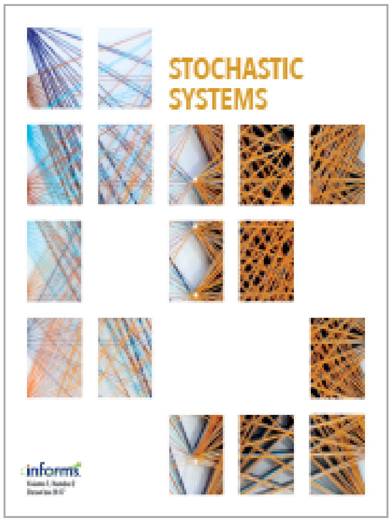


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Smoothed Variable Sample-Size Accelerated Proximal Methods for Nonsmooth Stochastic Convex Programs

Afrooz Jalilzadeh, Uday Shanbhag, Jose Blanchet, Peter W. Glynn

informs
https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 12, No. 4, December 2022, pp. 373–410
ISSN 1946-5238 (online)

Smoothed Variable Sample-Size Accelerated Proximal Methods for Nonsmooth Stochastic Convex Programs

Afrooz Jalilzadeh,[a] Uday Shanbhag,[b,*] Jose Blanchet,[c] Peter W. Glynn[c]

[a] Department of Systems and Industrial Engineering, University of Arizona, Tucson, Arizona 85721; [b] Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, Pennsylvania 16802; [c] Management Science and Engineering, Stanford University, Stanford, California 94305
*Corresponding author

Contact: afrooz@arizona.edu, https://orcid.org/0000-0002-3734-1082 (AJ); udaybag@psu.edu, https://orcid.org/0000-0001-5869-9561 (US); jblanche@stanford.edu (JB); glynn@stanford.edu (PWG)

Received: April 17, 2019
Revised: May 12, 2020; August 16, 2021; February 26, 2022
Accepted: March 7, 2022
Published Online in Articles in Advance: September 8, 2022

https://doi.org/10.1287/stsy.2022.0095

Abstract. We consider the unconstrained minimization of the function F, where $F = f + g$, f is an expectation-valued nonsmooth convex or strongly convex function, and g is a closed, convex, and proper function. (I) Strongly convex f. When f is μ-strongly convex in x, traditional stochastic subgradient schemes (SSG) often display poor behavior, arising in part from noisy subgradients and diminishing steplengths. Instead, we apply a variable sample-size accelerated proximal scheme (VS-APM) on F, the Moreau envelope of F; we term such a scheme as (mVS-APM) and in contrast with (SSG) schemes, (mVS-APM) utilizes constant steplengths and increasingly exact gradients. We consider two settings. (a) Bounded domains. In this setting, (mVS-APM) displays linear convergence in inexact gradient steps, each of which requires utilizing an inner (prox-SSG) scheme. Specifically, (mVS-APM) achieves an optimal oracle complexity in prox-SSG steps of $\mathcal{O}(1/\epsilon)$ with an iteration complexity of $\mathcal{O}(\log(1/\epsilon))$ in inexact (outer) gradients of F to achieve an ϵ-accurate solution in mean-squared error, computed via an increasing number of inner (stochastic) subgradient steps; (b) Unbounded domains. In this regime, under an assumption of state-dependent bounds on subgradients, an unaccelerated variant (mVS-APM) is linearly convergent where increasingly exact gradients $\nabla_x F(x)$ are approximated with increasing accuracy via (SSG) schemes. Notably, (mVS-APM) also displays an optimal oracle complexity of $\mathcal{O}(1/\epsilon)$; (II) Convex f. When f is merely convex but smoothable, by suitable choices of the smoothing, steplength, and batch-size sequences, smoothed (VS-APM) (or sVS-APM) achieves an optimal oracle complexity of $\mathcal{O}(1/\epsilon^2)$ to obtain an ϵ-optimal solution. Our results can be specialized to two important cases: (a) Smooth f. Since smoothing is no longer required, we observe that (VS-APM) admits the optimal rate and oracle complexity, matching prior findings; (b) Deterministic nonsmooth f. In the nonsmooth deterministic regime, (sVS-APM) reduces to a smoothed accelerated proximal method (s-APM) that is both asymptotically convergent and optimal in that it displays a complexity of $\mathcal{O}(1/\epsilon)$, matching the bound provided by Nesterov in 2005 for producing ϵ-optimal solutions. Finally, (sVS-APM) and (VS-APM) produce sequences that converge almost surely to a solution of the original problem.

Funding: The first and second authors would like to acknowledge support from NSF CMMI-1538605, DOE ARPA-E award DE-AR0001076, and the Gary and Sheila Bello Chair funds.

Keywords: stochastic approximation • nonsmooth optimization • stochastic optimization • smoothing

1. Introduction

We consider the following stochastic nonsmooth convex optimization problem:

$$\min_{x\in\mathbb{R}^n} F(x), \text{ where } F(x) \triangleq f(x) + g(x), \tag{1}$$

where $f(x) \triangleq \mathbb{E}[\tilde{f}(x, \xi(\omega))]$, $\xi : \Omega \to \mathbb{R}^o, \tilde{f} : \mathbb{R}^n \times \mathbb{R}^o \to \mathbb{R}$; g is a closed, convex, and proper deterministic function with an efficient proximal evaluation; $(\Omega, \mathcal{H}, \mathbb{P})$ denotes the associated probability space; and $\mathbb{E}[\bullet]$ denotes the expectation with respect to the probability measure $\mathbb{P}$. Throughout, we refer to $\tilde{f}(x, \xi(\omega))$ by $\tilde{f}(x, \omega)$, whereas

$\tilde{F}(x,\omega) \triangleq \tilde{f}(x,\omega) + g(x)$. We consider settings in which $\tilde{f}(\cdot,\omega)$ is nonsmooth strongly convex/convex in x for every ω, generalizing the focus beyond the *structured nonsmooth* setting in which the "stochastic part" is smooth. Specifically, structured nonsmooth problems require minimizing $f(x) + g(x)$, where f is smooth, whereas g is nonsmooth with an efficient prox evaluation (allows for capturing constrained problems over closed and convex sets).

Among the earliest avenues for resolving (1) is stochastic approximation (Robbins and Monro 1951, Kushner and Yin 2003), and it is proven to be effective on a breadth of stochastic computational problems, including convex optimization problems. Polyak and Juditsky (1992) develop an averaging scheme in convex differentiable settings, deriving the optimal convergence rate of $\mathcal{O}(1/\sqrt{K})$ under classic assumptions, where k is the number of iterations. Among the cleanest of early complexity requirements for the minimization of expectation-valued μ-strongly convex and convex functions over a closed and convex set X are given as $\left(\max\left\{\frac{M^2}{\mu^2}, \|x_0 - x^*\|^2\right\}\frac{1}{\epsilon}\right)$ (to ensure that $\mathbb{E}[\|x_k - x^*\|^2] \le \epsilon$) and $\mathcal{O}\left(\frac{MD_X}{\epsilon^2}\right)$ (to ensure that the expected optimality gap is less than ϵ), respectively, where $S(x,\omega)$ denotes a measurable selection from $\partial_x \tilde{f}(x,\omega)$, $\sup_{x\in X}\mathbb{E}[\|S(x,\omega)\|^2] \le M^2$, and $D_X \triangleq \max_{x\in X}\|x_0 - x\|$. Of these, the former is presented by Shapiro et al. (2009), whereas the latter is the result of an optimal robust constant step length stochastic approximation scheme suggested by Nemirovski et al. (2009). When f is both L-smooth and μ-strongly convex, an improved complexity requirement (from a constant factor standpoint) of $\mathcal{O}\left(\sqrt{\frac{L\|x_0 - x^*\|^2}{\epsilon}} + \frac{\nu^2}{\mu\epsilon}\right)$ is provided by Ghadimi and Lan (2013). This contrasts sharply with the deterministic regime in which $\mathcal{O}(\log(1/\epsilon))$ and $\mathcal{O}(1/\sqrt{\epsilon})$ steps are required in smooth strongly convex and smooth convex regimes to compute an ϵ-accurate solution (ϵ-solution in terms of mean squared error) and ϵ-optimal solution (ϵ-solution in terms of expected suboptimality), respectively. In structured nonsmooth regimes, there is an effort to employ the stochastic generalization of an accelerated proximal gradient method to minimize $f + g$ when f is smooth. Reliant on a first order oracle that produces a sampled gradient $\nabla_x \tilde{f}(x,\omega)$ and given an x_0, our proposed variable sample-size accelerated proximal gradient scheme (VS-APM) (also see Ghadimi and Lan 2016, Jofré and Thompson 2017) is stated as follows in which the true gradient is replaced by a sample average $(\nabla_x f(x_k) + \bar{w}_{k,N_k})$ with batch size N_k.

