Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

IMPROVING DIMENSION DEPENDENCE IN COMPLEXITY GUARANTEES FOR ZEROTH-ORDER METHODS VIA EXPONENTIALLY-SHIFTED GAUSSIAN SMOOTHING

Mingrui Wang[1], Prakash Chakraborty[1], and Uday V. Shanbhag[2]

[1]Dept. of Industrial and Manufacturing Engg., Pennsylvania State University, State College, PA, USA
[2]Dept. of Industrial and Operations Engg., University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Smoothing-enabled zeroth-order (ZO) methods for nonsmooth convex stochastic optimization have assumed increasing relevance. A shortcoming of such schemes is the dimension dependence in the complexity guarantees, a concern that impedes truly large-scale implementations. We develop a novel exponentially-shifted Gaussian smoothing (**esGS**) gradient estimator by leveraging a simple change-of-variable argument. The moment bounds of the (**esGS**) estimator are characterized by a muted dependence on dimension. When the (**esGS**) estimator is incorporated within a ZO framework, the resulting iteration complexity bounds are reduced to $\mathcal{O}(n\varepsilon^{-2})$ from $\mathcal{O}(n^2\varepsilon^{-2})$, the latter being the best available for the existing two-point estimator with Gaussian smoothing. More specifically, we provide asymptotic and rate statements for nonsmooth convex and strongly convex regimes. Preliminary comparisons with existing schemes appear promising.

1 INTRODUCTION

Zeroth-order (ZO) methods have gained traction over the last two decades in the context of deterministic (cf. (Larson et al. 2019; Conn et al. 2009)) and stochastic optimization problems (Spall 2005). An obvious advantage of such an approach lies in its reliance on function oracles, rather than gradient information, thereby allowing for accommodating nonsmooth functions, where either the availability of the gradient or a subgradient (or a generalized subgradient in nonconvex settings) is unnecessary. Our framework is reliant on a smoothing framework originating from (Steklov 1907). Consider the following optimization problem.

$$\min_{x\in X} \; f(x) \triangleq \mathbb{E}_{\boldsymbol{\xi}}\left[\, F(x,\boldsymbol{\xi})\,\right], \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$ is a bounded convex set, $\boldsymbol{\xi} : \Omega \to \mathbb{R}^d$ is a random variable taking realizations denoted by $\xi, \Xi \triangleq \{\boldsymbol{\xi}(\omega) \mid \omega \in \Omega\}$, $F : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is a real-valued function, $F(\bullet, \xi)$ is convex and L_0-Lipschitz continuous for almost every $\xi \in \Xi$. Consider f_η for $\eta > 0$, the smoothed counterpart of f, where

$$f_\eta(x) \triangleq \mathbb{E}_Z\left[\, f(x+\eta Z)\,\right], \tag{1.2}$$

where Z is a mean-zero unit-variance random variable taking realizations $z \in \mathbb{R}^n$. Under suitable assumptions, f_η is $\mathcal{O}(1/\eta)$-smooth; i.e. g_η is $\mathcal{O}(1/\eta)$-Lipschitz, where $g_\eta(x) = \nabla f_\eta(x)$ and is expectation-valued. When Z is normally distributed with correlation matrix B^{-1}, $\tilde{g}_\eta$ denotes a gradient estimator of f_η, defined as

$$\tilde{g}_\eta(x,z) \triangleq \left(\tfrac{f(x+\eta z)-f(x)}{\eta}\right) Bz. \tag{1.3}$$

1.1 Prior research.

Convolution-based smoothing appears to have originated from (Steklov 1907) and has found utility in developing schemes for nonsmooth convex (Lakshmanan and Farias 2008; Yousefian, Nedić, and Shanbhag 2012; Nemirovski and Yudin 1983) and nonconvex (Ghadimi and Lan 2013; Nesterov and Spokoiny 2017)

regimes. A commonly employed distribution is the Gaussian distribution and the second moment of the associated gradient estimator may be bounded by $\mathcal{O}(L_0^2 n^2)$ (Nesterov and Spokoiny 2017); specifically, the authors employ such estimators within a zeroth-order framework, obtaining worst-case iteration and sample-complexity bounds for nonsmooth convex problems, both of which are given by $\mathcal{O}(n^2\varepsilon^{-2})$. An alternate zeroth-order avenue referred to as simultaneous perturbation stochastic approximation (SPSA) obviated smoothing, relying instead on finite difference approaches and was suggested by Spall (Spall 1992). More recently, such techniques have examined nonsmooth and nonconvex (Shanbhag and Yousefian 2021) and possibly hierarchical regimes (Cui et al. 2023; Qiu et al. 2023) by leveraging spherical smoothing. A comprehensive theoretical comparison, supported by empirical studies, of gradient approximations is provided in (Berahas et al. 2022).

1.2 Gap, contributions, and organization

Key concern. The dimension dependence in iteration complexity, i.e. $\mathcal{O}(n^2)$, significantly impacts practical implementations, when n is large and the projection onto X is computationally challenging.

Contribution. We develop an exponentially-shifted Gaussian smoothing (**esGS**) gradient estimator for f_η, given by $g_\eta(x,v,z) \triangleq (g_\eta^1, g_\eta^2, \cdots, g_\eta^n)^\top$,

$$g_\eta^i(x,v,z) \triangleq \frac{1}{\eta\sqrt{2\pi}}\left[f\left(x_i+\eta\sqrt{2v}, x^{-i}-z^{-i}\right) - f\left(x_i-\eta\sqrt{2v}, x^{-i}-z^{-i}\right)\right], \qquad \textbf{(esGS)}$$

where $u^{-i} \triangleq (u_j)_{j\neq i}$, V and Z are random variables following $\mathcal{E}xp(1)$ and $\mathcal{N}(0,\eta^2 I)$ taking realizations v and z, respectively. In fact, we show that $\mathbb{E}\left[\|g_\eta(x,V,Z)\|^2\right] \leq \mathcal{O}(L_0^2 n)$, **reducing the dimension dependence to n from n^2**; recall that the corresponding second moment of the two-point gradient estimator (1.3) is $\mathcal{O}(L_0^2 n^2)$. This improvement in dimension dependence is further manifested in the iteration complexity $\mathcal{O}(n\varepsilon^{-2})$ for a ZO framework for resolving nonsmooth convex stochastic optimization problems, which is superior in terms of dimension dependence to $\mathcal{O}(n^2\epsilon^{-2})$ for ZO schemes leveraging the two-point gradient estimator. Our scheme can accommodate diminishing smoothing parameter sequences, implying that the sequences tend to asymptotically exact solutions. In particular, we derive almost-sure and mean-square convergence claims for the iterate sequences and provide iteration and oracle (sample) complexity guarantees in both convex and strongly convex settings. Table 1 compares the iterate and oracle complexity guarantees of related ZO methods on a convex stochastic optimization problem; notably, **esGS** leads to the best known iteration complexity for ZO methods in this setting. **Outline.** The remainder of the paper is organized into four sections. In Section 2, we provide an overview on convolution-based smoothing and introduce our new gradient estimator. Convergence and complexity guarantees for an associated smoothed ZO framework reliant on precisely such an estimator are derived in Section 3, while some empirical validation is provided in Section 4. The paper concludes with some remarks.

Table 1: Comparison of complexity of zeroth-order methods

Literature	Iterate complexity	Oracle complexity	nonsmooth
Nesterov and Spokoiny (2017)	$n^2\varepsilon^{-2}$	$n^2\varepsilon^{-2}$	✓
Ghadimi and Lan (2013)	$\max\{n\varepsilon^{-1}, n\sigma^2\varepsilon^{-2}\}$	$\max\{n\varepsilon^{-1}, n\sigma^2\varepsilon^{-2}\}$	✗
Cui et al. (2023)	$n^2\varepsilon^{-2}$	$n^2\varepsilon^{-2}$	✓
This work	$n\varepsilon^{-2}$	$n^2\varepsilon^{-2}$	✓

2 GRADIENT ESTIMATION

In this section, we introduce the framework for convolution, derive our proposed exponentially-shifted Gaussian smoothing (**esGS**) estimator, and provide some properties for the estimator.

