, which can also be observed from our experimental results in the next section.

Theorem 3.4. *Let $\{x_k\}$ be a sequence generated by Algorithm 3.1 with the diminishing stepsize rule. Then*

$$\liminf_{k\to\infty} f(x_k) = f_* \quad \textit{almost surely},$$

that is,

$$P\left(\left\{\omega \in \Omega : \liminf_{k\to\infty} f(x_k)(\omega) = f_*\right\}\right) = 1.$$

Furthermore, if $\sum_{k=0}^{\infty} v_k^2 < \infty$, then $\{x_k\}$ converges to an optimal solution of (1.1) *almost surely.*

Proof. The proof of the first statement uses the property of the diminishing stepsize rule (cf. (3.2)) and a line of analysis similar to that of Theorem 3.2. Hence we omit the details.

Recall that Lemma 3.1 describes the basic inequality of the sequence $\{x_k\}$

$$\mathbf{E}\left\{\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right\} \le \|x_k - x^*\|^2 - 2v_k\left(\frac{f(x_k) - f_*}{L}\right)^{\frac{1}{p}} + v_k^2.$$

Note that $\sum_{k=0}^{\infty} v_k = \infty$ and $\sum_{k=0}^{\infty} v_k^2 < \infty$. It follows from [15, Theorem 2] that $\{x_k\}$ converges to an optimal solution of (1.1) almost surely. □

Remark 3.5. Theorem 3.4 describes the exact convergence of the stochastic subgradient method for the quasi-convex optimization problem (1.1) by using the diminishing stepsize rule, which shares the same convergence property as that of the exact subgradient method (see [22, Theorem 1]) almost surely.

Theorem 3.6. *Let $\{x_k\}$ be a sequence generated by Algorithm 3.1 with the dynamic stepsize rule. Then there exist $\bar{x} : \Omega \to X^*$ such that $\{x_k\}$ converges to $\bar{x}$ almost surely, that is,*

$$P\left(\{\omega \in \Omega : \{x_k(\omega)\} \textit{ converges to } \bar{x}(\omega)\}\right) = 1.$$

Proof. It follows from Lemma 3.1 and (3.3) that, for any $x^* \in X^*$,

$$\begin{aligned} \mathbf{E}\left\{\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right\} &\le \|x_k - x^*\|^2 - 2v_k \left(\tfrac{f(x_k)-f_*}{L}\right)^{\frac{1}{p}} + v_k^2 \\ &= \|x_k - x^*\|^2 - \gamma_k(2-\gamma_k)\left(\tfrac{f(x_k)-f_*}{L}\right)^{\frac{2}{p}} \\ &\le \|x_k - x^*\|^2 - \underline{\gamma}(2-\overline{\gamma})\left(\tfrac{f(x_k)-f_*}{L}\right)^{\frac{2}{p}}. \end{aligned} \tag{3.5}$$

Thus by Lemma 2.5, we can show that $\{\|x_k - x^*\|\}$ is convergent and $\lim_{k\to\infty} f(x_k) = f_*$ almost surely. Let Z be a countable and dense subset of X^*. Choose

$$\Theta_z := \{\omega : \{\|x_k(\omega) - z\|\} \text{ converges}\} \quad \text{for each } z \in Z,$$

and

$$\Theta := \bigcap_{z\in Z} \Theta_z.$$

By the elements of probability theory, one has that

$$P(\Theta) = 1 - P(\Theta^c) = 1 - P\left(\bigcup_{z\in Z} \Theta_z^c\right) \ge 1 - \sum_{z\in Z} P(\Theta_z^c) = 1.$$

For any $\omega \in \Theta$ and any $z \in Z$, the sequence $\{\|x_k(\omega) - z\|\}$ converges; hence $\{x_k(\omega)\}$ is bounded and must has the cluster point. Define $\bar{x} : \Omega \to \mathbb{R}^n$ be such that

$$\bar{x}(\omega) \text{ is a cluster point of } \{x_k(\omega)\} \text{ for any } \omega \in \Theta.$$

Since $\lim_{k\to\infty} f(x_k) = f_*$ almost surely, without loss of generality, we can assume that $f(x_k(\omega)) \to f_*$ for any $\omega \in \Theta$. Then it follows from the continuity of f that

$$\bar{x}(\omega) \in X^* \quad \text{for any } \omega \in \Theta.$$

Fix $\epsilon > 0$ and $\omega \in \Theta$. Then there exists $z(\omega) \in Z$ such that

