Convex Nondifferentiable Stochastic Optimization: A Local Randomized Smoothing Technique

Farzad Yousefian, Angelia Nedić, and Uday V. Shanbhag

Abstract— **We consider a class of stochastic nondifferentiable optimization problems where the objective function is an expectation of a random convex function, that is not necessarily differentiable. We propose a local smoothing technique, based on random local perturbations of the objective function, that leads to a differentiable approximation of the function. Under the assumption that the local randomness originates from a uniform distribution, we establish a Lipschitzian property for the gradient of the approximation. This facilitates the development of a stochastic approximation framework, which now requires sampling in the product space of the original measure and the artificially introduced distribution. We show that under suitable assumptions, the resulting stochastic subgradient algorithm, with two samples per iteration, converges to an optimal solution of the problem for a diminishing stepsize when the subgradients are bounded.**

I. Introduction

In this paper, we consider a stochastic gradient method for minimizing a convex constrained problem with the objective function given as the expectation of a random convex function. These methods have a long tradition starting with research by Robbins and Monro [1] in 1951, and then followed by the early work of Ermoliev on quasigradient methods [2]–[5].[1] In classical models, for a diminishing stepsize proportional to $\frac{1}{k}$, the asymptotic rate of convergence is the order of the stepsize. To improve the convergence rate, simple averaging schemes have been proposed in [7], [8].

Randomization has been used for some specially structured large-scale convex problems in order to improve the performance of subgradient algorithms [9], [10]. The distributed consensus-based stochastic subgradient methods for minimizing a convex objective over a network has been recently developed and studied in [11]–[13]. In other work [14], two types of stochastic gradient methods for solving stochastic variational inequality problems have been proposed, with convergence analysis reliant on strong monotonicity property on the mapping.

In this paper, we study the convergence behavior of stochastic gradient methods for convex constrained problem with diminishing stepsizes, assuming that the gradient has a Lipschitz property. This is closely related to work in [15], where an unconstrained but nonconvex problem is studied.ODE-based approaches for stochastic approximation have been studied by Borkar and Meyn [16] (also see [17]). We note however that, when the constraint set is present, the ODE analysis is far more involved. Our problem of interest is one where the objective function is nondifferentiable but has bounded subgradients. In this case, we propose a local smoothing technique which leads to a globally differentiable approximation of the original function. Furthermore, the gradient of the resulting differentiable function is Lipschitz continuous. This smoothing technique is motivated by the *global* smoothing technique proposed in [18]. In the remainder of the paper, we apply the local smoothing approach to both deterministic and stochastic convex optimization problems. The novelty of our work lies in developing a local, rather than a global, smoothing framework. This change is crucial in that it precludes the need for the function to be defined everywhere and has particular relevance in constrained settings. Furthermore, for such a local smoothing, we derive a Lipshitz bound on the gradients and shows that it grows modestly with the size of the problem. This Lipshitzian property facilitates the construction of a stochastic approximation framework whose convergence is subsequently proved under a cocoercivity requirement on the gradient.

This paper is organized as the follows. In Section II, we consider the classical stochastic optimization algorithm for a differentiable convex function with Lipschitz gradients, and establish the almost-sure convergence of the algorithm. In Section III, we introduce a local randomized smoothing technique for nondifferentiable convex optimization, and we show its global approximation properties. In Section IV, we apply this smoothing technique to a deterministic nondifferentiable optimization problem, while in Section V, we apply it to a nondifferentiable stochastic problem. We conclude with a discussion in Section VI.

Notation: We view vectors as columns, and write x^T to denote the transpose of a vector x. We use $\|x\|$ to denote the Euclidean vector norm, i.e., $\|x\| = \sqrt{x^T x}$. We write $\Pi_X(x)$ to denote the Euclidean projection of a vector x on a set X. i.e., $\|x - \Pi_X(x)\| = \min_{y\in X} \|x-y\|$. For two sets X and Y, we write $X \subset Y$ to denote that X is a proper subset of Y. Given a nonempty set $X \subseteq \mathbb{R}^n$ and $\varepsilon > 0$, we let X_ε be the set defined by:

$$X_\varepsilon = \{y \mid y = x+z,\ x \in X,\ z \in \mathbb{R},\ \|z\| \leq \varepsilon\}.$$

For a convex function f with domain $\mathrm{dom} f$, a subgradient is defined as follows: a vector g is a *subgradient of f at* $\bar{x} \in \mathrm{dom} f$ if the following relation holds[2]:

$$f(\bar{x}) + g^T(x - \bar{x}) \leq f(x) \qquad \text{for all } x \in \mathrm{dom} f.$$

The authors are with the Department of Industrial and Enterprise Systems Engineering, University of Illinois, Urbana, IL 61801, USA, {yousefi1,angelia,udaybag}@illinois.edu. Nedić and Shanbhag gratefully acknowledge the support of the NSF through the award NSF CMMI 09-48905 ARRA.

[1] See the book by Polyak [6] for an elegant exposition of these methods.

[2] For convex and differentiable f, the inequality holds with $g = \nabla f(\bar{x})$.

The subdifferential set of f at $x=\bar{x}$, denoted by $\partial f(\bar{x})$, is the set of all subgradients of f at $x=\hat{x}$.