$$\begin{aligned} y_{k+1} &:= \mathbf{P}_{\gamma_k g}(x_k - \gamma_k(\nabla_x f(x_k) + \bar{w}_{k,N_k})) \\ x_{k+1} &:= y_{k+1} + \beta_k(y_{k+1} - y_k), \end{aligned} \tag{2}$$

where $\bar{w}_{k,N_k} \triangleq \frac{\sum_{j=1}^{N_k}(\nabla_x \tilde{f}(x_k,\omega_{j,k}) - \nabla_x f(x_k))}{N_k}$, $\mathbf{P}_{\eta g}(y) \triangleq \arg\min_x\left\{\frac{1}{2}\|x - y\|^2 + \frac{1}{2\eta}g(x)\right\}$, γ_k, and β_k are suitably defined step lengths. Our approach produces linearly convergent iterates in strongly convex regimes and achieves an iteration complexity of $\mathcal{O}(1/K^2)$ in merely convex and smooth regimes, where K is the total number of iterations, matching the deterministic results seen in the work by Beck and Teboulle (2009) and Nesterov (1983). The avenue represented by (2) has two key distinctions: (i) increasingly exact gradients through increasing batch sizes N_k of sampled gradients, allowing for progressive variance reduction, and (ii) larger (nondiminishing) step sizes in accordance with deterministic accelerated schemes. Collectively, (i) and (ii) allow for recovering fast (i.e., deterministic) convergence rates (in an expected value sense) when N_k grows sufficiently fast. Additionally, such schemes have a more muted reliance on the condition number $\kappa = L/\mu$ (in μ-strongly convex and L-smooth regimes); specifically, in accelerated schemes, such dependence reduces to $\sqrt{\kappa}$ in comparison with κ in unaccelerated counterparts (cf. Nesterov 2014).

1.1. Prior Research

1.1.1. Stochastic Gradient Schemes. In nonsmooth convex stochastic optimization problems, Nemirovski et al. (2009) derive an optimal rate of $\mathcal{O}(1/\sqrt{K})$ in terms of expected suboptimality via an optimal constant step length (also see Shamir and Zhang 2013), whereas in strongly convex regimes, they derive a rate of $\mathcal{O}(1/K)$ in a mean squared sense. Structured nonsmooth problems (or composite problems) as defined by (1) are examined extensively (cf. Ghadimi and Lan 2012, Lan 2012), and rates of $\mathcal{O}(L/K^2 + 1/\sqrt{K})$ and $\mathcal{O}(L/K + 1/\sqrt{K})$ are developed by Dang and Lan (2015) via a mirror-descent framework for strongly convex and convex problems with L-smooth objectives, respectively. In related work, Devolder et al. (2014) derive oracle complexities with a deterministic oracle of fixed inexactness, which is extended to a stochastic oracle by Dvurechensky and Gasnikov (2016).

Randomized smoothing techniques are also employed by Yousefian et al. (2012) together with recursive step lengths (see Newton et al. 2018 for a review).

1.1.2. Variance Reduction. In strongly convex regimes (without acceleration), a linear rate of convergence in expected error is first shown for variance-reduced gradient methods by Shanbhag and Blanchet (2015) and revisited by Jofré and Thompson (2017), whereas similar rates are provided for extragradient methods by Jalilzadeh and Shanbhag (2016); the accelerated counterpart (VS-APM) mutes the dependence on κ, improving the bound to $\mathcal{O}(\sqrt{L/\mu}\log(1/\epsilon))$. In smooth regimes, an accelerated scheme is first presented by Ghadimi and Lan (2016), in which every iteration requires two prox evaluations, admitting the optimal iteration complexity and oracle complexity of $\mathcal{O}(1/\sqrt{\epsilon})$ and $\mathcal{O}(1/\epsilon^2)$, respectively. Jofré and Thompson (2017) extend this scheme to allow for state-dependent noise. An extragradient-based variable sample-size framework is suggested by Jalilzadeh and Shanbhag (2016) with a rate of $\mathcal{O}(1/K)$.

1.1.3. Smoothing Techniques for Nonsmooth Problems. For a subclass of deterministic nonsmooth problems, Nesterov (2005b) proves that an ϵ-optimal solution is computable in $\mathcal{O}(1/\epsilon)$ gradient steps by applying an accelerated method to a smoothed problem (primal smoothing with fixed smoothing parameter). Subsequently, Nesterov (2005a) considers primal–dual smoothing in deterministic regimes (extended to composite problems by Tran-Dinh et al. 2018) with a diminishing smoothing parameter, leading to rates of $\mathcal{O}(1/K^2)$ and $\mathcal{O}(1/K)$ for strongly convex and convex deterministic problems, respectively (also see Devolder et al. 2012, Boţ and Hendrich 2013). Adaptive smoothing, considered by Tran-Dinh (2017), is shown to have an iteration complexity of $\mathcal{O}(1/\epsilon)$, whereas Ouyang and Gray (2012) show that smoothing-based minimization of $\mathbb{E}[\tilde{f}(x,\omega)]+\mathbb{E}[\tilde{g}(x,\omega)]$ leads to rates $\mathcal{O}(1/K)$ and $\mathcal{O}(1/\sqrt{K})$ when $\tilde{g}(\cdot,\omega)$ is nonsmooth for almost every (a.e.) ω, whereas $\tilde{f}(\cdot,\omega)$ is either strongly convex or merely convex for a.e. ω (extended by Zhong and Kwok 2014).[1]

1.2. Gaps and Contributions

Unfortunately when $\tilde{f}(\cdot,\omega)$ is a nonsmooth strongly convex/convex function, stochastic subgradient schemes (subsequently defined in (SSG)), while a de facto standard, generally display poor empirical behavior because they utilize diminishing step lengths and noisy gradients. We develop two distinct avenues for combining smoothing with acceleration and variance reduction in strongly convex and convex regimes that ameliorate these concerns while achieving optimal rates.

1.2.1. mVS-APM for Strongly Convex Nonsmooth *f*. In Section 2, our smoothing framework is reliant on a variable sample-size accelerated proximal method (VS-APM), which requires smoothness of f while displaying linear convergence and optimal oracle complexity. In two distinct settings, we propose applying VS-APM (or an unaccelerated variant) on the Moreau envelope of F, denoted by F_η, where F_η is $\frac{1}{\eta}$-smooth and retains the minimizers of F.

1.2.1.1. Compact Domains. Under the assumption that the domain of g is bounded and $\mathbb{E}[\|S(x,\omega)\|^2]\leq M^2$ for all $x\in\mathbb{R}^n$, where $S(x,\omega)$ is a measurable selection from $\partial\tilde{f}(x,\omega)$; i.e. $S(x,\omega)\in\partial\tilde{f}(x,\omega)$, we show that (mVS-APM) produces a linearly convergent sequence with an iteration complexity of $\mathcal{O}(\log(1/\epsilon))$ in inexact gradient steps $\nabla_x F_\eta(x_k)$, where increasingly exact gradients $\nabla_x F_\eta(x)$ are obtained by employing a (prox-SSG) scheme. In particular, our variance-reduced scheme endeavors to get increasingly exact gradients by progressively reducing the bias in the gradients (because we utilize an increasing number of SSG steps); such a benefit does not appear in a naive implementation of SSG. Moreover, the overall complexity in subgradient evaluations (and consequently sample or oracle complexity) is $\mathcal{O}(1/\epsilon)$, matching the optimal complexity in subgradient steps achieved by (SSG) schemes.

1.2.1.2. Unbounded Domains. When domains are possibly unbounded, assuming that $\mathbb{E}[\|S(x,\omega)\|^2]\leq\bar{M}^2\|x\|^2+M^2$, where $S(x,\omega)\in\partial\tilde{F}(x,\omega)$, the proposed (unaccelerated) variable sample-size proximal method (mVS-PM) achieves an iteration complexity of $\mathcal{O}(\log(1/\epsilon))$ (in gradient steps with $\nabla_x F_\eta$) and overall complexity in subgradient steps of $\mathcal{O}(1/\epsilon)$.