2.1 Convolution and mollification

The convolution of f and g, real-valued measurable functions on $\mathbb{R}^n$, denoted by $f * g$, is defined as

$$f * g(x) \triangleq \int_{\mathbb{R}^n} f(x-y)g(y)dy,$$

for all x such that the integral exists. The following is a special case of a classical result on convolution that relates the partial derivative of the convolution to the convolution of one of the functions and the partial of the second. Note that $g \in C^1$ on a set U implies that $\nabla g(x)$ exists for any $x \in U$, where $\nabla g(x) = \left(\frac{\partial g(x)}{\partial x_i}\right)_{i=1}^n$. If $f : \mathbb{R}^n \to \mathbb{R}$, we generally use ∇f for its gradient map.

Proposition 2.1. (Folland 1999, Prop 8.10) Let $f \in L^1$, $g \in C^1$ and ∇g be bounded. Then $f * g \in C^1$ and $\nabla(f * g) = f * (\nabla g)$.

Now consider a nonnegative function ϕ on $\mathbb{R}^n$. This function ϕ forms the basis for our mollification and we provide its explicit form in Section 2.2. Let $\eta > 0$ and let ϕ_η be defined as

$$\phi_\eta(z) \triangleq \eta^{-n}\phi\left(\eta^{-1}z\right). \tag{2.1}$$

If $\phi \in L^1$ then $\int_{\mathbb{R}^n} \phi_\eta(z)dz$ is independent of η, which immediately follows by change of variables, i.e.

$$\int_{\mathbb{R}^n} \phi_\eta(z)dz = \int_{\mathbb{R}^n} \phi\left(\eta^{-1}z\right)\eta^{-n}dz = \int_{\mathbb{R}^n} \phi(v)dv. \tag{2.2}$$

We now introduce the smoothing or mollification of f via ϕ, as referred to as the ϕ_η-mollified f,

$$f_\eta(x) := f * \phi_\eta(x). \tag{2.3}$$

In fact, mollification preserves strong convexity. Recall that a function is called μ-strongly convex if for $u_1, u_2 \in \mathbb{R}^n$, $f(\lambda u_1 + (1-\lambda)u_2) \leq \lambda f(u_1) + (1-\lambda)f(u_2) - \frac{\lambda(1-\lambda)\mu}{2}\|u_1 - u_2\|^2$ for any $\lambda \in [0,1]$.

Lemma 2.1. Assume $f \in L^1$ is μ-strongly convex with $\mu \geq 0$, and ϕ is smooth satisfying $\int_{\mathbb{R}^n} \phi(u)du = 1$. Then $f_\eta = f * \phi_\eta$ is μ-strongly convex for all $\eta > 0$.

Proof. Notice that by definition,

$$\begin{aligned}
&f_\eta\left(\lambda u_1 + (1-\lambda)u_2\right)\\
&= \int_{\mathbb{R}^n} f(\lambda u_1 + (1-\lambda)u_2 - z)\phi_\eta(z)dz = \int_{\mathbb{R}^n} f(\lambda(u_1 - z) + (1-\lambda)(u_2 - z))\phi_\eta(z)dz\\
&\leq \int_{\mathbb{R}^n} \left[\lambda f(u_1 - z) + (1-\lambda)f(u_2 - z)\right]\phi_\eta(z)dz - \frac{\lambda(1-\lambda)\mu}{2}\|u_1 - u_2\|^2 \int_{\mathbb{R}^n} \phi_\eta(z)dz\\
&= \lambda f_\eta(u_1) + (1-\lambda)f_\eta(u_2) - \frac{\lambda(1-\lambda)\mu}{2}\|u_1 - u_2\|^2,
\end{aligned}$$

where the inequality follows by invoking strong convexity and the last inequality leverages the definition of f_η, (2.2), and $\int_{\mathbb{R}^n} \phi_\eta(z)dz = 1$ by assumption. We have thus shown that f_η is μ-strongly convex. □

2.2 An exponentially-shifted Gaussian smoothing (esGS) gradient estimator

Suppose ϕ, introduced in Section 2.1, is chosen to be the standard Gaussian density on $\mathbb{R}^n$, i.e.

$$\phi(z) \triangleq \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{\sum_{i=1}^n z_i^2}{2}}, \tag{2.4}$$

From (2.1), (2.4), and coordinate-wise decomposability of the standard Gaussian, ϕ_η may be defined as

$$\phi_\eta(z) \,=\, \prod_{i=1}^{n} \rho_\eta(z_i), \text{ where } \rho_\eta(z_i) \,\triangleq\, \tfrac{1}{\eta\sqrt{(2\pi)}} e^{\frac{-z_i^2}{2\eta^2}} \text{ for } i = 1, \cdots, n. \tag{2.5}$$

In addition, we define $\phi_\eta^{(-i)}(z^{-i}) \,\triangleq\, \prod_{j\neq i} \rho_\eta(z_j)$.

Remark 2.1. Note that ϕ_η is infinitely smooth and $\nabla\phi_\eta$ is uniformly bounded.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be as in (1.1) and let f_η denote the mollified function introduced in (2.3). We will need the following assumption on F and X for our future discourse.

Assumption 2.1.

1. $F(\cdot, \xi)$ is a Lipschitz continuous function with Lipschitz constant L_0 for every $\xi \in \Xi$.
2. $F(\cdot, \xi)$ is a convex function for every $\xi \in \Xi$.
3. $X \subseteq \mathbb{R}^n$ is a bounded and convex set.

By invoking Jensen's inequality and Assumption 2.1.1, we may claim that f is L_0-Lipschitz on X. This ensures f is $L^1(\mathbb{R}^n; \mathbb{R})$ since f has bounded support. The following proposition derives the gradient of f_η via convolution and this representation is crucial for our gradient estimates.

Proposition 2.2. Let ϕ and ϕ_η be as in (2.4) and (2.5). Then for $1 \leq i \leq n$, the i-th partial derivative of f_η is given by the following, where $V \sim \mathcal{E}xp(1)$ and $Z \sim \mathcal{N}_n(0, \eta^2 I)$.

$$\tfrac{\partial_i f_\eta(x)}{\partial x_i} \,=\, \tfrac{1}{\eta\sqrt{2\pi}} \mathbb{E}_{V,Z^{-i}} \left[f\left(x_i + \eta\sqrt{2V}, x^{-i} - Z^{-i}\right) - f\left(x_i - \eta\sqrt{2V}, x^{-i} - Z^{-i}\right)\right] \tag{2.6}$$

$$= \tfrac{1}{\eta\sqrt{2\pi}} \mathbb{E}_{V,Z^{-i},\boldsymbol{\xi}} \left[F\left(x_i + \eta\sqrt{2V}, x^{-i} - Z^{-i}, \boldsymbol{\xi}\right) - F\left(x_i - \eta\sqrt{2V}, x^{-i} - Z^{-i}, \boldsymbol{\xi}\right)\right]. \tag{2.7}$$

Proof. Using Remark 2.1 to check the hypothesis of Proposition 2.1, we have that $f_\eta \,\triangleq\, f * \phi_\eta$. Consequently, if $\frac{\partial h(x)}{\partial x_i}$ is denoted by h_i', then by Proposition 2.1, for $i = 1, \cdots, n$, we have

$$f'_{\eta,i} \,=\, (f * \phi_\eta)'_i \,=\, f * \phi'_{\eta,i} \implies \tfrac{\partial f_\eta(x)}{\partial x_i} \,=\, \int_{\mathbb{R}^n} f(x - z) \tfrac{\partial \phi_\eta(z)}{\partial z_i} dz. \tag{2.8}$$