We write *a.s.* for "almost sure", and use $\mathsf{E}[\cdot]$ to denote the expectation.

II. Problem Formulation

We consider the following stochastic optimization problem

$$\min_{x\in X} f(x) = \mathsf{E}[F(x,\xi)], \tag{1}$$

where $F:\mathscr{D}\times\Omega\to\mathbb{R}$ is a function, the set $\mathscr{D}\subseteq\mathbb{R}^n$ is open, and the set X is nonempty with $X\subset\mathscr{D}$. The vector $\xi:\Omega\to\mathbb{R}^d$ is a random vector with a probability distribution on a set $\Omega\subseteq\mathbb{R}^d$, and the expectation $\mathsf{E}[F(x,\xi)]$ is taken with respect to ξ. We use X^* to denote the optimal set of problem (1) and f^* to denote its optimal value. We assume the following.

Assumption 1: The set $X\subset\mathscr{D}$ is convex and closed. The function $F(\cdot,\xi)$ is convex on $\mathscr{D}$ for every $\omega\in\Omega$, and the expected value $\mathsf{E}[F(x,\xi)]$ is finite for every $x\in\mathscr{D}$.

Under Assumption 1, the function f is convex over X and the following relation holds

$$\partial f(x) = \mathsf{E}[\partial_x F(x,\xi)], \tag{2}$$

where $\partial_x F(x,\xi)$ denotes the set of all subgradients of $F(x,\xi)$ with respect to the variable x (see [19], [20][3]).

Our primary interest is in problem (1) when f is nondifferentiable function. For such a problem, we will consider a local smoothing technique yielding a differentiable function that approximates f over X. For this reason, we start our discussion by focusing on a differentiable problem (1) and by considering the following iterative algorithm:

$$\begin{aligned} x_{k+1} &= \Pi_X\left(x_k-\gamma_k(\nabla f(x_k)+w_k)\right) \qquad \text{for all } k\geq 0,\\ w_k &= \nabla_x F(x_k,\xi_k)-\nabla f(x_k), \end{aligned} \tag{3}$$

where $x_0\in X$ and γ_k is a (deterministic) stepsize. The random vector w_k is the difference between the sampled gradient $\nabla_x F(x,\xi_k)$ and its expectation $\mathsf{E}[\nabla_x F(x,\xi)]$ at $x=x_k$.

We let $\mathscr{F}_k$ denote the history of the method up to time k, i.e., $\mathscr{F}_k=\{\xi_0,\xi_1,\ldots,\xi_{k-1}\}$ with $\mathscr{F}_0=\emptyset$. By Assumption 1 and relation (2), it follows that $\nabla f(x_k)=\mathsf{E}[\nabla_x F(x_k,\xi)]$ for a differentiable F, implying that w_k has zero-mean, i.e.,

$$\mathsf{E}[w_k\mid\mathscr{F}_k]=0 \qquad \text{for all } k\geq 0.$$

Next,we state some additional assumptions and a result that we use in establishing the convergence of method (3).

Assumption 2: The following conditions hold:

(a) The stepsize is such that $\gamma_k>0$ for all k, $\sum_{k=0}^{\infty}\gamma_k=\infty$, and $\sum_{k=0}^{\infty}\gamma_k^2<\infty$.

(b) The errors w_k satisfy

$$\sum_{k=1}^{\infty}\gamma_k^2\mathsf{E}\left[\|w_k\|^2\mid\mathscr{F}_k\right]<\infty \quad a.s.$$

Assumption 2(b) is satisfied, for example, when $\sum_{k=0}^{\infty}\gamma_k^2<\infty$, and $\|w_k\|\leq c$ a.s. for all k and some scalar c.

[3]In both of these articles, the analysis is for a function defined over $\mathbb{R}^n\times\Omega$, but can be extended to the case of a function defined over $\mathscr{D}\times\Omega$ for a closed convex set $\mathscr{D}\subseteq\mathbb{R}^n$.

The following lemma can be found in [6], page 50.

Lemma 1: (Robbins-Siegmund) Let v_k, u_k, α_k, and β_k be nonnegative random variables, and let

$$\mathsf{E}[v_{k+1}\mid\mathscr{F}_k]\leq(1+\alpha_k)v_k-u_k+\beta_k \quad a.s. \text{ for all } k,$$

$$\sum_{k=0}^{\infty}\alpha_k<\infty \quad a.s., \qquad \sum_{k=0}^{\infty}\beta_k<\infty \quad a.s.,$$

where $\mathscr{F}_k$ denotes the collection $v_0,\ldots,v_k$, $u_0,\ldots,u_k$, $\alpha_0,\ldots,\alpha_k$, $\beta_0,\ldots,\beta_k$. Then, almost surely we have

$$\lim_{k\to\infty}v_k=v, \qquad \sum_{k=0}^{\infty}u_k<\infty.$$

where $v\geq 0$ is some random variable.

Under the preceding assumptions, we study the behavior of the method for a function f with Lipschitz gradients over X, i.e., ∇f such that for a scalar $L>0$,

$$\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\| \qquad \text{for all } x,y\in X.$$

In the next proposition, we show that method (3) converges to an optimal solution of problem (1).