1.2.2. sVS-APM for Convex Nonsmooth *f*. In this setting, in Section 3, we develop an iterative smoothing-based extension of VS-APM, denoted by sVS-APM. By reducing the smoothing and step length parameters at a suitable rate, $\mathbb{E}[F(y_K) - F(x^*)] \leq \mathcal{O}(1/K)$. Notably sVS-APM produces asymptotically accurate solutions (unlike the scheme by Nesterov (2005b), which produces approximate solutions via a fixed smoothing parameter) and is characterized by the optimal oracle complexity of $\mathcal{O}(1/\epsilon^2)$. When f is convex and smooth, we may specialize these results to obtain an optimal rate of $\mathcal{O}(1/K^2)$ and display an optimal sample complexity of $\mathcal{O}(1/\epsilon^2)$. When f is deterministic but nonsmooth, s-APM matches the rate by Nesterov (2005b) but produces asymptotically exact solutions. Additionally, we prove that, for suitable (but distinct) choices of step length and smoothing sequences, sVS-APM and VS-APM produce sequences that converge almost surely (a.s.) to a solution of (1), a convergence statement that was unavailable thus far, matching deterministic results by Orabona et al. (2012) and Boţ and Hendrich (2015) that leverage Moreau smoothing; we provide a result for (α, β)-smoothable functions (see Beck 2017).

1.2.3. Notation. A vector x is assumed to be a column vector, whereas $\|x\|$ denotes the Euclidean vector norm, that is, $\|x\| = \sqrt{x^T x}$. $\mathbf{P}_{\eta g}(x)$ denotes the prox with respect to g with prox parameter $\frac{1}{2\eta}$ at x. $\mathbb{E}[z]$ denotes the expectation of a random variable z. We let X^* denote the set of optimal solutions of (1).

2. Nonsmooth Strongly Convex Problems

In this section, we develop rate and complexity analysis for nonsmooth strongly convex optimization problems via techniques that combine smoothing, acceleration, and variance reduction. In Section 2.1, we review a linearly convergent variance-reduced accelerated proximal scheme (VS-APM) for smooth stochastic convex optimization; this scheme serves as our subproblem solver. In Section 2.2, we present a Moreau-smoothed variant of VS-APM, referred to as (mVS-APM), which relies on minimizing the Moreau envelope F_η of the strongly convex nonsmooth function F by VS-APM. In Section 2.3, we then derive rate and complexity guarantees for (mVS-APM), where $\nabla_x F_\eta$ is approximated with increasing accuracy by a stochastic subgradient (SSG) scheme. Finally, in Section 2.4, we derive analogous statements when applying an unaccelerated variable sample-size proximal method (mVS-PM) under possibly non-compact domains and under a (weaker) state-dependent bound on the subgradient (see Table 1 for a summary of findings).

2.1. Background on VS-APM

Consider (1), in which f, g, and the initial point x_0 satisfy the following assumption.

Assumption 1. *(i) f is a μ-strongly convex function, and g is a closed, convex, and proper deterministic function. (ii) There exist $C, D > 0$ such that $\mathbb{E}[\|x_0 - x^*\|^2] \leq C$ and $\mathbb{E}[\|F(x_0) - F(x^*)\|] \leq D$, where $F(x) \triangleq f(x) + g(x)$ and x^* solves (1).*

In a subset of regimes, we impose an L-smoothness assumption on f.

Assumption 2. *The function f is continuously differentiable with a Lipschitz continuous gradient with constant L; i.e., $\|\nabla_x f(x) - \nabla_x f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$.*

We utilize a variable sample-size accelerated proximal scheme (VS-APM) as defined in Algorithm 1, which can process such problems and differs from a standard accelerated proximal method in that we employ an

Table 1. Comparison of Schemes in Nonsmooth (NS) and Strongly Convex Regimes in Terms of Convergence Rate and Complexity of Iterations, Proximal Evals., and Oracle Evaluations ($\kappa = L/\mu$), Where $\rho \in (0, 1)$

Smooth	Conv. rate iter. comp.	Prox. eval. oracle comp.	Comments
VS-APM (2.1) f is L-smooth	$\mathcal{O}(\rho^k)$ $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$	$\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$ $\mathcal{O}(\kappa/\epsilon)$	Optimal rate and complexity
Nonsmooth	Conv. rate iter. comp.	Oracle comp.	Comments
(mVS-APM) (2.3) dom(g) is bounded; $\mathbb{E}[\|R(x,\omega)\|^2] \leq M^2$ $\forall R(x,\omega) \in \partial\tilde{f}(x,\omega)$	$\mathcal{O}(\rho^k)$ $\mathcal{O}(\log(1/\epsilon))$	$\mathcal{O}(1/\epsilon)$	Minimize Moreau env. $F_\eta(x)$ via VS-APM Nondiminishing outer steps; Approx. $\nabla_x F_\eta$ by (prox-SSG) with increasing exactness

mVS-PM (2.4)
$\mathbb{E}[\|S(x,\omega)\|^2] \leq \bar{M}^2\|x\|^2 + M^2$
$\forall S(x,\omega) \in \partial\tilde{f}(x,\omega)$
	$\mathcal{O}(\rho^k)$
	$\mathcal{O}(\log(1/\epsilon))$
		$\mathcal{O}(1/\epsilon)$	Minimize Moreau env. $F_\eta(x)$ via (VS-PM)
			Nondminishing outer steps;
			Approx. $\nabla_x F_\eta(x)$ by **(SSG)** with increasing exactness;

inexact gradient $\nabla_x f(x_k) + \bar{w}_{k,N_k}$, where the bound on the second moment of $\bar{w}_{k,N_k} \triangleq \nabla_x f(x_k) - \frac{\sum_{k=0}^{N_k} \nabla_x f(x_k, \omega_k)}{N_k}$ is diminishing with k, a consequence of using variance reduction.

Algorithm 1 (Variable Sample-Size Accelerated Proximal Method)

```
(0) Given $x_0, y_0 = x_0, \kappa$, and positive sequences $\{\gamma_k, N_k\}$, set $\lambda_1 \in (1, \sqrt{\kappa}\,]$, $k := 1$.
(1) $y_{k+1} := \mathbf{P}_{\gamma_k g}(x_k - \gamma_k(\nabla_x f(x_k) + \bar{w}_{k,N_k}))$.
(2) $\lambda_{k+1} := \frac{1}{2}\left(1 - \frac{\lambda_k^2}{\kappa} + \sqrt{\left(1 - \frac{\lambda_k^2}{\kappa}\right)^2 + 4\lambda_k^2}\right)$.
(3) $x_{k+1} := y_{k+1} + \left(\frac{(\lambda_k - 1)\left(1 - \frac{1}{4\kappa}\lambda_{k+1}\right)}{\left(1 - \frac{1}{4\kappa}\right)\lambda_{k+1}}\right)(y_{k+1} - y_k)$.
(4) If $k > K$, then stop; else $k := k + 1$; return to step 1.
```

We outline the assumptions on the first and second moments of $\bar{w}_k$.

Assumption 3. *(i) Conditional boundedness of second moments: there exists $\nu > 0$ such that $\mathbb{E}[\|\bar{w}_k\|^2 \mid \mathcal{H}_k] \le \frac{\nu^2}{N_k}$ holds a.s. for all k and $\mathcal{H}_k \triangleq \sigma\{x_0, x_1, \ldots, x_{k-1}\}$. (ii) Conditional unbiasedness of first moments: $\mathbb{E}[w_k \mid \mathcal{H}_k] = 0$ holds a.s., where $w_k \triangleq \nabla_x f(x_k, \omega_k) - \nabla_x f(x_k)$.*

VS-APM can be shown to achieve linear convergence akin to that by Nesterov (2014) by combining inexact gradients in which the inexactness is driven to zero by increasing the sample-size in estimating the gradients. This avenue also allows for achieving the optimal oracle complexity to obtain an ϵ-accurate solution. These differences lead to a slightly modified set of update rules in contrast with that developed by Nesterov (2014) and require that $\gamma_k = 1/2L$ rather than $1/L$. This scheme serves as a subproblem solver in subsequent sections, and we now state a lemma and the associated complexity statement of VS-APM. The proof is similar to that by Nesterov (2014) and is in the appendix. Importantly, this scheme allows for a possibly biased estimate of the gradient.