It is readily checked that

$$\tfrac{\partial \phi_\eta(z)}{\partial z_i} \,=\, \left(\phi_\eta^{(-i)}\left(z^{-i}\right)\right) \tfrac{\partial \rho_\eta(z_i)}{\partial z_i} \text{ where } \tfrac{\partial \rho_\eta(z_i)}{\partial z_i} = \tfrac{-z_i}{\eta^3\sqrt{2\pi}} e^{-\frac{z_i^2}{2\eta^2}}. \tag{2.9}$$

From (2.8) we thus have

$$\tfrac{\partial f_\eta(x)}{\partial x_i} \,=\, \int_{\mathbb{R}^{n-1}} \left(\int_{\mathbb{R}} f(x - z) \tfrac{\partial \rho_\eta(z_i)}{\partial z_i} dz_i \right) \phi_\eta^{(-i)}\left(z^{-i}\right) dz^{-i}. \tag{2.10}$$

The inner integral in (2.10) can be split as follows:

$$\int_{\mathbb{R}} f(x - z) \tfrac{\partial \rho_\eta(z_i)}{\partial z_i} dz_i = \int_0^{+\infty} f(x - z) \tfrac{\partial \rho_\eta(z_i)}{\partial z_i} dz_i + \int_0^{+\infty} f(x + z) \tfrac{\partial \rho_\eta(-z_i)}{\partial z_i} dz_i. \tag{2.11}$$

By change of variables, where $z_i \to v = \frac{z_i^2}{2\eta^2}$, from (2.9) and by defining h as $h(v) \,\triangleq\, e^{-v}$, we obtain

$$\begin{aligned} \int_{\mathbb{R}} f(x - z) \tfrac{\partial \rho_\eta(z_i)}{\partial z_i} dz_i =& \tfrac{1}{\eta\sqrt{2\pi}} \left[\int_0^{+\infty} f\left(x_i + \eta\sqrt{2v}, x^{-i} - z^{-i}\right) h(v) dv \right. \\ & \left. - \int_0^{+\infty} f\left(x_i - \eta\sqrt{2v}, x^{-i} - z^{-i}\right) h(v) dv \right], \end{aligned} \tag{2.12}$$

where h can be observed as the density function of the $\mathcal{E}xp(1)$ distribution. Plugging (2.12) into (2.10), we obtain our desired result (2.6). Since $f(x_i, x_{-i}) = \mathbb{E}_{\boldsymbol{\xi}}[F(x_i, x_{-i}, \boldsymbol{\xi})]$, by Fubini's theorem and by leveraging (Folland 1999, Thm 2.27), we observe that (2.7) holds. □

Remark 2.2. The (**esGS**) gradient estimator is obtained as a coordinate-wise difference of function values evaluated at points shifted by an η-scaling of the square-root of an exponential random variable with parameter unity, while maintaining Gaussian smoothing at all other coordinates. Though our result is obtained by a direct analysis, we realized that this framework is reminiscent of weak derivatives and their estimators as in (Fu 2006; Jie and Fu 2022), where derivatives are considered with respect to some parameter of choice. However, we consider derivatives with respect to z_i on the support of ρ_η. In addition, in (2.8), the differentiability requirement is transferred from f to ϕ_η, weakening smoothness requirements on f. This is crucial in obtaining (2.6). We now show that second moment bounds of (**esGS**) have distinctly better dimensional dependence. □

We use $\tilde{g}_\eta(x, v, z, \xi)$ to denote a gradient estimate of $f_\eta(x)$, where the ith component is defined as

$$\tilde{g}^i_\eta(x, v, z, \xi) = \frac{1}{\eta\sqrt{2\pi}}\left[F\left(x_i + \eta\sqrt{2v}, x^{-i} - z^{-i}, \xi\right) - F\left(x_i - \eta\sqrt{2v}, x^{-i} - z^{-i}, \xi\right)\right], \tag{2.13}$$

and $\tilde{g}_\eta(x, v, z, \xi) = (\tilde{g}^1_\eta, \tilde{g}^2_\eta, \cdots, \tilde{g}^n_\eta)^T$. From (2.6), $\tilde{g}_\eta$ is an unbiased estimator for ∇f_η:

$$\mathbb{E}_{V,Z,\boldsymbol{\xi}}\left[\tilde{g}_\eta(x, V, Z, \boldsymbol{\xi})\right] = \nabla f_\eta(x). \tag{2.14}$$

We now prove that the second moment of $\tilde{g}_\eta$ is bounded and scales with n, rather than n^2.

Proposition 2.3. Suppose Assumption 2.1 holds. Then for any x,

$$\mathbb{E}_{V,Z,\boldsymbol{\xi}}[\|\tilde{g}_\eta(x, V, Z, \boldsymbol{\xi})\|^2] \leq \tfrac{4}{\pi} L_0^2 n. \tag{2.15}$$

Proof. Since $F(\bullet, \xi)$ is L_0-Lipschitz continuous,

$$\begin{aligned}\mathbb{E}\left[\left|\tilde{g}^i_\eta(x, V, Z, \boldsymbol{\xi})\right|^2\right] &= \tfrac{1}{\eta^2 2\pi}\mathbb{E}\left[\left|F\left(x_i + \eta\sqrt{2V}, x^{-i} - Z^{-i}, \boldsymbol{\xi}\right) - F\left(x_i - \eta\sqrt{2V}, x^{-i} - Z^{-i}, \boldsymbol{\xi}\right)\right|^2\right] \\ &\leq \tfrac{1}{\eta^2 2\pi}\mathbb{E}\left[\left|L_0\left\|(2\eta\sqrt{2V}, 0, \cdots, 0)\right\|\right|^2\right] \leq \tfrac{4L_0^2}{\pi}\mathbb{E}[V] = \tfrac{4L_0^2}{\pi}.\end{aligned}$$

Then the result follows by noting that $\mathbb{E}\left[\|\tilde{g}_\eta(x, V, Z, \boldsymbol{\xi})\|^2\right] = \sum_{i=1}^n \mathbb{E}\left[\left|\tilde{g}^i_\eta(x, V, Z, \boldsymbol{\xi})\right|^2\right] \leq \frac{4nL_0^2}{\pi}$. □

2.3 Properties of the smoothed function

Lemma 2.2. Suppose Ass. 2.1 holds and f, ϕ_η, and f_η are defined as in (1.1), (2.1), and (2.3), respectively. Then the following hold.

a. The components of ∇f_η are given by (2.6).
b. $|f_\eta(x) - f_\eta(y)| \leq L_0\|x - y\|$ for any $x, y \in X$.
c. $|f_\eta(x) - f(x)| \leq L_0\sqrt{n+1}\eta$ for any $x \in X$.
d. If in addition f is convex, then f_η is convex and for any $x \in X$,

$$f(x) \leq f_\eta(x) \leq f(x) + L_0\sqrt{n+1}\eta. \tag{2.16}$$

e. $\|\nabla f_\eta(x) - \nabla f_\eta(y)\| \leq \frac{2L_0\sqrt{n}}{\eta\sqrt{2\pi}}\|x - y\|$ for any $x, y \in X$.
f. If f is differentiable and L_1-smooth, then $\|\nabla f_\eta(x) - \nabla f(x)\| \leq L_1\eta\sqrt{n}$ for any $x \in X$.