Proposition 1: Let Assumptions 1–2 hold, and let f be differentiable over the set X with Lipschitz gradients. Assume that the optimal set X^* of problem (1) is nonempty. Then, the sequence $\{x_k\}$ generated by method (3) converges almost surely to some point in X^*.

Proof: By definition of the method and the nonexpansive property of the projection operation, we obtain for any $x^*\in X^*$ and $k\geq 0$,

$$\begin{aligned} \|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*-\gamma_k(\nabla f(x_k)+w_k)\|^2\\ &= \|x_k-x^*\|^2-2\gamma_k(\nabla f(x_k)+w_k)^T(x_k-x^*)\\ &\quad+\gamma_k^2\|\nabla f(x_k)+w_k\|^2. \end{aligned}$$

By the convexity of f and the gradient inequality, we have

$$\begin{aligned} \|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*\|^2-2\gamma_k(f(x_k)-f(x^*))\\ &\quad-2\gamma_k w_k^T(x_k-x^*)+\gamma_k^2\|\nabla f(x_k)+w_k\|^2. \end{aligned}$$

Since $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$ for any $a,b\in\mathbb{R}^n$, by using $f^*=f(x^*)$, and by adding and subtracting $\nabla f(x^*)$ in the last term, we obtain

$$\begin{aligned} \|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*\|^2-2\gamma_k(f(x_k)-f^*)\\ &\quad-2\gamma_k w_k^T(x_k-x^*)+2\gamma_k^2\|\nabla f(x_k)-\nabla f(x^*)\|^2\\ &\quad+2\gamma_k^2\|\nabla f(x^*)+w_k\|^2. \end{aligned}$$

Taking the conditional expectation given $\mathscr{F}_k$, using $\mathsf{E}[w_k\mid\mathscr{F}_k]=0$ and the Lipschitzian property of the gradient, we have

$$\begin{aligned} \mathsf{E}\left[\|x_{k+1}-x^*\|^2\mid\mathscr{F}_k\right] &\leq (1+2L^2\gamma_k^2)\|x_k-x^*\|^2\\ &\quad-2\gamma_k(f(x_k)-f^*)\\ &\quad+2\gamma_k^2\left(\|\nabla f(x^*)\|^2+\mathsf{E}\left[\|w_k\|^2\mid\mathscr{F}_k\right]\right). \end{aligned}$$

Under Assumption 2, the conditions of Lemma 1 are satisfied. Therefore, almost surely, the sequence $\{\|x_{k+1}-x^*\|\}$ is convergent for any $x^*\in X^*$, and $\sum_{k=0}^{\infty}\gamma_k(f(x_k)-f(x^*))<\infty$.

The former relation implies that $\{x_k\}$ is bounded a.s., while the latter implies $\liminf_{k\to\infty} f(x_k) = f^*$ a.s. Since the set X is closed and convex, all accumulation points of $\{x_k\}$ lie in X. Furthermore, since $f(x_k) \to f^*$ along a subsequence a.s., by continuity of f, it follows that $\{x_k\}$ has a subsequence converging to some random point in X^* a.s. Moreover, since $\{\|x_{k+1} - x^*\|\}$ is convergent for any $x^* \in X^*$ a.s., the whole sequence $\{x_k\}$ converges to some point in X^* a.s. ■

III. Local Randomized Smoothing Technique

In this section, we consider a nondifferentiable function $f(x)$ and introduce a smooth approximation for $f(x)$, denoted by $\hat{f}$ and defined by

$$\hat{f}(x) \triangleq \mathsf{E}[f(x+z)], \tag{4}$$

where the expectation is with respect to $z \in \mathbb{R}^n$, a random vector with uniform distribution over the n-dimensional ball centered at the origin and with radius ε, i.e., z has the following density function:

$$p(z) = \begin{cases} \frac{1}{c_n \varepsilon^n} & \text{for } \|z\| \le \varepsilon, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $c_n = \dfrac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$, and Γ is the gamma function given by

$$\Gamma\left(\frac{n}{2}+1\right) = \begin{cases} \left(\frac{n}{2}\right)! & \text{if } n \text{ is even,} \\ \sqrt{\pi}\,\frac{n!!}{2^{(n+1)/2}} & \text{if } n \text{ is odd.} \end{cases}$$

Now, we make the following assumption.

Assumption 3: The subgradients of f over X_ε are uniformly bounded, i.e., there is an $L > 0$ such that $\|g\| \le L$ for all $g \in \partial f(x)$ and $x \in X_\varepsilon$.

Assumption 3 is satisfied, for example, when X is bounded.

In the sequel, we use the notation $\mathsf{E}[f(z+z)]$ to denote the vector-valued integral of an element from the set of subdifferentials, and is precisely given by

$$\mathsf{E}[f(x+z))] = \left\{ \bar{g} = \int_{\mathbb{R}^n} g(x+z)p(z)dz \,\middle|\, g(x+z) \in \partial f(x+z)\, a.s. \right\}. \tag{6}$$

The following lemma shows that, under the boundedness of the subgradients of f, the set $\mathsf{E}[f(x+z)]$ defined above is a singleton. In particular, the lemma shows that $\hat{f}$ is convex and differentiable approximation of f with Lipschitz gradients.