Lemma 1. *Suppose Assumptions 1–3(i) hold. Consider the iterates generated by VS-APM, where $\gamma_k = \frac{1}{2L}$ for all $k \ge 0$, $\kappa = \frac{L}{\mu}$, and $\bar{\alpha} = \frac{1}{2\sqrt{\kappa}}$. Then, the following holds for all K.*

$$\mathbb{E}[F(y_K) - F^*] \le \left(D + \frac{\mu}{2}C^2\right)(1 - \bar{\alpha})^{K-1} + \sum_{i=0}^{K-1} \frac{(1 - \bar{\alpha})^i \left(\frac{2}{L} + \frac{1}{\mu}\right)\nu^2}{N_{k-i}} + \sum_{i=0}^{K-2} \frac{(1 - \bar{\alpha})^{i+1}\left(\frac{2}{L} + \frac{1}{\mu}\right)\nu^2}{N_{k-i-1}}. \tag{3}$$

The following theorem characterizes the iteration and oracle complexity of VS-APM.

Theorem 1 (Rate and Oracle Complexity of VS-APM Under Biased Oracles). *Suppose Assumptions 1–3(i) hold. Consider the iterates generated by VS-APM, where $\gamma_k \triangleq \frac{1}{2L}$, $N_k \triangleq \lfloor \rho^{-k} \rfloor$, $\theta \triangleq \left(1 - \frac{1}{2\sqrt{\kappa}}\right)$, $\rho \triangleq \left(1 - \frac{1}{2a\sqrt{\kappa}}\right)$ for all $k \ge 0$ and $a > 2$.*

i. *For all K, we have that $\mathbb{E}[F(y_K) - F^*] \le \tilde{C}\rho^{K-1}$ where $\tilde{C} \triangleq \left(D + \frac{\mu}{2}C^2\right) + \frac{4\nu^2}{\mu} + \frac{2\nu^2\sqrt{\kappa}}{\mu}$.* (4)

In addition, VS-APM needs $\mathcal{O}(\sqrt{\kappa}\log(\frac{1}{\epsilon}))$ steps to obtain an ϵ-accurate solution, that is, $\mathbb{E}[F(y_{K+1}) - F^*] \le \epsilon$.

ii. *To compute an ϵ-accurate solution, $\sum_{k=1}^{K} N_k \le \left(\left(D + \frac{\mu C^2}{2}\right) + \frac{4\nu^2}{\mu} + \frac{2\nu^2\sqrt{\kappa}}{\mu}\right)\mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$.*

We know of no other result for variance-reduced accelerated proximal schemes in strongly convex (or even convex) smooth regimes that allows for biased oracles. For instance, Schmidt et al. (2011) impose unbiasedness in strongly convex regimes. Next, we show that adding the unbiasedness requirement, that is, $\mathbb{E}[w_k \mid \mathcal{H}_k] = 0$ a.s. for all k improves the constants in these bounds.

Corollary 1 (Rate and Oracle Complexity of VS-APM Under Unbiased Oracles). *Suppose Assumptions 1–3(i,ii) hold. Consider the iterates generated by VS-APM, where $\gamma_k \triangleq \frac{1}{2L}$, $N_k \triangleq \lfloor \rho^{-k} \rfloor$, $\theta \triangleq \left(1 - \frac{1}{2\sqrt{\kappa}}\right)$, $\rho \triangleq \left(1 - \frac{1}{2a\sqrt{\kappa}}\right)$ for all $k \ge 0$ and $a > 2$.*

i. *For all K, we have that $\mathbb{E}[F(y_K) - F^*] \le \tilde{C}\rho^{K-1}$ where $\tilde{C} \triangleq \left(D + \frac{\mu}{2}C^2\right) + \frac{4\nu^2}{\mu}$.* (5)

In addition, VS-APM needs $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$ steps to obtain an ϵ-accurate solution.

ii. *To compute an* ϵ*-accurate solution,* $\sum_{k=1}^{K} N_k \le \left(\left(D+\frac{\mu C^2}{2}\right)+\frac{4\nu^2}{\mu}\right)\mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$.

The application of VS-APM is afflicted by the need for the L-smoothness of f as well as the availability of L, the Lipschitz constant. Naturally, in many settings, the problem may not be smooth, and even if L-smoothness holds, an estimate of L may be unavailable. Consequently, to broaden the reach of the scheme, an approach that obviates the need for L or the imposition of the smoothness assumption is necessitated. This prompts the subsequent smoothed scheme, (mVS-APM). This scheme can always be implemented if the strong convexity modulus (denoted by μ) is known but the function is either nonsmooth or smooth with an unknown Lipschitz constant L. It is worth noting that estimating μ is challenging, and if μ is indeed unknown, then in Section 3, we introduce an iteratively smoothed VS-APM (sVS-APM) method that necessitates neither the knowledge of the Lipschitz constant L nor the smoothness of f nor the strong convexity modulus μ.

2.2. A Moreau-Smoothed Inexact Accelerated Framework (mVS-APM)

When $\tilde f(\cdot,\omega)$ is a nonsmooth strongly convex function for almost every ω, then the standard approach lies in utilizing stochastic subgradient schemes (SSG) in which convergence relies on choosing square-summable but nonsummable step length sequences. The choice of the parameters in such sequences can have a debilitating impact on performance in some settings (cf. Shapiro et al. 2009). Specifically, choosing γ_k as $1/(\mu\lambda)$ minimizes the mean squared error but overestimating μ can have catastrophic impact as seen in Shapiro et al. (2009, section 5.9, example 5.36). More generally, such choices are often characterized by poor asymptotic behavior, a consequence that arises in part from the diminishing nature of step length sequences and the noisy subgradients. We consider a distinct avenue reliant on minimizing the Moreau envelope of a closed, convex, and proper function F (cf. Moreau 1965), denoted by F_η and defined next.

$$F_\eta(x) \triangleq \min_u \left\{F(u)+\frac{1}{2\eta}\|u-x\|^2\right\}. \tag{6}$$

Notably, this smoothing retains the minimizer of F when F is strongly convex.

Lemma 2 (Planiden and Wang 2016, Lemma 2.19). *Consider a convex, closed, and proper function F and its Moreau envelope $F_\eta(x)$. Then, the following hold: (i) x^* is a minimizer of F over $\mathbb{R}^n$ if and only if x^* is a minimizer of F_η; (ii) F is μ-strongly convex on $\mathbb{R}^n$ if and only if F_η is $\bar\mu$-strongly convex on $\mathbb{R}^n$, where $\bar\mu \triangleq \frac{\mu}{\eta\mu+1}$.*

Consequently, we minimize the $\bar\mu$-strongly convex and $\frac{1}{\eta}$-smooth function F_η, which is not necessarily an easy task because computing $\nabla_x F_\eta(x)$ necessitates solving nonsmooth stochastic optimization problems. We adopt an inexact accelerated proximal scheme for minimizing F_η. But, in contrast with (SSG) schemes applied to minimizing F, we control the smoothness of the outer problem by choosing η and utilize (i) larger nondiminishing step lengths, (ii) acceleration, and (iii) increasingly exact gradients, all of which are distinct from (SSG), as shown next.

$$\overbrace{\left[\begin{array}{l} x_{k+1} := x_k - \gamma_k u_k \\ u_k \in \partial \tilde F(x_k,\omega_k). \end{array} \qquad (\mathbf{SSG})\right]}^{\gamma_k \to 0,\ u_k \text{ is noisy subgradient.}} \qquad \overbrace{\left[\begin{array}{l} y_{k+1} := x_k - \gamma_k(\nabla_x F_\eta(x_k)+\bar w_{k,N_k}), \\ x_{k+1} := y_{k+1}+\beta_k(y_{k+1}-y_k). \end{array} \qquad (\mathbf{mVS-APM})\right]}^{\text{Non-diminishing } \gamma_k \text{ + increasingly exact gradients + Acceleration}}$$