Table 2: Comparison of different gradient estimators

Estimator	$\lvert f_\eta(x) - f(x)\rvert$	$\lVert\nabla f_\eta(x) - \nabla f_\eta(y)\rVert$	$\lVert\nabla f_\eta(x) - \nabla f(x)\rVert$	$\mathbb{E}[\lVert g_\eta\rVert^2]$	Smoothing Method
Nesterov and Spokoiny (2017)	$L_0\sqrt{n}\eta$	$\frac{L_0\sqrt{n}}{\eta}\lVert x-y\rVert$	$\frac{L_1\eta}{2}(n+3)^{3/2}$	$L_0^2(n+4)^2$	Gaussian
Cui et al. (2023)	$L_0\eta$	$\frac{L_0 n}{\eta}\lVert x-y\rVert$	$L_1\eta n$	$L_0^2 n^2$	Spherical
This work	$L_0\sqrt{n+1}\eta$	$\frac{2L_0\sqrt{n}}{\eta\sqrt{2\pi}}\lVert x-y\rVert$	$L_1\eta\sqrt{n}$	$\frac{4}{\pi}L_0^2 n$	Gaussian

Proof. a. By Proposition 2.1, this property follows.

b. By definition of f_η in (2.3), triangle inequality for integrals, and Lipschitz continuity of f, respectively

$$\begin{aligned}|f_\eta(x) - f_\eta(y)| &= \left|\int_{\mathbb{R}^n}\left(f(x-z) - f(y-z)\right)\phi_\eta(z)dz\right| \le \int_{\mathbb{R}^n}|f(x-z) - f(y-z)|\,\phi_\eta(z)dz \\ &\le L_0\|x-y\|\int_{\mathbb{R}^n}\phi_\eta(z)dz = L_0\|x-y\|.\end{aligned}$$

c. By definition of f_η in (2.3), triangle inequality for integrals and Lipschitz continuity of f, respectively

$$\begin{aligned}|f_\eta(x) - f(x)| &= \left|\int_{\mathbb{R}^n} f(x-z)\phi_\eta(z)dz - f(x)\right| = \left|\int_{\mathbb{R}^n}\left(f(x-z) - f(x)\right)\phi_\eta(z)dz\right| \\ &\le \int_{\mathbb{R}^n}|f(x-z) - f(x)|\,\phi_\eta(z)dz \le L_0\int_{\mathbb{R}^n}\|z\|\phi_\eta(z)dz, \qquad (2.17)\end{aligned}$$

By setting $r = \|z\| = \sqrt{\sum_{i=1}^n z_i^2}$, and denoting $S_{n-1} = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}$ as the surface area of the $n-$dimensional unit sphere, by standard spherical coordinate transformation we obtain,

$$\begin{aligned}\int_{\mathbb{R}^n}\|z\|\phi_\eta(z)dz &= \int_0^{+\infty} r\frac{1}{\eta^n(2\pi)^{n/2}}e^{-\frac{r^2}{2\eta^2}}S_{n-1}r^{n-1}dr \overset{u=\frac{r}{\eta}}{=} \frac{\eta\sqrt{2\pi}}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})}\int_0^{+\infty} u^n\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du \\ &= \frac{\eta\sqrt{2\pi}}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})}\frac{1}{2}\mathbb{E}_{U\sim\mathcal{N}(0,1)}\left[|U|^n\right] = \frac{\sqrt{2}\eta\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}, \qquad (2.18)\end{aligned}$$

where we utilize the fact that $\mathbb{E}_{U\sim\mathcal{N}(0,1)}[|U|^n] = \frac{2^{n/2}\Gamma(\frac{n+1}{2})}{\sqrt{\pi}}$ and $U\sim\mathcal{N}(0,1)$. Plugging (2.18) in (2.17) and using Gautschi's inequality $x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s}$, for $x = \frac{n-1}{2}$ and $s = \frac{1}{2}$, it is readily checked that $|f_\eta(x) - f(x)| \le L_0\eta\sqrt{n+1}$.

d. Convexity of f_η follows by setting $\mu = 0$ in Proposition 2.1. By Jensen's inequality and $Z\sim\mathcal{N}_n(0,\eta^2 I)$,

$$f_\eta(x) = \mathbb{E}[f(x-Z)] = \mathbb{E}[f(x+Z)] \ge f(x+\mathbb{E}[Z]) = f(x). \qquad (2.19)$$

Now our desired inequality (2.16) is obtained by combining (2.19) and Lemma 2.2 (c.) .

e. From Proposition 2.1 and by Lipschitz continuity of f,

$$\left|\frac{\partial f_\eta(x)}{\partial x_i} - \frac{\partial f_\eta(y)}{\partial x_i}\right| = \left|\int_{\mathbb{R}}^n\left(f(x-z) - f(y-z)\right)\frac{\partial\phi_\eta(z)}{\partial z_i}dz\right| \le L_0\|x-y\|\int_{\mathbb{R}^n}\left|\frac{\partial\phi_\eta(z)}{\partial z_i}\right|dz \qquad (2.20)$$

From (2.9),

$$\begin{aligned}\int_{\mathbb{R}^n}\left|\frac{\partial\phi_\eta(z)}{\partial z_i}\right|dz &= \int_{\mathbb{R}^n}\phi_\eta^{(-i)}(z^{-i})\left|\frac{\partial\rho_\eta(z)}{\partial z_i}\right|dz = \int_{\mathbb{R}}\left|\frac{\partial\rho_\eta(z)}{\partial z_i}\right|dz_i \qquad (2.21) \\ &= \frac{1}{\eta^2}\mathbb{E}\left[|Z|\right] = \frac{1}{\eta}\sqrt{\frac{2}{\pi}} \implies \left|\frac{\partial f_\eta(x)}{\partial x_i} - \frac{\partial f_\eta(y)}{\partial x_i}\right| \le \frac{1}{\eta}\sqrt{\frac{2}{\pi}}L_0\|x-y\|. \qquad (2.22)\end{aligned}$$

From (2.21),

$$\|\nabla f_\eta(x) - \nabla f_\eta(y)\| = \sqrt{\sum_{i=1}^n \left|\tfrac{\partial f_\eta(x)}{\partial x_i} - \tfrac{\partial f_\eta(y)}{\partial x_i}\right|^2} \leq \sqrt{\sum_{i=1}^n \left(\tfrac{2L_0}{\eta\sqrt{2\pi}}\|x-y\|\right)^2} = \tfrac{2L_0\sqrt{n}}{\eta\sqrt{2\pi}}\|x-y\|.$$

f. Since $f \in C^1, \phi_\eta \in L^1$, by Prop. 2.1, $\int_{\mathbb{R}^n} \phi_\eta(z)dz = 1$, and Jensen's inequality,

$$\left|\tfrac{\partial f_\eta(x)}{\partial x_i} - \tfrac{\partial f_\eta(y)}{\partial x_i}\right| = \left|\int_{\mathbb{R}^n} \left(\tfrac{\partial f(x-z)}{\partial x_i} - \tfrac{\partial f(x)}{\partial x_i}\right)\phi_\eta(z)dz\right| \leq \left(\int_{\mathbb{R}^n} \left|\tfrac{\partial f(x-z)}{\partial x_i} - \tfrac{\partial f(x)}{\partial x_i}\right|^2 \phi_\eta(z)dz\right)^{1/2}$$

By the above inequality and by using $\|\nabla f(x-z) - \nabla f(x)\|^2 \leq L_1^2\|z\|^2$

$$\begin{aligned}\|\nabla f_\eta(x) - \nabla f(x)\|^2 = \sum_{i=1}^n \left(\tfrac{\partial f_\eta(x)}{\partial x_i} - \tfrac{\partial f(x)}{\partial x_i}\right)^2 &\leq \int_{\mathbb{R}^n} \sum_{i=1}^n \left(\tfrac{\partial f(x-z)}{\partial x_i} - \tfrac{\partial f(x)}{\partial x_i}\right)^2 \phi_\eta(z)dz \\ &\leq L_1^2 \int_{\mathbb{R}^n} \|z\|^2 \phi_\eta(z)dz. \end{aligned} \tag{2.23}$$

Now by a spherical transformation,

$$\begin{aligned}\int_{\mathbb{R}^n} \|z\|^2\phi_\eta(z)dz &= \int_0^{+\infty} r^2 \tfrac{1}{\eta^n(2\pi)^{n/2}} e^{-\frac{r^2}{2\eta^2}} S_{n-1} r^{n-1} dr \overset{u=\frac{r}{\eta}}{=} \tfrac{\eta^2\sqrt{2\pi}}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^{+\infty} u^{n+1} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\ &= \tfrac{\eta\sqrt{2\pi}}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \cdot \frac{1}{2}\mathbb{E}_{U\sim\mathcal{N}(0,1)}[|U|^{n+1}] = \tfrac{2\eta^2\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n}{2})} = \tfrac{2\eta^2\frac{n}{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n}{2})} = \eta^2 n, \end{aligned} \tag{2.24}$$

where we recall that $\Gamma(a+1) = a\Gamma(a)$. Substituting (2.24) in (2.23), we obtain our result.