Lemma 2: For a convex set $X \subseteq \mathbb{R}^n$, let $f(x)$ be defined and convex on the set X_ε, where $\varepsilon > 0$ is the parameter characterizing the distribution of the random vector z as given in (5). Also, let Assumption 3 hold. Then, for the function $\hat{f}$ given in (4), we have:

(a) $\hat{f}$ is convex and differentiable over X, with gradient

$$\nabla \hat{f}(x) = \mathsf{E}[g(x+z)], g(x+z) \in \partial f(x+z), \text{ for all } x \in X.$$

Furthermore, $\|\nabla \hat{f}(x)\| \le L$ for all $x \in X$.

(b) $f(x) \le \hat{f}(x) \le f(x) + \varepsilon L$ for all $x \in X$.

(c) $\|\nabla \hat{f}(x) - \nabla \hat{f}(y)\| \le \kappa \dfrac{n!!}{(n-1)!!}\dfrac{L}{\varepsilon}\|x-y\|$ for all $x, y \in X$, where $\kappa = \frac{2}{\pi}$ if n is even, and otherwise $\kappa = 1$.

Proof: (a) For the convexity and differentiability of $\hat{f}$ see the proof[4] of Lemma 3.3(a) in [18]. The gradient boundedness follows by Assumption 3, relation (6), and $\nabla \hat{f}(x) = \mathsf{E}[\partial f(x+z)]$.

(b) By (5), the vector z has zero mean, i.e., $\mathsf{E}[x+z] = x$, so that $f(\mathsf{E}[x+z]) = f(x)$. Therefore, by Jensen's inequality and the definition of $\hat{f}$, we have

$$f(x) = f(\mathsf{E}[x+z]) \le \mathsf{E}[f(x+z)] = \hat{f}(x) \qquad \text{for all } x \in X.$$

To show relation $\hat{f}(x) \le f(x) + \varepsilon L$, we use the subgradient inequality for f, which in particular implies that, for every $\bar{x} \in X_\varepsilon$ and $g \in \partial f(\bar{x})$, we have

$$f(\bar{x}) \le f(x) + \|g\|\,\|x - \bar{x}\| \qquad \text{for all } x \in X_\varepsilon.$$

Since $\bar{x} \in X_\varepsilon$, we have $\bar{x} = x + z$ for some $x \in X$ and z with $\|z\| \le \varepsilon$. Using this and the subgradient boundedness, from the preceding relation we obtain

$$f(x+z) \le f(x) + L\varepsilon \qquad \text{for all } x \in X.$$

Thus, by taking the expectation, we get

$$\hat{f}(x) = \mathsf{E}[f(x+z)] \le f(x) + L\varepsilon \qquad \text{for all } x \in X.$$

(c) From part (a) and relation (6), for any $x \in X$, there is a vector $g(z+x)$ such that $g(z+x) \in \partial f(x+z)$ a.s. and

$$\nabla \hat{f}(x) = \int_{\mathbb{R}^n} g(x+z)p(z)dz = \int_{\mathbb{R}^n} g(v)p(v-x)dv,$$

where the last equality follows by letting $v = x+z$. Therefore,

$$\begin{aligned} \|\nabla \hat{f}(x) - \nabla \hat{f}(y)\| &= \left\| \int_{\mathbb{R}^n} (p(z-x) - p(z-y))g(z)dz \right\| \\ &\le \int_{\mathbb{R}^n} |p(z-x) - p(z-y)|\|g(z)\|dz \\ &\le L \int_{\mathbb{R}^n} |p(z-x) - p(z-y)|dz, \end{aligned} \tag{7}$$

where the last inequality follows by noting that the actual integration is over points z satisfying $z \in X_\varepsilon$ (see (5)) and by using the boundedness of the subgradients of f over X_ε.

For estimating $\int_{\mathbb{R}^n} |p(z-x) - p(z-y)|dz$ in (7), we consider cases where $\|x-y\| > 2\varepsilon$ and $\|x-y\| \le 2\varepsilon$.

Case 1 ($\|x-y\| > 2\varepsilon$): For every z with $\|z-x\| < \varepsilon$, we have $\|z-y\| \ge \varepsilon$, implying that $p(z-y) = 0$, so that $\int_{\|z-x\|<\varepsilon} |p(z-x) - p(z-y)|dz = 1$. Likewise, for every z with $\|z-y\| < \varepsilon$, we have $p(z-x) = 0$, implying

$$\int_{\|z-y\|<\varepsilon} |p(z-x) - p(z-y)|dz = 1.$$

Therefore,

$$\begin{aligned} \int_{R^n} |p(z-x) - p(z-y)|dz = &\quad \int_{\|z-x\|<\varepsilon} |p(z-x) - p(z-y)|dz + \\ + &\quad \int_{\|z-y\|<\varepsilon} |p(z-x) - p(z-y)|dz = 2. \end{aligned}$$

[4]There, the vector z has a normal zero-mean distribution. Furthermore, the proof is applicable to a convex function defined over $\mathbb{R}^n$. However, the analysis can be extended in a straightforward way to the case when f is defined over an open convex set $\mathscr{D} \subset \mathbb{R}^n$, since the directional derivative $f'(x;d)$ is finite for each $x \in \mathscr{D}$ and for any direction $d \in \mathbb{R}^n$ (Theorem 23.1 in [21]).