Importantly, $\nabla_x F_\eta(x_k)+\bar w_{k,N_k}$ represents an *approximation* of the gradient of the Moreau envelope. The true gradient of the Moreau envelope F_η is defined as $\nabla_x F_\eta(x) = \frac{1}{\eta}(x-\mathrm{prox}_{\eta F}(x))$, where

$$\mathrm{prox}_{\eta F}(x) \triangleq \arg\min_u \left\{F(u)+\frac{1}{2\eta}\|x-u\|^2\right\}. \tag{7}$$

But $\mathrm{prox}_{\eta F}(x)$ cannot be computed in finite time because F is a nonsmooth, expectation-valued convex function. Instead, via stochastic approximation, we compute an approximate solution of $\mathrm{prox}_{\eta F}(x)$ denoted by $\widehat{\mathrm{prox}}_{\eta F}(x)$, implying that the inexact gradient of $F_\eta(x)$ is given by $\frac{1}{\eta}(x-\widehat{\mathrm{prox}}_{\eta F}(x))$. In Algorithm 1, the inexact gradient

$\nabla_x F_\eta(x_k) + \bar{w}_{k,N_k}$ is defined as

$$\nabla_x F_\eta(x_k) + \bar{w}_{k,N_k} = \frac{1}{\eta}(x_k - \mathrm{prox}_{\eta F}(x_k)) + \overbrace{\frac{1}{\eta}(\mathrm{prox}_{\eta F}(x_k) - \widehat{\mathrm{prox}}_{\eta F}(x_k))}^{\triangleq \bar{w}_{k,N_k}}. \tag{8}$$

We now proceed to develop (mVS-APM) for compact domains in Section 2.3 and then weaken compactness requirements in Section 2.4 for an unaccelerated variant.

2.3. Linear Convergence of (mVS-APM): Compact Domains

When $F(x) = \mathbb{E}[\tilde{f}(x,\omega)] + g(x)$, $\mathrm{prox}_{\eta F}(x)$, defined as (7), is generally unavailable in closed form and requires solving a strongly convex nonsmooth stochastic optimization problem exactly. Instead, one may solve (6) inexactly using (prox-SSG), a slightly extended variant of (SSG) (Shapiro et al. 2009). In particular, we propose (mVS-APM) with the following update rules for $k \geq 1$:

$$y_{k+1} := x_k - \frac{\gamma_k}{\eta}(x_k - \widehat{\mathrm{prox}}_{\eta F}(x_k)), \tag{9a}$$

$$x_{k+1} := y_{k+1} + \beta_k(y_{k+1} - y_k), \tag{9b}$$

where $\widehat{\mathrm{prox}}_{\eta F}(x_k)$ is obtained by taking finite number of steps of (prox-SSG) with a sample size of one at each step and having the following update rule for $j = 0, \ldots, N_k - 1$:

$$z_{k,j+1} := \mathbf{P}_{\eta/j,g}\left(z_{k,j} - \frac{\eta}{j}u_j\right), \quad u_j \in \partial\tilde{f}(z_{k,j}, \omega_j). \tag{prox-SSG}$$

Next, we state our assumptions and present the main result of this section. The constant in the rate and complexity bounds is dependent on $\tilde{\kappa}$; unlike the condition number κ in smooth regimes, $\tilde{\kappa}$ is user-specified and can be relatively small. For instance, $\tilde{\kappa} = 2$ when $\eta = 1/\mu$. We employ a measurable selection from $\partial\tilde{f}(x,\omega)$ as a stochastic subgradient in (SSG) and impose the following assumption.

Assumption 4. *For any $x \in \mathbb{R}^n$, consider a measurable selection $R(x,\omega) \in \partial\tilde{f}(x,\omega)$. Unbiasedness: we have that $\mathbb{E}[R(x,\omega)] = R(x) \in \partial f(x)$. Subgradient boundedness: there exists $M > 0$ such that for any x, $\mathbb{E}[\|R(x,\omega)\|^2] \leq M^2$. Compact domain: the function g has a compact domain, that is, there exists $\Delta > 0$ such that $\|x\| \leq \Delta$ for any $x \in \mathrm{dom}(g)$.*

Theorem 2 (Rate and Oracle Complexity of (mVS-APM)). *Suppose Assumptions 1 and 4 hold. Consider the iterates generated by VS-APM applied on $F_\eta(x)$ defined as (6), where $\theta \triangleq \left(1 - \frac{1}{2\sqrt{\tilde{\kappa}}}\right)$, $\rho \triangleq \left(1 - \frac{1}{2a\sqrt{\tilde{\kappa}}}\right)$, $\tilde{\kappa} = \frac{\mu\eta+1}{\mu\eta}$, $a > 2$, and $\gamma_k = \eta/2$, $N_k = \lfloor \rho^{-k} \rfloor$ for all $k \geq 0$. Then, the following hold for $Q \triangleq \max\{\eta^2 M^2, 4\Delta^2\}$.*

i. *Rate: for all $K \geq 1$, we have that*

$$\mathbb{E}[\|y_K - x^*\|^2] \leq \hat{C}\rho^{K-1} \textit{ where } \hat{C} \triangleq 2D\eta\tilde{\kappa} + C^2 + 8\tilde{\kappa}^{5/2}Qa. \tag{10}$$

ii. *Outer iteration complexity: the iteration complexity of (mVS-APM) in gradient steps (of $\nabla_x f_\eta(x_k)$) to obtain an ϵ-accurate solution is $\mathcal{O}(\sqrt{\tilde{\kappa}}\log(\hat{C}/\epsilon))$.*

iii. *Oracle complexity: to compute y_K such that $\mathbb{E}[\|y_K - x^*\|^2] \leq \epsilon$, the complexity of SSG steps is bounded as follows: $\sum_{k=1}^K N_k \leq \frac{2a^2\sqrt{\tilde{\kappa}}\hat{C}}{(a-1)\epsilon} = \mathcal{O}(1/\epsilon)$.*

Proof.

i. Recall that F_η is $\frac{\mu}{\mu\eta+1}$-strongly convex with $\frac{1}{\eta}$-Lipschitz continuous gradients. At iteration k of Algorithm 1, (prox-SSG) with single sampling can be used to inexactly solve $\min_u\left\{\mathbb{E}[\tilde{f}(u,\omega)] + g(u) + \frac{1}{2\eta}\|u - x_k\|^2\right\}$. In particular, let $\{z_{k,j}\}_{j=1}^{N_k}$ be the sequence generated by (prox-SSG) starting from $z_{k,0} = x_k$ and let z_k^* denote the unique optimal solution of the subproblem. Therefore, at step 1 of Algorithm 1, $\bar{w}_{k,N_k} = \frac{1}{\eta}(z_k^* - z_{k,N_k})$, and by the convergence rate of (prox-SSG) (Shapiro et al. 2009), $\mathbb{E}[\|\bar{w}_{k,N_k}\|^2] \leq \frac{\bar{Q}_k}{\eta^2 N_k}$, where $\bar{Q}_k \triangleq \max\{\eta^2 M^2, \|z_{k,0} - z_k^*\|^2\} \leq Q$ because $\|z_{k,0} - z_k^*\|^2 \leq 4\Delta^2$. The results in Lemma 1 hold when $F(x)$ is replaced by $F_\eta(x)$, by letting $L = \frac{1}{\eta}$, replacing μ by $\frac{\mu}{\mu\eta+1}$, ν^2 by $\frac{Q}{\eta^2}$, and setting $\bar{\alpha} = 1/(2\sqrt{\tilde{\kappa}})$, where $\tilde{\kappa} = \frac{\mu\eta+1}{\eta\mu}$:

$$\mathbb{E}[F_\eta(y_K) - F_\eta^*] \le \left(D + \frac{\mu}{2(\mu\eta+1)}C^2\right)(1-\bar{\alpha})^{K-1} + \sum_{i=0}^{K-1}\frac{(1-\bar{\alpha})^i\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2 N_{K-i}} + \sum_{i=0}^{K-2}\frac{(1-\bar{\alpha})^{i+1}\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2 N_{K-i-1}}. \tag{11}$$