□

3 RATE, CONVERGENCE, AND COMPLEXITY GUARANTEES

We now consider (1.1), where we remind the reader that Assumption 2.1 is assumed to hold. For any given scalar $\eta > 0$, we consider the smoothing of the integrand $F(\bullet, \xi)$, given by $F_\eta(\bullet, \xi)$ and defined as

$$F_\eta(x, \xi) = \int_{\mathbb{R}^n} F(x - z, \xi)\phi_\eta(z)dz, \tag{3.1}$$

where ϕ_η is defined in (2.5). Consider the following update rule for generating $\{x_k\}$, given $x_0 \in X$,

$$x_{k+1} \;=\; \Pi_X\left[x_k - \gamma_k\left(\nabla f_{\eta_k}(x_k) + w_k\right)\right], \quad \text{for all } k \geq 0 \tag{3.2}$$

where $\eta_k > 0$, $w_k = \tilde{g}_{\eta_k}(x_k, v_k, z_k, \xi_k) - \nabla f_{\eta_k}(x_k)$. Here $\{V_k\}_{k=1}^N$ are i.i.d $\mathcal{E}xp(1)$ and $\{Z_k\}_{k=1}^N$ are i.i.d. $\mathcal{N}_n(0, \eta^2 I)$ with realizations $\{v_k\}$ and $\{z_k\}$, respectively. From Prop. 2.2, note that $\tilde{g}_\eta$ is an unbiased estimator for ∇F_η. Denote by $\mathcal{F}_k$, the natural filtration generated by the first k iterates $\{x_0, \ldots, x_{k-1}\}$. Since $f(x) = \mathbb{E}_{\boldsymbol{\xi}}[F(x, \boldsymbol{\xi})]$ and $g_\eta(x, V, Z, \boldsymbol{\xi})$ is an unbiased estimator of $\nabla F_\eta(x, \xi)$, by the independence of V, Z and ξ

$$\begin{aligned}\mathbb{E}[w_k \mid \mathcal{F}_k] &= \mathbb{E}[\tilde{g}_{\eta_k}(x_k, V_k, Z_k, \boldsymbol{\xi}_k) - \nabla F_{\eta_k}(x_k, \boldsymbol{\xi}_k) + \nabla F_{\eta_k}(x_k, \boldsymbol{\xi}_k) - \nabla f_{\eta_k}(x_k) \mid \mathcal{F}_k] \\ &= \mathbb{E}_{V_k, Z_k}[\tilde{g}_{\eta_k}(x_k, V_k, Z_k, \boldsymbol{\xi}_k) - \nabla F_{\eta_k}(x_k, \boldsymbol{\xi}_k)] + \mathbb{E}_{\boldsymbol{\xi}_k}[\nabla F_{\eta_k}(x_k, \boldsymbol{\xi}_k) - \nabla f_{\eta_k}(x_k) \mid \mathcal{F}_k] = 0. \end{aligned}$$

Let us now introduce a few assumptions on the terms of our stochastic approximation algorithm (3.2)

Assumption 3.1. The sequences $\{\gamma_k, \eta_k\}$ are positive, satisfying $\sum_{k=0}^{\infty} \gamma_k = \infty$, $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ and $\sum_{k=0}^{\infty} \gamma_k \eta_k < \infty$.

Akin to (Yousefian, Nedić, and Shanbhag 2012), we derive consistency claims for Algo. (3.2) on (1.1).

Proposition 3.1 (a.s. convergence). If Assumption 2.1 holds and the optimal set X^* of problem (1.1) is nonempty, the sequence $\{x_k\}$ generated by (3.2) converge almost surely to some point $x^* \in X^*$.

Proof. By non-expansive property of projection, for any $x^* \in X^*$ and $k \geq 0$,

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^* - \gamma_k \left(\nabla f_{\eta_k}(x_k) + w_k\right)\|^2 \\ &= \|x_k - x^*\|^2 - 2\gamma_k (\nabla f_{\eta_k}(x_k) + w_k)^\top (x_k - x^*) + \gamma_k^2 \|\nabla f_{\eta_k}(x_k) + w_k\|^2.\end{aligned}$$

By Proposition 2.1, f_{η_k} is convex for any $\eta_k > 0$. Consequently, for any $k \geq 0$,

$$\begin{aligned}\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 &- 2\gamma_k \left(f_{\eta_k}(x_k) - f_{\eta_k}(x^*)\right) - 2\gamma_k w_k^\top (x_k - x^*) \\ &+ \gamma_k^2 \left\|\tilde{g}_{\eta_k}(x_k, v_k, z_k, \xi_k)\right\|^2.\end{aligned}$$

Moreover, by Lemma 2.2 (d.), $f(x_k) \leq f_{\eta_k}(x_k)$ and $f_{\eta_k}(x^*) \leq f(x^*) + L_0\sqrt{n+1}\eta_k$, implying

$$\begin{aligned}\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 &- 2\gamma_k \left(f(x_k) - f(x^*)\right) + 2L_0\sqrt{n+1}\gamma_k\eta_k - 2\gamma_k w_k^\top (x_k - x^*) \\ &+ \gamma_k^2 \left\|\tilde{g}_{\eta_k}(x_k, v_i, z_i, \xi_k)\right\|^2.\end{aligned}$$

Taking conditional expectations of both sides with respect to V_k, Z_k, and $\boldsymbol{\xi}_k$, given $\mathcal{F}_k$, we get

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &\leq \|x_k - x^*\|^2 - 2\gamma_k \left(f(x_k) - f(x^*)\right) + 2L_0\sqrt{n+1}\gamma_k\eta_k \\ &\quad + \gamma_k^2 \mathbb{E}[\|\tilde{g}_{\eta_k}(x_k, V_k, Z_k, \boldsymbol{\xi}_k)\|^2 \mid \mathcal{F}_k] \\ &\leq \|x_k - x^*\|^2 - 2\gamma_k \left(f(x_k) - f(x^*)\right) + 2L_0\sqrt{n+1}\gamma_k\eta_k + \tfrac{4L_0^2 n}{\pi}\gamma_k^2, \qquad (3.3)\end{aligned}$$

where (3.3) follows from Prop. 2.3. Applying the Robbins-Siegmund Lemma (Robbins and Siegmund 1971), for any $x^* \in X^*$, the sequence $\{\|x_k - x^*\|\}$ converges and $\sum_{k=0}^{\infty} \gamma_k \left(f(x_k) - f(x^*)\right) < \infty$ a.s.. The latter implies that $\liminf_{k\to\infty} f(x_k) = f^*$ a.s. in view of the condition $\sum_{k=0}^{\infty} \gamma_k = \infty$. Therefore, there exists a subsequence $\{x_{k_j}\}$ of $\{x_k\}$ such that $f(x_{k_j}) \to f^*$ a.s. By the continuity of f, the sublevel sets $\{x \in \mathbb{R}^n : f(x) \leq c\}$ are closed, implying that all limit points of $\{x_{k_j}\}$ belong to the optimal set X^*. Therefore there exists a subsubsequence of $\{x_k\}$ that converges to some point in X^* a.s. Combining with the a.s. convergence of $\{\|x_k - x^*\|\}$, the entire sequence $\{x_k\}$ converges to a point in X^* a.s. . □

Proposition 3.2 (Rate of convergence). Let Assumptions 2.1(1),(2) hold and the optimal set X^* of (1.1) is nonempty. Suppose $\gamma_k = \eta_k = \frac{1}{\sqrt{n(k+1)}}$, $\bar{x}_K \triangleq \frac{\sum_{k=0}^{K-1} \gamma_k x_k}{\sum_{k=0}^{K-1} \gamma_k}$ and $a_0 = \|x_0 - x^*\|$. Then, for all K,

$$\mathbb{E}\left[f\left(\bar{x}_K\right) - f^*\right] \leq \frac{a_0^2\sqrt{n} + \left(2L_0\frac{\sqrt{n+1}}{\sqrt{n}} + \frac{4L_0^2\sqrt{n}}{\pi}\right)(1+\ln(K))}{4\sqrt{K+1}} \lesssim \mathcal{O}\left(\frac{\sqrt{n}\ln(K)}{\sqrt{K}}\right).$$

Proof. Let us rewrite equation (3.3) from the previous proof.