$$\|\bar{x}(\omega) - z(\omega)\| \le \epsilon/3, \tag{3.6}$$

because $\bar{x}(\omega) \in X^*$ and $Z \subseteq X^*$ is dense. Let $\{x_{k_i}(\omega)\}$ be a subsequence of $\{x_k(\omega)\}$ such that $x_{k_i}(\omega) \to \bar{x}(\omega)$. Hence, $\lim_{i\to\infty} \|x_{k_i}(\omega) - z(\omega)\| \le \epsilon/3$ (by (3.6)). By the definition of Θ, one has that $\{\|x_k(\omega) - z(\omega)\|\}$ converges; so that $\lim_{k\to\infty} \|x_k(\omega) - z(\omega)\| \le \epsilon/3$. Then there exists $N \in \mathbb{N}$ such that

$$\|x_k(\omega) - z(\omega)\| \le 2\epsilon/3 \quad \text{for any } k \ge N.$$

Consequently, by (3.6), we obtain that

$$\|x_k(\omega) - \bar{x}(\omega)\| \le \|x_k(\omega) - z(\omega)\| + \|\bar{x}(\omega) - z(\omega)\| \le \epsilon \quad \text{for any } k \ge N.$$

Therefore we proved $\{x_k(\omega)\}$ converges to $\bar{x}(\omega)$ for any $\omega \in \Theta$, where $P(\Theta) = 1$, and the proof is complete. □

4. Application

This section illustrates an application in fractional programming. In general, the objective function in a fractional programming is a certain indicator (e.g. efficiency), characterized by a ratio of technical terms. Fractional programming is widely applied in various areas, such as economics, information theory, management science and applied physics. For details, one can refer to [3, 9, 17, 35] and references therein.

Here, we consider the Cobb-Douglas production efficiency problem introduced by Bradley and Frey [7]. The problem is to maximize the profit/cost ratio, which is an efficiency indicator, i.e., the ratio between the revenue and the expenditure, subject to a variety of constraints on funding levels. In particular, for a set of m projects and a collection of n production factors, the total profit value assigned to these projects can be expressed as the following Cobb-Douglas production function

$$\text{Profit} = a_0 \prod_{j=1}^{n} x_j^{a_j}, \quad \text{where } \sum_{j=1}^{n} a_j = 1,$$

where the variables x_j designate the production factors. The total cost is formulated as a linear function of the levels of investment in these projects, i.e.,

$$\text{Cost} = \sum_{j=1}^{n} c_j x_j + c_0.$$

With the definitions of total profit and total cost, the Cobb-Douglas production efficiency model is expressed as

$$\begin{array}{ll} \max & f(x) := \frac{a_0 \prod_{j=1}^{n} x_j^{a_j}}{\sum_{j=1}^{n} c_j x_j + c_0} \\ \text{s.t.} & \sum_{j=1}^{n} b_{ij} x_j \geq p_i, \quad i = 1, \ldots, m, \\ & x \geq 0, \end{array} \tag{4.1}$$

where p_i represents the profit that must be obtained at project i and b_{ij} is the contribution of the production factor j to project i to realize the profit p_i. According to the circumstance of the Cobb-Douglas production efficiency problem, all parameters on profit (a_j) and cost (c_j) are positive. From [35, Theorems 2.3.3 and 2.5.1], it is obvious that (4.1) is a quasi-concave maximization problem.

Two popular techniques for solving the nonlinear fractional programming are the bisection method [21] and the Dinkelbach's method (also called the parametric method) [11, 35]. However, the Dinkelbach's method is not applicable for solving the Cobb-Douglas production efficiency problem (4.1), because the subproblem of Dinkelbach's method is nonconcave and thus difficult to solve. Note that the subproblem of the bisection method is a nonconvex feasibility problem. We will apply the linearized proximal algorithm in our recent work [18], which can efficiently solve the nonconvex feasibility problem, to solve subproblems of the bisection method.

In order to facilitate the presentation of numerical results, we list the abbreviations of algorithms used for solving the Cobb-Douglas production efficiency problem in Table 1.

Table 1. List of the algorithms for solving the Cobb-Douglas production efficiency problem.

Abbreviations	Algorithms
BSM	**B**i**S**ection **M**ethod in [21].
QSM	Exact **Q**uasi-**S**ubgradient **M**ethod in [22].
AQSM	**A**pproximate **Q**uasi-**S**ubgradient **M**ethod in [19].
StoSM	**Sto**chastic **S**ubgradient **M**ethod.