From Lemma 2, x^* is a minimizer of function F if and only if x^* is a minimizer of function F_η. Because F_η is $\frac{\mu}{\mu\eta+1}$-strongly convex, $\frac{\mu}{2(\mu\eta+1)}\|y_K - x^*\|^2 \le F_\eta(y_K) - F_\eta(x^*)$, implying (11) can be written as

$$\frac{\mu\mathbb{E}[\|y_K - x^*\|^2]}{2(\mu\eta+1)} \le \left(D + \frac{\mu}{2(\mu\eta+1)}C^2\right)(1-\bar{\alpha})^{K-1} + \sum_{i=0}^{K-1}\frac{(1-\bar{\alpha})^i\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2 N_{K-i}} + \sum_{i=0}^{K-2}\frac{(1-\bar{\alpha})^{i+1}\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2 N_{K-i-1}}. \tag{12}$$

From (11), by definition of θ and recalling the increasing nature of $\{N_k\}$, we may claim the following:

$$\begin{aligned}
\frac{\mu\mathbb{E}[\|y_K - x^*\|^2]}{2(\mu\eta+1)} &\le \left(D + \frac{\mu}{2(\mu\eta+1)}C^2\right)\theta^{K-1} + \sum_{j=0}^{K-1}\theta^j\frac{\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2 N_{K-j-1}} + \sum_{j=0}^{K-1}\theta^{j+1}\frac{\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2 N_{K-j-1}} \\
&= \left(D + \frac{\mu}{2(\mu\eta+1)}C^2\right)\theta^{K-1} + \sum_{j=0}^{K-1}\frac{\theta^j(1+\theta)\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2 N_{K-j-1}} \\
&\overset{(1+\theta)\le 2}{\le} \left(D + \frac{\mu}{2(\mu\eta+1)}C^2\right)\theta^{K-1} + \sum_{j=0}^{K-1}\frac{2\theta^j\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2 N_{K-j-1}}.
\end{aligned} \tag{13}$$

If $N_{K-j-1} = \lfloor \rho^{-(K-j-1)} \rfloor$, by using Lemma A.1, we have the following:

$$\sum_{i=0}^{K-1}\frac{2\theta^j(2\eta+1/\mu)Q}{\eta^2\lfloor\rho^{-(K-j-1)}\rfloor} \le \sum_{i=0}^{K-1}\frac{\theta^j\left(2\eta+\frac{1}{\mu}\right)Q}{\eta^2\rho^{-(K-j-1)}} \le \frac{\left(2\eta+\frac{1}{\mu}\right)Q\rho^{K-1}}{\eta^2}\sum_{i=0}^{K-1}\left(\frac{\theta}{\rho}\right)^i \le \left(\frac{\left(2\eta+\frac{1}{\mu}\right)Q\rho}{\eta^2(\rho-\theta)}\right)\rho^{K-1}. \tag{14}$$

By substituting (14) in (13) and using $\frac{\rho}{\rho-\theta} = \frac{1-\frac{1}{2a\sqrt{\tilde{\kappa}}}}{\frac{1}{2\sqrt{\tilde{\kappa}}}-\frac{1}{2a\sqrt{\tilde{\kappa}}}} = \frac{(2a\sqrt{\tilde{\kappa}}-1)}{a-1} \le 2a\sqrt{\tilde{\kappa}}$, (13) becomes

$$\begin{aligned}
\mathbb{E}[\|y_K - x^*\|^2] &\le \frac{2(\mu\eta+1)}{\mu}\left(D + \frac{\mu}{2(\mu\eta+1)}C^2\right)\theta^{K-1} + \left(\frac{2(\mu\eta+1)}{\mu}\right)\frac{2}{\eta^2}\left(2\eta+\frac{1}{\mu}\right)Qa\sqrt{\tilde{\kappa}}\rho^{K-1} \\
&\le \left(\left(D\frac{2(\eta\mu+1)}{\mu}\right) + C^2 + \left(8\left(\frac{1+\eta\mu}{\eta\mu}\right)^2 Qa\right)\sqrt{\tilde{\kappa}}\right)\rho^{K-1} \\
&= \hat{C}\rho^{K-1}, \quad \text{where} \quad \hat{C} \triangleq 2D\eta\tilde{\kappa} + C^2 + 8\tilde{\kappa}^{5/2}Qa.
\end{aligned} \tag{15}$$

ii. We may derive the number of gradient steps K (of $\nabla_x f_\mu$) to obtain an ϵ-accurate solution:

$$\frac{1}{\rho} = \frac{1}{\left(1 - \frac{1}{2a\sqrt{\tilde{\kappa}}}\right)} = \frac{2a\sqrt{\tilde{\kappa}}}{(2a\sqrt{\tilde{\kappa}}-1)} \Rightarrow \frac{\log(\hat{C}) - \log(\epsilon)}{\log(1/\rho)} \le \frac{\log(\hat{C}) - \log(\epsilon)}{(1-\rho)} = (2a\sqrt{\tilde{\kappa}})\log(\hat{C}/\epsilon) \le K.$$

iii. To compute a vector y_K satisfying $\mathbb{E}[\|y_K - x^*\|^2] \le \epsilon$, we have $\hat{C}\rho^K \le \epsilon$, implying that $K = \lceil \log_{(1/\rho)}(\hat{C}/\epsilon) \rceil \le 1 + \log_{(1/\rho)}(\hat{C}/\epsilon)$. To obtain the oracle complexity, we require $\sum_{k=1}^K N_k$ gradients. If $N_k = \lfloor \rho^{-k} \rfloor \le \rho^{-k}$, we obtain the following because $(1-\rho) = (1\,/(2a\sqrt{\tilde{\kappa}}))$.

$$\sum_{k=1}^{K} \rho^{-k} \le \frac{\left(\frac{1}{\rho}\right)^{2+K}}{\left(\frac{1}{\rho}-1\right)} \le \frac{\left(\frac{1}{\rho}\right)^{3+\log_{1/\rho}(\hat{C}/\epsilon)}}{\left(\frac{1}{\rho}-1\right)} \le \frac{\hat{C}}{\rho^2(1-\rho)\epsilon} = \frac{2a\sqrt{\tilde{\kappa}}\hat{C}}{\rho^2\epsilon}. \tag{16}$$

Note that $\rho = 1 - \frac{1}{2a\sqrt{\tilde{\kappa}}}$, implying that

$$\rho^2 = 1 - 2/(2a\sqrt{\tilde{\kappa}}) + 1/(4a^2\tilde{\kappa}) = \frac{4a^2\tilde{\kappa} - 4a\sqrt{\tilde{\kappa}} + 1}{4a^2\tilde{\kappa}} \ge \frac{4a^2\tilde{\kappa} - 4a\tilde{\kappa}}{4a^2\tilde{\kappa}} = \frac{(a^2-a)}{a^2}$$

$$\Rightarrow \frac{\sqrt{\tilde{\kappa}}}{\rho^2} \le \frac{a^2\sqrt{\tilde{\kappa}}}{(a^2-a)} = \frac{a}{a-1}\sqrt{\tilde{\kappa}} \Rightarrow \text{ by (16), } \sum_{k=1}^{\log_{(1/\rho)}(\hat{C}/\epsilon)+1} \rho^{-k} \le \frac{2a^2\sqrt{\tilde{\kappa}}\hat{C}}{(a-1)\epsilon}. \quad \square$$

Remark 1. In Theorem 2, choosing $\eta = 1/\mu$ leads to $\mathbb{E}[\|y_K - x^*\|^2] \le \left(\frac{4D}{\mu} + C^2 + 12\sqrt{2}aQ\right)\rho^{K-1}$, and an oracle complexity of $\mathcal{O}\left(\frac{\max\{M^2/\mu^2, \|\tilde{x}_0 - \tilde{x}^*\|^2\}}{\epsilon}\right)$, matching the result by Shapiro et al. (2009).

Minimizing the convergence bound in (15) in η is possible via a less obvious coercivity and strict convexity claim for the nonsmooth function $\hat{C}(\eta)$ (see appendix for proof).