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 - 2\gamma_k \left(f(x_k) - f(x^*)\right) + 2L_0\sqrt{n+1}\gamma_k\eta_k + \tfrac{4L_0^2 n}{\pi}\gamma_k^2.$$

Taking unconditional expectations, we have

$$2\mathbb{E}[\gamma_k \left(f(x_k) - f(x^*)\right)] \leq \mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\|x_{k+1} - x^*\|^2] + 2L_0\sqrt{n+1}\gamma_k\eta_k + \tfrac{4L_0^2 n}{\pi}\gamma_k^2.$$

Summing over $k = 0, \cdots, K-1$,

$$\sum_{k=0}^{K-1} 2\mathbb{E}[\gamma_k \left(f(x_k) - f(x^*)\right)] \leq \|x_0 - x^*\|^2 + \sum_{k=0}^{K-1} \left(2L_0\sqrt{n+1}\gamma_k\eta_k + \tfrac{4L_0^2 n}{\pi}\gamma_k^2\right). \tag{3.4}$$

Since $\bar{x}_K \triangleq \frac{\sum_{k=0}^{K-1}\gamma_k x_k}{\sum_{k=0}^{K-1}\gamma_k}$ and f is convex, we can apply Jensen's inequality to (3.4) to obtain

$$2\mathbb{E}[(f(\bar{x}_K) - f(x^*))] \leq 2\frac{\sum_{k=0}^{K-1}\mathbb{E}[\gamma_k(f(x_k)-f^*)]}{\sum_{k=0}^{K-1}\gamma_k} \leq \frac{a_0^2+\sum_{k=0}^{K-1}\left(2L_0\sqrt{n+1}\gamma_k\eta_k+\frac{4L_0^2n}{\pi}\gamma_k^2\right)}{\sum_{k=0}^{K-1}\gamma_k}. \tag{3.5}$$

Recall that $\gamma_k = \eta_k = \frac{1}{\sqrt{n(k+1)}}$ for any k, implying that $\sum_{k=0}^{K-1}\gamma_k \geq \frac{1}{\sqrt{n}}(\int_1^{K+1}\frac{1}{\sqrt{x}}dx)$, while $\sum_{k=0}^{K-1}\gamma_k\eta_k = \sum_{k=0}^{K-1}\gamma_k^2 \leq \frac{1}{n}(1+\int_1^K \frac{1}{x}dx)$. Using these inequalities in (3.5), our desired result follows as shown next.

$$2\mathbb{E}[(f(\bar{x}_K) - f(x^*))] \leq \frac{a_0^2+\frac{1}{n}\left(2L_0\sqrt{n+1}+\frac{4L_0^2n}{\pi}\right)\left(1+\int_1^K\frac{1}{x}dx\right)}{\frac{1}{\sqrt{n}}\left(\int_1^{K+1}\frac{1}{\sqrt{x}}dx\right)} \leq \frac{a_0^2\sqrt{n}+\left(2L_0\frac{\sqrt{n+1}}{\sqrt{n}}+\frac{4L_0^2\sqrt{n}}{\pi}\right)(1+\ln(K))}{2\sqrt{K+1}}.$$

□

From the above analysis, one can easily derive the following when $\eta_k = \eta$ for all k.

Corollary 3.1. Let $\eta_k = \eta$ for any k. For any $K > 0$, denote $S_K = \sum_{k=0}^{K-1}\gamma_k$.Then the following holds.

$$\mathbb{E}[f(\bar{x}_K)] - f(x^*) \leq \frac{1}{S_K}\sum_{k=0}^{K-1}\mathbb{E}[\gamma_k\left(f(x_k) - f(x^*)\right)] \leq L_0\sqrt{n+1}\eta + \frac{1}{S_K}\left[\frac{a_0^2}{2} + \frac{2L_0^2 n}{\pi}\sum_{k=0}^{K-1}\gamma_k^2\right].$$

Remark 3.1. Compare Cor. 3.1 to (Nesterov and Spokoiny 2017, Thm 6). In fact, we may also choose $\gamma_k = \frac{R}{L_0\sqrt{nK}}$ as in (Nesterov and Spokoiny 2017) where R is a constant. Similar arguments show that our scheme yields ε-error in $\mathcal{O}(n\varepsilon^{-2})$ iterations, as opposed to $\mathcal{O}(n^2\varepsilon^{-2})$ in (Nesterov and Spokoiny 2017). □

We provide an intermediate lemma for claiming convergence of a suitable recursion (Polyak 1987).

Lemma 3.1. Let $\{v_k\}$ be a sequence of nonnegative random variables, where $\mathbb{E}[v_0] < \infty$, and let $\{u_k\}$ and $\{\mu_k\}$ be deterministic scalar sequences such that the following hold. (i) $\mathbb{E}[v_{k+1} \mid v_0, \ldots, v_k] \leq (1-u_k)v_k + \mu_k$ a.s. for all $k \geq 0$; (ii) $0 \leq u_k \leq 1, \mu_k \geq 0$, for all $k \geq 0$; (iii) $\sum_{k=0}^{\infty} u_k = \infty, \sum_{k=0}^{\infty}\mu_k < \infty, \lim_{k\to\infty}\frac{\mu_k}{u_k} = 0$. Then, $v_k \to 0$ almost surely as $k \to \infty$.

Proposition 3.3 (a.s. convergence for strongly convex f). Suppose Assumption 2.1 holds and the optimal set X^* of problem (1.1) is nonempty. If in addition, $F(\bullet, \xi)$ is μ-strongly convex on X almost surely, the sequence $\{x_k\}$ generated by (3.2) converges almost surely to a unique optimal solution x^*.

Proof. By the non-expansive property of the Euclidean projection, for any $x^* \in X^*$ and $k > 0$,

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^* - \gamma_k\left(\nabla f_{\eta_k}(x_k) + w_k\right)\|^2 \\ &= \|x_k - x^*\|^2 - 2\gamma_k(\nabla f_{\eta_k}(x_k) + w_k)^\top(x_k - x^*) + \gamma_k^2\|\nabla f_{\eta_k}(x_k) + w_k\|^2.\end{aligned}$$

If $F(\bullet, \xi)$ is μ-strongly convex ξ-a.s., then f, defined as $f(x) = \mathbb{E}_{\boldsymbol{\xi}}[F(x, \boldsymbol{\xi})]$, is μ-strongly convex, implying

$$\begin{aligned}\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 &- 2\gamma_k\left(f_{\eta_k}(x_k) - f_{\eta_k}(x^*)\right) - \mu\gamma_k\|x_k - x^*\|^2 \\ &- 2\gamma_k w_k^\top(x_k - x^*) + \gamma_k^2\|\tilde{g}_{\eta_k}(x_k, v_k, z_k, \xi_k)\|^2.\end{aligned}$$

Moreover, by Lemma 2.2 (d.) we have $f(x_k) \leq f_{\eta_k}(x_k)$ and $f_{\eta_k}(x^*) \leq f(x^*) + L_0\sqrt{n+1}\eta_k$. Thus