Since $2<\|x-y\|/\varepsilon$, it follows that

$$\int_{R^n} |p(z-x)-p(z-y)|dz \le \frac{\|x-y\|}{\varepsilon}. \tag{8}$$

It can be further seen that $\kappa\frac{n!!}{(n-1)!!} \ge 1$ for all n, which combined with (8) and (7) yields the result.

Case 2 ($\|x-y\| \le 2\varepsilon$): We decompose the integral in (7) over several regions, as follows:

$$\begin{aligned}
&\int_{\mathbb{R}^n} |p(z-x)-p(z-y)|dz \\
&= \int_{\|z-x\|\le\varepsilon \ \&\ \|z-y\|\le\varepsilon} |p(z-x)-p(z-y)|dz \\
&\quad + \int_{\|z-x\|\le\varepsilon \ \&\ \|z-y\|\ge\varepsilon} |p(z-x)-p(z-y)|dz \\
&\quad + \int_{\|z-x\|\ge\varepsilon \ \&\ \|z-y\|\le\varepsilon} |p(z-x)-p(z-y)|dz \\
&\quad + \int_{\|z-x\|\ge\varepsilon \ \&\ \|z-y\|\ge\varepsilon} |p(z-x)-p(z-y)|dz.
\end{aligned}$$

The first and the last integrals are zero, since $p(z-x) = p(z-y)$ for z in the integration region there. Furthermore, in the other two integrals, the supports of $p(z-x)$ and $p(z-y)$ do not intersect, so that we have $|p(z-x)-p(z-y)| = 1/(c_n\varepsilon^n)$ for z in the integration region there. Using this and the symmetry of these integrals, we obtain

$$\int_{\mathbb{R}^n} |p(z-x)-p(z-y)|dz = \frac{1}{c_n\varepsilon^n} 2V_S, \tag{9}$$

where V_S denotes the volume of the set S defined by

$$S \triangleq \{z \in \mathbb{R}^n \mid \|z-x\| \le \varepsilon \text{ and } \|z-y\| \ge \varepsilon\}.$$

Now we want to find an upper bound for V_S in terms of $\|y-x\|$. Let $V_{cap}(d)$ denote the volume of a spherical cap with the distance d from the center of the sphere. Therefore,

$$V_S = c_n\varepsilon^n - 2V_{cap}\left(\frac{\|x-y\|}{2}\right). \tag{10}$$

The volume of an n-dimensional spherical cap with distance d from the center of the sphere can be calculated in terms of the volumes of $n-1$-dimensional spheres, as follows:

$$V_{cap}(d) = \int_d^\varepsilon c_{n-1}\left(\sqrt{\varepsilon^2-\rho^2}\right)^{n-1} d\rho \qquad \text{for } d \in [0,\varepsilon],$$

with $c_n = \dfrac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$ for $n \ge 1$. We have for $d \in [0,\varepsilon]$,

$$\begin{aligned}
V'_{cap}(d) &= -c_{n-1}(\varepsilon^2-d^2)^{\frac{n-1}{2}} \le 0, \\
V''_{cap}(d) &= (n-1)c_{n-1}d(\varepsilon^2-d^2)^{\frac{n-3}{2}} \ge 0,
\end{aligned}$$

where V'_{cap} and V''_{cap} denote the first and the second derivative, respectively, with respect to d. Hence, $V_{cap}(d)$ is convex over $[0,\varepsilon]$, and by the subgradient inequality we have

$$V_{cap}(0) + V'_{cap}(0)\,d \le V_{cap}(d) \qquad \text{for } d \in [0,\varepsilon].$$

Since $V_{cap}(0) = \frac{1}{2}c_n\varepsilon^n$ and $V'_{cap}(0) = -c_{n-1}\varepsilon^{n-1}$, it follows

$$\frac{1}{2}c_n\varepsilon^n - c_{n-1}\varepsilon^{n-1}d \le V_{cap}(d) \qquad \text{for } d \in [0,\varepsilon]. \tag{11}$$

Noting that $\|x-y\|/2 \le \varepsilon$ (since $\|x-y\| \le 2\varepsilon$), we can let $d = \|x-y\|/2 \le \varepsilon$ in (11). By doing so and using (10), we obtain

$$V_S = c_n\varepsilon^n - 2V_{cap}\left(\frac{\|x-y\|}{2}\right) \le 2c_{n-1}\varepsilon^{n-1}\frac{\|x-y\|}{2}.$$

Finally, substituting the preceding relation in (9), we have

$$\int_{\mathbb{R}^n} |p(z-x)-p(z-y)|dz \le \frac{2c_{n-1}}{c_n}\frac{\|x-y\|}{\varepsilon}.$$

Since $c_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$, it can be seen that

$$\frac{2c_{n-1}}{c_n} = \kappa\frac{n!!}{(n-1)!!},$$

with $\kappa = \frac{2}{\pi}$ if n is even, and otherwise $\kappa = 1$. Thus, we have

$$\int_{\mathbb{R}^n} |p(z-x)-p(z-y)|dz \le \kappa\frac{n!!}{(n-1)!!}\frac{\|x-y\|}{\varepsilon}. \tag{12}$$