Lemma 3. *Consider $\hat{C}(\eta)$ defined as $\hat{C}(\eta) \triangleq 2D\eta\tilde{\kappa}(\eta) + C^2 + 8\tilde{\kappa}(\eta)^{5/2}Q(\eta)a$, where $Q \triangleq \max\{\eta^2M^2, 4\Delta^2\}$. Then, the following hold.*

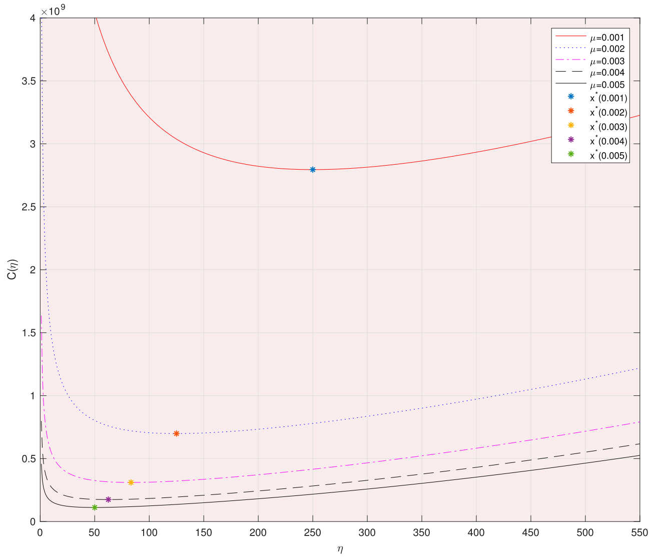
i. *$\hat{C}(\eta)$ is a coercive function on $\{\eta \mid \eta \ge 0\}$.*

ii. *$\hat{C}(\eta)$ is a strictly convex function on $\{\eta \mid \eta \ge 0\}$.*

iii. *The minimizer of $\hat{C}(\eta)$ on $\{\eta \mid \eta \ge 0\}$ is unique.*

Remark 2. Lemma 3 allows for claiming that $\hat{C}(\eta)$ has a unique minimizer η^*; in fact, such a minimizer can be computed by a standard semismooth Newton method (Facchinei and Pang 2003). Figure 1 provides a schematic of $\hat{C}(\eta)$ for different values of μ, whereas η^* is computed by semismooth Newton method. We note that, when μ is larger, $\eta^*(\mu)$ tends to be smaller. In such cases, obtaining an optimal η^* is particularly useful. However, when $\mu \ll 1$, we observe that $\eta^*(\mu) \gg 1$; consequently, this leads to rescaling of the step γ_k to $\frac{\gamma_k}{\eta}$, resulting in poorer behavior. Therefore, if $\mu \ll 1$, we employ $\eta = 1$, and this has far better empirical behavior as seen in the numerics.

Figure 1. Schematic of $\hat{C}(\eta)$ When $D = 10, M = 10, C = 100, a = 2.1, \Delta = 1$ for $\mu \in \{0.001, \cdots, 0.005\}$



2.4. Linear Convergence of mVS-PM: Non-compact Domains

In this section, we derive rate and complexity guarantees when (VS-PM), an unaccelerated variant of VS-APM, is applied on a Moreau-smoothed problem under possibly noncompact domains and under a (weaker) state-dependent bound on the subgradient (Assumption 5). When the subgradient of g is characterized by a state-dependent bound, the bound on the cumulative error in the accelerated method builds up because of a recursive relation, see (A.16). Hence, in this section, we consider a more general case in which Assumption 5 imposes a state-dependent bound, weakening Assumption 4. By employing an unaccelerated method, we derive a similar oracle complexity as in Section 2.3. To obtain rate results, we apply (VS-PM) with the following update rule:

$$x_{k+1} := x_k - \gamma(\nabla_x F_\eta(x_k) + \bar{w}_{k,N_k}), \tag{VS-PM}$$

where $\nabla_x F_\eta(x_k) + \bar{w}_{k,N_k}$ can be obtained by solving $\min_{u\in\mathbb{R}^n}[\mathbb{E}[\tilde{F}(u,\omega)] + \frac{1}{2\eta}\|u - x_k\|^2]$ inexactly taking N_k (stochastic) subgradient steps. Consider the sequence of iterates $\{x_k\}$ generated by applying an inexact gradient scheme on the following strongly convex smooth optimization problem.

$$\min_{x\in\mathbb{R}^n} F_\eta(x), \;\; \text{where } F_\eta(x) \triangleq \min_{u\in\mathbb{R}^n}\left[\mathbb{E}[\tilde{f}(u,\omega)] + g(u) + \frac{1}{2\eta}\|x-u\|^2\right].$$

In effect, given an $x_0 \in \mathbb{R}^n$, the inexact gradient scheme generates a sequence $\{x_k\}$ such that

$$x_{k+1} := x_k - \gamma(\nabla_x F_\eta(x_k) + \bar{w}_k). \tag{IG}$$

Given an x_k, we denote the update with the exact gradient by $\bar{x}_{k+1}$, which is defined as follows.

$$\bar{x}_{k+1} := x_k - \gamma\nabla_x F_\eta(x_k).$$

Recall that $\nabla_x F_\eta(x_k)$ is defined as $\nabla_x F_\eta(x_k) = \frac{1}{\eta}(x_k - z_k^*)$, where z_k^* is the unique minimizer of the following problem, that is,

$$z_k^* \triangleq \arg\min_{u\in\mathbb{R}^n}\left[\mathbb{E}[\tilde{F}(u,\omega)] + \frac{1}{2\eta}\|x_k - u\|^2\right]. \tag{17}$$

In other words, z_k^* is defined as

$$z_k^* \triangleq \text{prox}_{\eta F}(x_k) \text{ while } x^* = \text{prox}_{\eta F}(x^*).$$

Because $\text{prox}_{\eta F}(x_k)$ is unavailable in closed form, we may compute increasingly exact analogs; given $z_{k,0} = x_k$, we construct the sequence $\{z_{k,j}\}_{j=1}^{N_k}$ based on (SSG).

$$z_{k,j+1} = z_{k,j} - \sigma_j G(z_{k,j},\omega_{k,j}), \quad j \geq 0, \;\; \text{where} \quad G(z_{k,j},\omega_{k,j}) \in \partial\tilde{F}(z_{k,j},\omega_{k,j}) + \frac{1}{\eta}(z_{k,j} - x_k). \tag{SSG}$$

Consequently, at major iteration k, the inexact gradient of $F_\eta(x)$ is given by $\frac{1}{\eta}(x_k - z_{k,N_k})$, implying that $\bar{w}_k$ is defined as $\frac{1}{\eta}(z_k^* - z_{k,N_k})$. Consequently, we have that

$$x_{k+1} = x_k - \gamma\left(\frac{1}{\eta}(x_k - z_{k,N_k})\right) = \left(1 - \frac{\gamma}{\eta}\right)x_k + \frac{\gamma}{\eta}z_{k,N_k}.$$

We proceed to derive a bound on the conditional second moment of $G(z_{k,j},\omega_{k,j}) = S(z_{k,j},\omega_{k,j}) + \frac{1}{\eta}(z_{k,j} - x_k)$, where $S(z_{k,j},\omega_{k,j}) \in \partial\tilde{F}(z_{k,j},\omega_{k,j})$, $M_1^2 \triangleq 2\bar{M}^2 + \frac{4}{\eta^2}$, $M_2^2 \triangleq \frac{4}{\eta^2}$, and $M_3^2 \triangleq 2M^2$. This requires defining the history up to iteration j at outer iteration k by $\mathcal{F}_{k,j}$ as follows.

$$\mathcal{F}_0 = \{x_0\}, \mathcal{F}_{0,j} = \mathcal{F}_0 \cup \{S(z_{0,0},\omega_{0,0}), \cdots, S(z_{0,j-1},\omega_{k,j-1})\}, \qquad j = 1, \cdots, N_0, \tag{18}$$

$$\mathcal{F}_k = \mathcal{F}_{k-1,N_{k-1}} \cup \{x_k\}, \mathcal{F}_{k,j} = \mathcal{F}_k \cup \{S(z_{k,0},\omega_{k,0}), \cdots, S(z_{k,j-1},\omega_{k,j-1})\}, j = 1, \cdots, N_k, k \geq 1. \tag{19}$$

We now outline an assumption on the bound on the stochastic subgradient that scales with the size of x allowing for noncompact domains.