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\gamma_k\left(f(x_k) - f(x^*)\right) + 2L_0\sqrt{n+1}\gamma_k\eta_k - \mu\gamma_k\|x_k - x^*\|^2 \\ - 2\gamma_k w_k^\top (x_k - x^*) + \gamma_k^2\|\tilde{g}_{\eta_k}(x_k, v_k, z_k, \xi_k)\|^2. \tag{3.6}$$

Since x^* is optimal, $f(x_k) - f(x^*) \geq 0$. Thus, from (3.6), we obtain

$$\|x_{k+1} - x^*\|^2 \leq (1 - \mu\gamma_k)\|x_k - x^*\|^2 + 2L_0\sqrt{n+1}\gamma_k\eta_k - 2\gamma_k w_k^\top (x_k - x^*) + \gamma_k^2\|\tilde{g}_{\eta_k}(x_k, v_k, z_k, \xi_k)\|^2.$$

Taking expectations of both sides with respect to V_k, Z_k and $\boldsymbol{\xi}_k$, given $\mathcal{F}_k$, we obtain

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq (1 - \mu\gamma_k)\|x_k - x^*\|^2 + 2L_0\sqrt{n+1}\gamma_k\eta_k + \gamma_k^2\mathbb{E}\left[\|\tilde{g}_{\eta_k}(x_k, V_k, Z_k, \boldsymbol{\xi}_k)\|^2 \mid \mathcal{F}_k\right] \\ \leq (1 - \mu\gamma_k)\|x_k - x^*\|^2 + 2L_0\sqrt{n+1}\gamma_k\eta_k + \tfrac{4L_0^2 n}{\pi}\gamma_k^2. \tag{3.7}$$

From Assumption 2.1, $\sum_{k=0}^{\infty}\gamma_k = \infty$ and $\sum_{k=0}^{\infty}(2L_0\sqrt{n+1}\gamma_k\eta_k + \frac{4L_0^2 n}{\pi}\gamma_k^2) < \infty$. In addition note, $\lim_{k\to\infty}(2L_0\sqrt{n+1}\gamma_k\eta_k + \frac{4L_0^2 n}{\pi}\gamma_k^2)/\mu\gamma_k \to 0$. Thus by Lemma 3.1 the result holds. □

Lemma 3.2. (Shapiro, Dentcheva, and Ruszczyński 2014) Consider the following recursion: $b_{k+1} \leq (1 - 2c\theta/k)b_k + \frac{1}{2}\theta^2 M^2/k^2$, where θ and M are positive constants, $b_k \geq 0$, and $(1 - 2c\theta) < 0$. Then for $k \geq 1$, we have that $2b_k \leq \frac{\max\left(\frac{\theta^2}{2c\theta - \lfloor 2c\theta \rfloor}M^2, 2b_1\right)}{k}$.

We now prove a ZO variant of an analogous result from (Shapiro, Dentcheva, and Ruszczyński 2014).

Proposition 3.4 (Convergence in mean and rate statement under strong convexity). Under the setting of Proposition 3.3. In addition, suppose X is a bounded set such that $\|x - x^*\| \leq C$ for all $x \in X$ and $\gamma_k = \frac{\theta}{k}$, where $\theta > 1/\mu$. Then the sequence $\{x_k\}$ converges to x^* in mean and

$$\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq \frac{\max\left\{\theta^2(4L_0\sqrt{n+1} + \frac{8L_0^2 n}{\pi})(\mu\theta - 1)^{-1}, C^2\right\}}{k} \lesssim \mathcal{O}\left(\tfrac{n}{k}\right).$$

Proof. Recall the recursion obtained from (3.7):

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] \leq (1 - \mu\gamma_k)\|x_k - x^*\|^2 + 2L_0\sqrt{n+1}\gamma_k\eta_k + \tfrac{4L_0^2 n}{\pi}\gamma_k^2.$$

By setting $\eta_k = \gamma_k = \frac{\theta}{k}$, it follows that by taking unconditional expectations we have that

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] \leq (1 - \mu\gamma_k)\mathbb{E}\left[\|x_k - x^*\|^2\right] + (2L_0\sqrt{n+1} + \tfrac{4L_0^2 n}{\pi})\gamma_k^2.$$

Choosing $\theta > 1/\mu$ and $M^2/2 = (2L_0\sqrt{n+1} + \frac{4L_0^2 n}{\pi})$ in Lemma 3.2, and noting that $a_0^2 = \|x_0 - x^*\|^2 \leq C^2$, we obtain the result for all k. □

4 NUMERICS

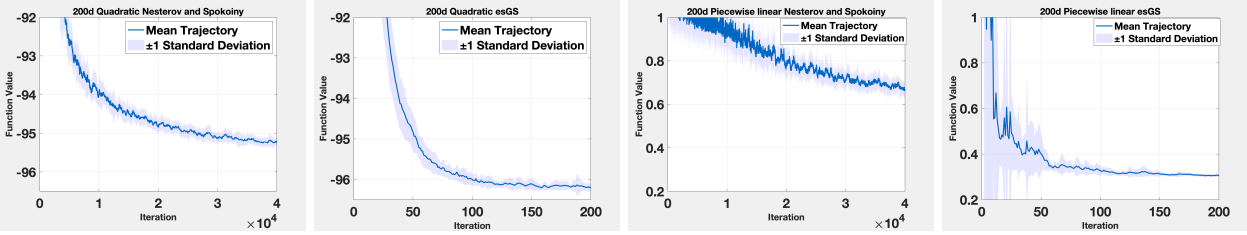
In this section, we provide some preliminary numerics, comparing the proposed scheme with three ZO schemes. To ensure parity, we terminated all algorithms after the same number of calls to the function oracle; specifically, we ran 200 iterations of our method and $200n$ steps of the other methods. We generated 20 replications of each scheme. For each problem considered, we provided a tabulation of error given by the empirical average of either $\mathbb{E}[f(\bar{x}_K) - f^*]$ or $\mathbb{E}[f(x_K) - f^*]$ over the replications while "time" represents the average runtime in `Matlab`. Our first problem is a nonsmooth convex stochastic quadratic optimization problem in which $F(x, \xi) = \frac{1}{2}x^\top\hat{Q}x + b(\xi)^\top x + 0.5\|x\|_1$, $X \triangleq [-1, 1]^n$, $\hat{Q} = Q + W(Q \in \mathbb{S}_{++}^n)$,

Table 3: Strongly Convex Quadratic Objective

	Nesterov and Spokoiny (2017)		Cui et al. (2023)		Spall (1992)		This work	
n	error	time	error	time	error	time	error	time
10	0.0700	0.0096	0.0558	0.0092	0.0562	0.0114	0.0280	0.0044
20	0.1580	0.0390	0.1444	0.0293	0.1334	0.0350	0.0285	0.0092
50	0.4047	0.5408	0.3273	0.4965	0.3432	0.5224	0.0383	0.0436
100	0.6066	3.0719	0.6360	3.0741	0.6337	3.0199	0.0615	0.1178
150	0.8232	11.5668	0.8070	12.6310	0.8680	11.9770	0.0828	0.2913
200	1.0683	39.0659	1.0025	38.0015	1.0228	37.4888	0.0955	0.6088
250	1.1982	69.4609	1.1731	71.8910	1.1298	70.4420	0.1185	0.9991

Table 4: Strongly Convex Piecewise-Linear Objective

	Nesterov and Spokoiny (2017)		Cui et al. (2023)		Spall (1992)		This work	
n	error	time	error	time	error	time	error	time
10	0.0672	0.0055	0.0844	0.0048	0.1008	0.0067	0.0205	0.0024
100	0.2111	0.0938	0.5434	0.0860	0.5292	0.1190	0.0228	0.0241
150	0.3264	0.1709	0.6346	0.1589	0.6507	0.2206	0.0314	0.0739
200	0.4014	0.2697	0.7299	0.2539	0.7627	0.3466	0.0400	0.1217
500	0.9795	1.2842	0.9311	1.3229	0.9448	1.7489	0.1098	0.1431
1000	1.6041	4.9934	1.8245	4.8908	1.7411	6.3167	0.1870	2.1960
2000	2.4267	16.9147	2.4379	17.1208	2.4572	22.4920	0.3253	2.5243
3000	3.1195	38.0462	3.1155	38.4532	3.0184	50.4454	0.5203	5.8374
4000	3.6107	92.6660	3.5918	101.2859	3.5943	133.5990	0.7237	10.1906



(a) (b) (c) (d)

Figures (a) and (b) represent replications of $n = 200$ quadratic problem with Nesterov-Spokoiny and (**esGS**), respectively, while Figures (c) and (d) represent replications of $n = 200$ piecewise-linear problem with Nesterov-Spokoiny and (**esGS**), respectively.