All numerical experiments are implemented in MATLAB R2009a and executed on a personal laptop (Intel Core i7, 2.00 GHz, 8.00 GB of RAM). In the numerical experiments, the parameters of the problem (4.1) are randomly chosen from different intervals,

$$a_j, b_{ij} \in [0,1], \quad a_0, c_0, c_j \in [0,10], \quad \text{and} \quad p_i \in [0, n/2].$$

The diminishing stepsize rule is chosen as

$$v_k = v/(1 + 0.1k),$$

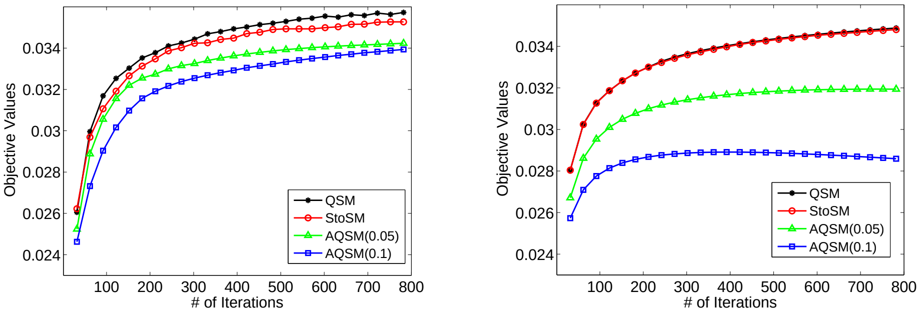
where v is always chosen between $[2,5]$, while the constant stepsize is selected between $[0.5, 2]$. The larger the problem size, the larger the stepsize.

We first compare the performances (in both the accuracy and CPU time) of the BSM, AQSM and StoSM by using the diminishing stepsize rule for different dimensions. The computation results are displayed in Table 2. The noise of the AQSM is set to be a deterministic vector with length being 0.05 or 0.1; while the one of the StoSM is randomly selected following a standard normal distribution. In this table, the columns of Projects and Factors represent the numbers of projects (m) and production factors (n) of problem (4.1) respectively, and f_{opt} and CPU time denote the obtained optimal value and the CPU time (seconds) cost to reach f_{opt} by each algorithm, respectively. From the results shown in Table 2, it is observed that the subgradient type methods are highly efficient for the Cobb-Douglas production efficiency problem (4.1); while the BSM is not suitable for the large-scale Cobb-Douglas production efficiency problem, since it takes too much time in solving the subproblems. We also note that the StoSM achieves a better optimal value than the AQSM in a little shorter time.

The second experiment is performed to compare the convergence behavior of the StoSM, QSM and AQSM by using the constant and diminishing stepsize rules, where the problem size is fixed to be 100×100. The numerical results, plotted in Figure 1, illustrate that the StoSM achieves a better estimation than the AQSM does, which is consistent with Remarks 3.3 and 3.5. In particular, Figure 1(a) and (b) demonstrate the exact convergence of the StoSM to an optimal objective value, while the AQSM only obtain the convergence to an approximate objective value.

Table 2. Computation results for maximizing the Cobb-Douglas production efficiency.

		BSM		AQSM		StoSM	
Projects	Factors	f_{opt}	CPU time	f_{opt}	CPU time	f_{opt}	CPU time
10	10	0.2171	6.75	0.2252	0.06	0.2267	0.04
50	50	0.0477	34.0	0.0502	0.08	0.0547	0.08
100	100	0.0285	66.5	0.0286	0.11	0.0347	0.12
500	500	0.0038	5297	0.0045	0.72	0.0055	0.69
1000	1000	fails	-	0.0021	1.66	0.0025	1.65
2000	2000	fails	-	0.0010	6.24	0.0012	5.86



(a) The constant stepsize rule.

(b) The diminishing stepsize rule.

Figure 1. The convergence behavior of the StoSM, QSM and AQSM.

Finally, we conduct 500 simulations to show the stability of the StoSM, which start from the same initial point and solve the same problem, but follow the different stochastic processes. Figure 2 plots the error bars of the StoSM in such 500 simulations. It is shown that the StoSM is highly stable and converges to an optimal objective value almost surely.

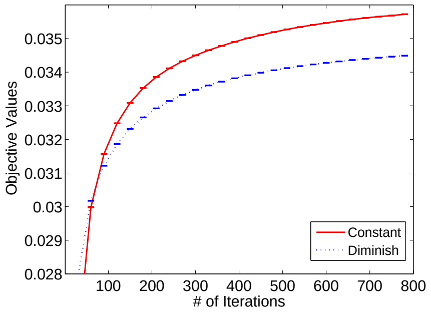


Figure 2. The error bars of the StoSM in 500 simulations.

References

[1] A. Auslender and M. Teboulle. Interior gradient and ϵ-subgradient descent methods for constrained convex minimization. *Mathematics of Operations Research*, 29(1):1–26, 2004.

[2] D. Aussel and A. Daniilidis. Normal characterization of the main classes of quasiconvex functions. *Set-Valued Analysis*, 8:219–236, 2000.

[3] M. Avriel, W. E. Diewert, S. Schaible, and I. Zang. *Generalized Concavity.* Plenum Press, New York, 1988.