By combining (12) with (7), we obtain the desired result. ■

It can be seen that the Lipschitz constant $\kappa\frac{n!!}{(n-1)!!}\frac{L}{\varepsilon}$ established in Lemma 2 for the differentiable approximation $\hat{f}$ grows at the rate of $\sqrt{n}$ with the number n of the variables, i.e.,

$$\lim_{n\to\infty} \frac{\kappa\frac{n!!}{(n-1)!!}}{\sqrt{n}} = \sqrt{\frac{\pi}{2}}.$$

This growth rate is worse than the growth rate $\sqrt{\ln(n+1)}$ obtained in [18] for the global smoothing approximation, which uses a normally distributed perturbation vector z. However, it should be emphasized that the smoothing technique in [18] requires the function f to be defined over the entire space since z is drawn from a normal distribution, a somewhat stringent requirement. Our proposed local smoothing technique removes such a requirement, but suffers from a worse growth rate.

IV. Deterministic nondifferentiable optimization

We apply the local smoothing technique to the minimization of a convex but not necessarily differentiable function f. In particular, suppose we want to minimize such a function f over some set X. We may first approximate f by a differentiable function $\hat{f}$ and then minimize $\hat{f}$ over f. In this case, by taking the minimum over $x \in X$ in the relation in Lemma 2(b), we see that $f^* \le \hat{f}^* \le f^* + \varepsilon L$. Thus, we may overestimate the optimal value f^* of the original problem by at most εL, where L is a bound on subgradient norms of f. So we consider the following optimization problem

$$\min_{x\in X} \hat{f}(x) = \mathsf{E}[f(x+z)]. \tag{13}$$

We may solve the problem by considering the method (3), which takes the following form

$$\begin{aligned}
x_{k+1} &= \Pi_X[x_k - \gamma_k(\nabla\hat{f}(x_k) + w_k)] \qquad \text{for } k \ge 0, \\
w_k &= g_k - \nabla\hat{f}(x_k) \qquad \text{with } g_k \in \partial f(x_k+z_k),
\end{aligned} \tag{14}$$

where $\{z_k\}$ is an i.i.d. sequence of random variables with uniform distribution over the n-dimensional sphere centered at the origin and with the radius $\varepsilon > 0$.

We have the following result.

Proposition 2: Let f be defined and convex over some open convex set $\mathscr{D} \subseteq \mathbb{R}^n$. Let X be a closed convex set and let $\varepsilon > 0$ be such that $X_\varepsilon \subset \mathscr{D}$, where ε is the parameter of the distribution of the random vector z as given in (5). Let Assumptions 2(a) and 3 hold. Then, the sequence $\{x_k\}$ generated by method (14) converges almost surely to some optimal solution of problem (13).

Proof: This result follows immediately from Prop 3. ■

V. Stochastic nondifferentiable optimization

In this section, we apply our local smoothing technique to a nondifferentiable stochastic problem of the form (1). Essentially, this amounts to putting the results of Sections II and III together. We thus consider the following problem:

$$\begin{array}{ll} \text{minimize} & \hat{f}(x) \\ \text{subject to} & x \in X \\ \text{where} \hat{f}(x) = \mathsf{E}[f(x+z)], & f(x) = \mathsf{E}[F(x,\xi)], \end{array} \tag{15}$$

F is the function as described in section II, and $\hat{f}$ is a smooth approximation of f with z having a uniform density p as discussed in Section III. In view of Lemma 2(a), we know that εL is an upper bound for the difference between the optimal value $f^* = \min_{x\in X} f(x)$ and $\hat{f}^* = \min_{x\in X} \hat{f}(x)$, under appropriate conditions to be stated shortly. Under these conditions, we are interested in solving the approximate problem in (15).

Note that

$$\hat{f}(x) = \mathsf{E}[f(x+z)] = \mathsf{E}[\mathsf{E}[F(x+z,\xi) \mid \xi]],$$

where the inner expectation is conditioned on ξ and is with respect to z and the outer expectation is with respect to ξ. We note that the variables ξ and z are independent, and by exchanging the order of the expectations, we obtain:

$$\hat{f}(x) = \mathsf{E}\big[\hat{F}(x,\xi)\big], \qquad \text{with } \hat{F}(x,\xi) = \mathsf{E}[F(x+z,\xi)].$$

Thus, the problem in (15) is equivalent to

$$\begin{array}{ll} \text{minimize} & \hat{f}(x) = \mathsf{E}\big[\hat{F}(x,\xi)\big] \\ \text{subject to} & x \in X \\ \hat{F}(x,\xi) = \mathsf{E}[F(x+z,\xi)]. & \end{array} \tag{16}$$

In the following lemma, we provide some conditions ensuring the differentiability of $\hat{F}$ with respect to x, as well as some other properties of $\hat{F}$. The lemma can be viewed as an immediate extension of Lemma 2 to the collection of functions $F(\cdot,\xi(\omega))$ with $\omega \in \Omega$.