$b(\xi) = b + \xi$, $\boldsymbol{\xi} \sim \mathcal{N}(0, 1)$, and $W = \frac{1}{n} B^\top B$ where the components of B are generated from $\mathcal{N}(0, 0.01)$. All algorithms utilize the same steplength and smoothing parameters $\gamma_k = \eta_k = 1/(\text{norm}(Q)\sqrt{k})$ and a deterministic starting point $x_0 = (5, 5, 5, 5, 5, 0, 0, \cdots)$. In our second problem, we consider a piecewise-linear objective given by $F(x, \xi) = \phi\left(\sum_{i=1}^n \left(\frac{i}{n} + \xi_i\right) x_i\right) + \frac{\mu}{2}\|x\|^2$, $X = \{x \mid \|x\| \leq 1\}$, where $\phi(t) = \max_{1 \leq j \leq m}(v_j + s_j t)$, and $\xi_i \sim \mathcal{N}(0, 1)$. We choose $\{v_j\}_{j=1}^5 = \{0.2, 0.3, 0.6, 0.5, 0.8\}$, $\{s_j\}_{j=1}^5 = \{0.9, 0.2, 0.1, 0.5, 0.5\}$ and $\mu = 1$. All algorithms employ the same steplength and smoothing sequences given by $\gamma_k = \eta_k = 1/k^{0.52}$.

Tables 3 and 4 represent the comparisons across the ZO schemes and our proposed framework with the (**esGS**) estimator for the quadratic and piecewise-linear problem, respectively. Figures (a) – (d) represent the trajectories and the spread across replications across the two problems, comparing our method with that by Nesterov and Spokoiny. **Insights.** (a) *Empirical error.* The empirical error produced by our scheme is nearly one order of magnitude better than competing schemes in the first example, while similar benefits emerge in the piecewise-linear setting. (b) *Computational time.* In terms of computational time, the distinctions are even more pronounced, a consequence of the iteration complexity of our scheme being $\mathcal{O}(n)$ better than its counterparts. For instance, in quadratic settings when $n = 200$, our scheme requires 0.609s while the scheme developed by Nesterov and Spokoiny takes approximately 39s, while producing an empirical error that is more than 10 times worse. The distinctions are significant, if not as pronounced, in the piecewise-linear setting. We further note that the distinctions in computational time and empirical error distinctions increase dramatically with dimension, in alignment with our theory.

5 CONCLUDING REMARKS

Conventional avenues for developing gradient estimators via Gaussian smoothing are afflicted by a key concern in that moment bounds on the estimators grow at the rate of $\mathcal{O}(n^2)$. As a consequence, iteration complexity bounds for computing a ε-solution in convex settings is $\mathcal{O}(n^2\varepsilon^{-2})$. This pronounced dependence on dimension impedes the application of zeroth-order schemes to large-scale settings. Via a simple change-of-variable argument, we develop an exponentially shifted Gaussian smoothing (**esGS**) estimator, reliant

on shifting via exponential random variables, whose moment bound grows at the rate of $\mathcal{O}(n)$ and the resulting iteration complexity is $\mathcal{O}(n\varepsilon^{-2})$. Preliminary numerics support these findings, with our scheme providing far more accurate solutions while requiring a fraction of the computational time, compared with competing schemes.

REFERENCES

Berahas, A. S., L. Cao, K. Choromanski, and K. Scheinberg. 2022. "A theoretical and empirical comparison of gradient approximations in derivative-free optimization". *Foundations of Computational Mathematics* 22(2):507–560.

Conn, A. R., K. Scheinberg, and L. N. Vicente. 2009. *Introduction to derivative-free optimization*. SIAM.

Cui, S., U. V. Shanbhag, and F. Yousefian. 2023. "Complexity guarantees for an implicit smoothing-enabled method for stochastic MPECs". *Mathematical Programming* 198(2):1153–1225.

Folland, G. B. 1999. *Real analysis: modern techniques and their applications*, Volume 40. John Wiley & Sons.

Fu, M. C. 2006. "Chapter 19 Gradient Estimation". In *Simulation*, edited by S. G. Henderson and B. L. Nelson, Volume 13 of *Handbooks in Operations Research and Management Science*, 575–616. Elsevier.

Ghadimi, S. and G. Lan. 2013. "Stochastic first-and zeroth-order methods for nonconvex stochastic programming". *SIAM journal on optimization* 23(4):2341–2368.

Jie, C. and M. C. Fu. 2022. "Using Importance Samping in Estimating Weak Derivative". *arXiv preprint arXiv:2209.13184*.

Lakshmanan, H. and D. Farias. 2008. "Decentralized Recourse Allocation In Dynamic Networks of Agents". *SIAM Journal on Optimization* 19(2):911–940.

Larson, J., M. Menickelly, and S. M. Wild. 2019. "Derivative-free optimization methods". *Acta Numerica* 28:287–404.

Nemirovski, A. and D. Yudin. 1983. *Problem Complexity and Method Efficiency in Optimization*. Wiley -Intersci. Ser. Discrete Math 15 John Wiley New York.

Nesterov, Y. and V. Spokoiny. 2017. "Random gradient-free minimization of convex functions". *Foundations of Computational Mathematics* 17(2):527–566.

Polyak, B. T. 1987. "Introduction to optimization".

Qiu, Y., U. Shanbhag, and F. Yousefian. 2023. "Zeroth-Order Methods for Nondifferentiable, Nonconvex, and Hierarchical Federated Optimization". *Advances in Neural Information Processing Systems* 36.

Robbins, H. and D. Siegmund. 1971. "A convergence theorem for non negative almost supermartingales and some applications". In *Optimizing methods in statistics*, 233–257. Elsevier.

Shanbhag, U. V. and F. Yousefian. 2021. "Zeroth-order randomized block methods for constrained minimization of expectation-valued Lipschitz continuous functions". In *2021 Seventh Indian Control Conference (ICC)*, 7–12. IEEE.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2014. *Lectures on Stochastic Programming: Modeling and Theory*, Volume 16. SIAM.

Spall, J. C. 1992. "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation". *IEEE transactions on automatic control* 37(3):332–341.

Spall, J. C. 2005. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons.

Steklov, V. A. 1907. "Sur les expressions asymptotiques decertaines fonctions définies par les équations différentielles du second ordre et leers applications au problème du dévelopement d'une fonction arbitraire en séries procédant suivant les diverses fonctions". *Comm. Charkov Math. Soc.* 2(10):97–199.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2012. "On Stochastic Gradient and Subgradient Methods with adaptive steplength sequences". *Automatica* 48(1):56–67.

AUTHOR BIOGRAPHIES

MINGRUI WANG is a PhD student in the Department of Industrial and Manufacturing Engineering at Pennsylvania State University. His email address is mvw5822@psu.edu.

PRAKASH CHAKRABORTY is an Assistant Professor in the Department of Industrial and Manufacturing Engineering at Pennsylvania State University. His website is https://prakashchakraborty.github.io and his email is prakashc@psu.edu.

UDAY V. SHANBHAG is a Professor in the Department of Industrial and Operations Engineering at University of Michigan in Ann Arbor. His website is https://udaybag2.github.io and his email is udaybag@umich.edu.