[4] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Cambridge, 1999.

[5] D. P. Bertsekas, A. Nedić, , and A. Ozdaglar. *Convex Analysis and Optimization.* Athena Scientific, Cambridge, 2003.

[6] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.

[7] S. P. Bradley and S. C. Frey. Fractional programming with homogeneous functions. *Operations Research*, 22(2):350–357, 1974.

[8] R. S. Burachik, R. N. Gasimov, N. A. Ismayilova, and C. Y. Kaya. On a modified subgradient algorithm for dual problems via sharp augmented Lagrangian. *Journal of Global Optimization*, 34:55–78, 2006.

[9] J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle. *Generalized Convexity, Generalized Monotonicity.* Kluwer Academic Publishers, Dordrecht, 1998.

[10] G. D'Antonio and A. Frangioni. Convergence analysis of deflected conditional approximate subgradient methods. *SIAM Journal on Optimization*, 20(1):357–386, 2009.

[11] W. Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967.

[12] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(2):2121–2159, 2011.

[13] Y. M. Ermoliev. Methods of solution of nonlinear extremal problems. *Cybernetics and Systems Analysis*, 2:1–14, 1966.

[14] Y. M. Ermoliev. *Stochastic Programming Methods.* Nauka, Moscow, 1976.

[15] Y. M. Ermoliev. Stochastic quasigradient methods and their application to system optimization. *Stochastics*, 9(1-2):1–36, 1983.

[16] H. J. Greenberg and W. P. Pierskalla. Quasiconjugate functions and surrogate duality. *Cahiers Centre Études Recherche Opertionnelle*, 15:437–448, 1973.

[17] N. Hadjisavvas, S. Komlósi, and S. Schaible. *Handbook of Generalized Convexity and Generalized Monotonicity.* Springer-Verlag, New York, 2005.

[18] Y. H. Hu, C. Li, and X. Q. Yang. On convergence rates of linearized proximal algorithms for convex composite optimization with applications. SIAM Journal on Optimization, in press.

[19] Y. H. Hu, X. Q. Yang, and C.-K. Sim. Inexact subgradient methods for quasi-convex optimization problems. *European Journal of Operational Research*, 240(2):315–327, 2015.

[20] B. Johansson, M. Rabi, and M. Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2009.

[21] Q. Ke and T. Kanade. Quasiconvex optimization for robust geometric reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1834–1847, 2007.

[22] K. C. Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical Programming*, 90:1–25, 2001.

[23] K. C. Kiwiel. Convergence of approximate and incremental subgradient methods for convex optimization. *SIAM Journal on Optimization*, 14(3):807–840, 2004.

[24] I. V. Konnov. On properties of supporting and quasi-supporting vectors. *Journal of Mathematical Sciences*, 71:2760–2763, 1994.

[25] T. Larsson, M. Patriksson, and A.-B. Strömberg. On the convergence of conditional ϵ-subgradient methods for convex programs and convex-concave saddle-point problems. *European Journal of Operational Research*, 151(3):461–473, 2003.

[26] S. Lee and A. Nedić. Distributed random projection algorithm for convex optimization. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):221–229, April 2013.

[27] J. E. Martínez-Legaz and P. H. Sach. A new subdifferential in quasiconvex analysis. *Journal of Convex Analysis*, 6(1):1–12, 1999.

[28] A. Nedić. Random projection algorithms for convex set intersection problems. Proceedings of *the 49th IEEE Conference on Decision and Control (CDC), 7655–7660, 2010.*

[29] A. Nedić and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[30] A. Nedić and D. P. Bertsekas. The effect of deterministic noise in subgradient methods. *Mathematical Programming*, 125:75–99, 2010.

[31] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009.

[32] B. T. Polyak. A general method for solving extremum problems. *Soviet Mathematics Doklady*, 8:593–597, 1967.

[33] S. S. Ram, A. Nedić, and V. V. Veeravalli. Incremental stochastic subgradient algorithms for convex optimization. *SIAM Journal on Optimization*, 20(2):691–717, 2009.

[34] N. Z. Shor. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag, New York, 1985.

[35] I. M. Stancu-Minasian. *Fractional Programming*. Kluwer Academic Publishers, Dordrecht, 1997.

[36] F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

(Y. H. Hu) College of Mathematics and Statistics, Shenzhen University, Shenzhen 518060, P. R. China
E-mail address: mayhhu@szu.edu.cn

(Carisa K. W. Yu) Department of Mathematics and Statistics, Hang Seng Management College, Hong Kong
E-mail address: carisayu@hsmc.edu.hk

(C. Li) Department of Mathematics, Zhejiang University, Hangzhou 310027, P. R. China
E-mail address: cli@zju.edu.cn