Lemma 3: Let the set X and function $F : \mathscr{D} \times \Omega \to \mathbb{R}$ satisfy Assumption 1. Let the parameter ε that characterizes the distribution of z be such that $X_\varepsilon \subset \mathscr{D}$. In addition, assume that the subdifferential set $\partial_x F(x,\xi)$ is uniformly bounded over the set $X_\varepsilon \times \Omega$, i.e., there is a constant L such that

$$\|s\| \le L \quad \text{for all } s \in \partial_x F(x,\xi(\omega)), \text{ and all } x \in X_\varepsilon \text{ and } \omega \in \Omega.$$

Then, for the function $\hat{F} : \mathscr{D} \times \Omega \to \mathbb{R}$ given by $\hat{F}(x,\xi) = \mathsf{E}[F(x+z,\xi)]$, we have:

(a) For every $\omega \in \Omega$, the function $\hat{F}(\cdot,\xi(\omega))$ is convex and differentiable with respect to x at every $x \in X$, and the gradient $\nabla_x \hat{F}(x,\xi)$ is given by

$$\nabla \hat{F}(x,\xi) = \mathsf{E}[\partial F(x+z,\xi)] \qquad \text{for all } x \in X.$$

Furthermore, $\|\nabla_x \hat{F}(x,\xi)\| \le L$ for all $x \in X$ and $\omega \in \Omega$.

(b) $F(x,\xi(\omega)) \le \hat{F}(x,\xi) \le F(x,\xi) + \varepsilon L$ for all $x \in X$ and $\omega) \in \Omega$.

(c) $\|\nabla_x \hat{F}(x,\xi(\omega)) - \nabla_x \hat{F}(y,\xi(\omega))\| \le \kappa \dfrac{n!!}{(n-1)!!} \dfrac{L}{\varepsilon} \|x-y\|$ for all $x, y \in X$ and $\omega \in \Omega$, where $\kappa = \frac{2}{\pi}$ if n is even, and otherwise $\kappa = 1$.

Proof: Under the given assumptions, each of the functions $F(\cdot,\xi(\omega))$ for $\omega \in \Omega$ satisfies the conditions of Lemma 2. Thus, the results follow by applying the lemma to each of the functions $F(\cdot,\xi(\omega))$ for $\omega \in \Omega$. ■

In the light of Lemma 2, the optimal value $\hat{f}^*$ of the approximate problem in (16) is an overestimate of the optimal value f^* of the original problem (1) within the error εL. In particular, by taking the expectation with respect to ξ in the relation of Lemma 2(b), we obtain

$$f^* \le \hat{f}^* \le f^* + \varepsilon L.$$

This motivates solving approximate problem (16). Since for every $\omega \in \Omega$, the function $\hat{F}(\cdot,\xi(\omega))$ is convex and differentiable over the set X, the function $\hat{f}(x) = \mathsf{E}\big[\hat{F}(x,\xi)\big]$ is also convex and differentiable over the set X (see [22]). Thus, the objective function $\hat{f}$ in (16) is differentiable. To solve the problem, we consider the method in (3), which takes the following form:

$$\begin{aligned} x_{k+1} &= \Pi_X[x_k - \gamma_k(\nabla \hat{f}(x_k) + w_k)] \qquad \text{for } k \ge 0, \\ w_k &= s_k - \nabla \hat{f}(x_k) \qquad \text{with } s_k \in \partial_x F(x_k + z_k, \xi_k). \end{aligned} \tag{17}$$

We have the following convergence result for the method.

Proposition 3: Let the assumptions of Lemma 3 hold, and let Assumption 2 hold. Then, the sequence $\{x_k\}$ generated by method (17) converges almost surely to some optimal solution of problem (16).

Proof: It suffices to show that the conditions of Proposition 1 are satisfied for the set X, and the functions $\hat{F}(x,\xi)$ and $\hat{f}(x)$. The result will then follow from Proposition 1.

We first verify that $\hat{F}(x,\xi)$ satisfies Assumption 1 and that $\hat{f}(x)$ has Lipschitz gradients over X. Under the given assumptions, Lemma 3 holds. By Lemma 3(a)–(b), the function $\hat{F}(x,\xi)$ satisfies Assumption 1. Furthermore, by Lemma 3(a) and (c), the function $\hat{F}(x,\xi)$ is differentiable and with Lipschitz gradients for every $\omega \in \Omega$. Hence, $\hat{f}(x) = \mathsf{E}\big[\hat{F}(x,\xi(\omega))\big]$ is also differentiable with the gradient given by $\nabla f(x) = \mathsf{E}\big[\nabla_x \hat{F}(x,\xi)\big]$(see [22]). To see that the gradients $\nabla \hat{f}$ are Lipschitz continuous, we take the expectation in the relation of Lemma 3(c), and we obtain for all $x, y \in X$,

$$\mathsf{E}\big[\|\nabla_x \hat{F}(x,\xi) - \nabla_x \hat{F}(y,\xi)\|\big] \le \kappa \frac{n!!}{(n-1)!!} \frac{L}{\varepsilon} \|x-y\|,$$

where $\kappa = \frac{2}{\pi}$ if n is even, and otherwise $\kappa = 1$. Using Jensen's inequality, we further have for all $x, y \in X$,

$$\|\mathsf{E}\big[\nabla_x \hat{F}(x,\xi)\big] - \mathsf{E}\big[\nabla_x \hat{F}(y,\xi)\big]\| \leq \kappa \frac{n!!}{(n-1)!!}\frac{L}{\varepsilon}\|x-y\|.$$

Since $\nabla \hat{f}(x) = \mathsf{E}\big[\nabla_x \hat{F}(x,\xi)\big]$, it follows that $\nabla \hat{f}(x)$ is Lipschitz over the set X. Thus, the objective function $\hat{f}$ satisfies the conditions of Proposition 1.

We now show that Assumption 2(b) is satisfied. In view of the assumption that $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ (Assumption 2(a)), it suffices to show that $\|w_k\|$ is uniformly bounded. By the definition of w_k in (17), we have for all k,

$$\|w_k\| \leq \|s_k\| + \|\nabla \hat{f}(x_k)\| \qquad \text{with } s_k \in \partial_x F(x_k + z_k, \xi_k),$$

where $x_k \in X$ and $\|z_k\| \leq \varepsilon$ for all k. Thus, $x_k + z_k \in X_\varepsilon$ for all k. By the assumptions of Lemma 3, the subdifferential set $\partial_x F(x,\xi)$ is uniformly bounded over $X_\varepsilon \times \Omega$, implying that

$$\|w_k\| \leq L + \|\nabla \hat{f}(x_k)\| \qquad \text{for all } k \geq 0. \tag{18}$$

We next prove that the gradients $\nabla \hat{f}(x)$ are uniformly bounded over the set X. Taking the expectation in the relation $\|\nabla_x \hat{F}(x, \xi(\omega))\| \leq L$ valid for any $x \in X$ and $\omega \in \Omega$ (Lemma 3(a)), and using Jensen's inequality, we obtain

$$\|\mathsf{E}\big[\nabla_x \hat{F}(x,\xi)\big]\| \leq \mathsf{E}\big[\|\nabla_x \hat{F}(x,\xi)\|\big] \leq L \qquad \text{for } x \in X.$$

Since $\nabla \hat{f}(x) = \mathsf{E}\big[\nabla_x \hat{F}(x,\xi)\big]$, we see that $\|\nabla \hat{f}(x)\| \leq L$ for $x \in X$. This and relation (18) yields

$$\|w_k\| \leq 2L \qquad \text{for all } k \geq 0.$$

thus showing that $\|w_k\|$ is uniformly bounded. ■

VI. Conclusions

In this paper, we have investigated convex stochastic optimization problems where the objective is not necessarily differentiable. Through the use of a local smoothing technique, we propose a stochastic gradient scheme. The convergence of this scheme is proved by showing that the smoothing leads to differentiable functions with Lipschitz gradients. As part of future work, we plan to further study local smoothing approaches with different distributions for the perturbation vector z with the goal of obtaining Lipschitz constants that show more modest growth with the number of variables, rather than the rate (namely $\sqrt{n}$ where n is the size of the decision vector) obtained in this paper.

References

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statistics*, vol. 22, pp. 400–407, 1951.

[2] Y. M. Ermoliev, "On the stochastic quasi-gradient method and stochastic quasi-feyer sequences," *Kibernetika, Kiev*, no. 2, pp. 73–83, 1969.

[3] ——, *Stochastic Programming Methods*. Nauka, Moscow, 1976.

[4] ——, "Stochastic quasigradient methods and their application to system optimization," *Stochastic*, vol. 9, pp. 1–36, 1983.

[5] ——, "Stochastic quasigradient methods," in *Numerical Techniques for Stochastic Optimization*. Sringer-Verlag, 1983, pp. 141–185.

[6] B. Polyak, *Introduction to optimization*. New York: Optimization Software, Inc., 1987.

[7] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, 1992.

[8] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[9] A. Nedić, "Subgradient methods for convex minimization," Ph.D. dissertation, MAssachussetts Institute of Technology, 2001.

[10] A. Nedić and D. P. Bertsekas, "Convergence rate of incremental algorithms," *Stochastic Optimization: Algorithms and Applications*, pp. 223–264, 2001.

[11] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Incremental stochastic subgradient algorithms for convex optimization," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 691–717, 2009.

[12] ——, "Distributed stochastic subgradient projection algorithms for convex optimization," 2009, submitted.

[13] ——, "Asynchronous gossip algorithms for stochastic optimization," 2009, accepted in IEEE Conference on Decision and Control.

[14] H. Jiang and H. Xu, "Stochastic approximation approaches to the stochastic variational inequality problem," *IEEE Transactions Automatic Control*, vol. 53, no. 6, pp. 1462–1475, 2008.

[15] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.

[16] V. S. Borkar and S. P. Meyn, "The O.D.E. method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 447–469 (electronic), 2000.

[17] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[18] H. Lakshmanan and D. Farias, "Decentralized recourse allocation in dynamic networks of agents," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 911–940, 2008.

[19] D. P. Bertsekas, "Stochastic optimization problems with nondifferentiable functionals with an application in stochastic programming," in *Proceedings of 1972 IEEE Conference on Decesion and Control*, 1972, pp. 555–559.

[20] ——, "Stochastic optimization problems with nondifferentiable cost functionals," *Journal of Optimization Theory and Applications*, vol. 12, no. 2, pp. 218–231, 1973.

[21] R. T. Rockafellar, *Convex Analysis*. Princeton, New Jersey: Princeton University Press, 1970.

[22] V. Strassen, "The existence of probability measures with given marginales," *Annals of Mathematical Statistics*, vol. 38, pp. 423–439, 1965.