Assumption 5. *Let $\{x_k\}$ be a sequence generated by (VS-PM), where $\nabla_x F_\eta(x_k) + \bar{w}_{k,N_k}$ is computed by taking N_k steps of (SSG), leading to a set of iterates $\{z_{k,1}, \cdots, z_{k,N_k}\}$. Let $\mathcal{F}_{k,j}$ be defined as (19) for $k \geq 1$ and $j = 1, \cdots, N_k$. For any $z_{k,j}$, let*

$S(z_{k,j},\omega_{k,j})$ denote a measurable selection $S(z_{k,j},\omega_{k,j})\in\partial\tilde F(z_{k,j},\omega_{k,j})$. With these constructs, the following are assumed to hold.

a. *Unbiasedness: we have that $\mathbb{E}[S(z_{k,j},\omega_{k,j})\,|\,\mathcal{F}_{k,j}]=S(z_{k,j})\in\partial F(z_{k,j})$ almost surely.*

b. *Subgradient boundedness: there exists $M,\bar M>0$ such that, for any x, $\mathbb{E}[\|S(z_{k,j},\omega_{k,j})\|^2\,|\,\mathcal{F}_{k,j}]\le\bar M^2\|z_{k,j}\|^2+M^2$ almost surely.*

Consequently, we have that

$$\begin{aligned}\|G(z_{k,j},\omega_{k,j})\|^2&\le 2\|S(z_{k,j},\omega_{k,j})\|^2+\frac{2}{\eta^2}\|z_{k,j}-x_k\|^2\le 2\|S(z_{k,j},\omega_{k,j})\|^2+\frac{4}{\eta^2}\|z_{k,j}\|^2+\frac{4}{\eta^2}\|x_k\|^2\\ \Rightarrow \mathbb{E}[\|G(z_{k,j},\omega_{k,j})\|^2\,|\,\mathcal{F}_{k,j}]&\overset{\text{Assump. 5}}{\le}\left(2\bar M^2+\frac{4}{\eta^2}\right)\|z_{k,j}\|^2+2M^2+\frac{4}{\eta^2}\|x_k\|^2\\ &=:M_1^2\|z_{k,j}\|^2+M_2^2\|x_k\|^2+M_3^2.\end{aligned}\tag{20}$$

Based on Assumption 5 and inspired by a proof technique from Chambolle and Pock (2011) among others, we derive a rate statement for (SSG) (see appendix for proof).

Proposition 1. *Consider (17) in which $F(\cdot,\omega)$ is a μ-strongly convex function and $S(z,\omega)\in\partial\tilde F(z,\omega)$ for any z. Suppose Assumption 5 holds and $\hat a^2\triangleq 4+4M_1^2+2M_2^2$ and $\hat b^2\triangleq(4M_1^2+2M_2^2)[\|x^*\|^2]+M_3^2$. Given x_k, consider a sequence generated by (SSG) in which $\tilde\mu=\mu+\frac{1}{\eta}$, $\bar J\triangleq\lceil\frac{2M_1^2}{\tilde\mu^2}-1\rceil$, and*

$$\sigma_j\triangleq\begin{cases}\min\left\{\dfrac{1}{(j+1)\log(j+1)},\dfrac{\tilde\mu}{M_1^2}\right\}, & j<\bar J\\ \dfrac{1}{(j+1)\log(j+1)}. & j\ge\bar J\end{cases}$$

Then, the following holds for $j\ge\bar J$.

$$\mathbb{E}[\|z_{k,j}-z_k^*\|^2\,|\,\mathcal{F}_k]\le\frac{\hat a^2\|x_k-x^*\|^2+\hat b^2}{j}.\tag{21}$$

We now show the convergence of mVS-PM when $\nabla_x F_\eta(x)$ is approximated via (SSG) (see appendix for proof).

Theorem 3 (mVS-PM Under State-Dependent Bound on Subgradients). *Suppose Assumptions 1 and 5 hold. Consider the iterates generated by (VS-PM) applied on $F_\eta(x)$, where $\tilde\kappa\triangleq 1+\frac{1}{\eta\mu}$, $\gamma=\eta$ and $N_k\triangleq\lfloor N_0\rho^{-k}\rfloor$ for all $k\ge 0$, $N_0>\max\left\{\frac{2\hat a^2}{(1-q/2)},\bar J\right\}$, $q\triangleq 1-\frac{1}{\tilde\kappa}$, $p_0\triangleq\frac{q}{2}+\frac{2\hat a^2}{N_0}$, and $\bar J\triangleq\lceil\frac{2M_1^2}{\bar\mu^2}-1\rceil$. Then, the following hold.*

i. *Rate: for all $k\ge 1$, we have that the following holds.*

$$\mathbb{E}[\|x_k-x^*\|^2]\le\mathcal{C}\hat p^k\ \text{where}\ \mathcal{C}\triangleq\left(\mathbb{E}[\|x_0-x^*\|^2]+\frac{\hat b\widehat D}{N_0}\right),\begin{cases}\rho\ne p_0, & \hat p=\max\{\rho,p_0\},\widehat D\triangleq\dfrac{1}{1-\dfrac{\min\{\rho,p_0\}}{\max\{\rho,p_0\}}}\\ \rho=p_0. & \hat p\in(p_0,1),\widehat D>\dfrac{1}{\ln(p_0/\hat p)^e}\end{cases}$$

ii. *Iteration complexity: the iteration complexity of mVS-PM in gradient steps of $(\nabla_x F_\eta(x_k))$ to obtain an ϵ-accurate solution is $\mathcal{O}(\tilde\kappa\log(\mathcal{C}/\epsilon))$.*

iii. *Oracle complexity in (SSG) steps: to compute x_K such that $\mathbb{E}[\|x_K-x^*\|^2]\le\epsilon$, the complexity in subgradient steps is bounded as $\sum_{k=1}^K N_k\le\mathcal{O}\left(\tilde\kappa\left(\frac{\mathcal{C}}{\epsilon}\right)^{\log_{1/\hat p}(1/\rho)}\right)$ for $\hat p\in[p_0,1)$, $\rho\le p_0$ and $\sum_{k=1}^K N_k\le\mathcal{O}(\tilde\kappa(\frac{\mathcal{C}}{\epsilon}))$ for $\rho>p_0$.*

Remark 3. We observe that, when $\rho>p_0$, we achieve the optimal oracle complexity in subgradient steps akin to the statement in the regime of bounded subgradients. Notably, $\tilde\kappa$ can be controlled because η is any nonnegative scalar. For instance, if $\eta=\frac{1}{\mu}$, $\tilde\kappa=2$.

3. Iteratively Smoothed VS-APM for Nonsmooth Convex Problems

Thus far, we consider settings in which f is a strongly convex function. However, there are many instances when the function f is neither smooth nor strongly convex. In fact, in strongly convex regimes, estimating the strong convexity parameter may often be challenging. In such settings, if the function f is subdifferentiable, then subgradient methods provide an avenue for resolving such problems in stochastic regimes but display a significantly poorer rate of convergence. Nesterov (2005b) shows that, for a subclass of problems, an accelerated gradient scheme may be applied to a suitably *smoothed* problem in which the smoothing leads to a differentiable problem with Lipschitz continuous gradients (with known Lipschitz constants). If the smoothing parameter is chosen suitably, the convergence rate to an approximate solution can be improved to $\mathcal{O}(1/K)$ from $\mathcal{O}(1/\sqrt{K})$ in terms of expected suboptimality. However, because the smoothing parameter is maintained as fixed, Nesterov's approach can provide approximate solutions at best but not asymptotically exact solutions. Subsequently, Nesterov (2005a) considers a primal–dual smoothing technique in which the smoothing parameter is reduced at every step, whereas extensions and generalizations are considered more recently by Tran-Dinh et al. (2018) and Van Nguyen et al. (2017). In this section, we develop an iteratively smoothed, variable sample-size, accelerated proximal gradient scheme that can contend with expectation-valued objectives and is asymptotically convergent. This can be viewed as a variant of the primal smoothing scheme introduced by Nesterov (2005b), in which the smoothing parameter is reduced after every step; this scheme is shown to admit a rate of $\mathcal{O}(1/K)$, matching the finding by Nesterov (2005b); however, our scheme is blessed with asymptotic guarantees rather than providing approximate solutions. In Section 3.1, we derive rate and complexity statements, in Section 3.2 for the iteratively smoothed VS-APM (or sVS-APM), recovering the optimal rate of $\mathcal{O}(1/K^2)$ with the optimal oracle complexity of $\mathcal{O}(1/\epsilon^2)$ under smoothness. Finally, in Section 3.3, under suitable choices of smoothing sequences, sVS-APM produces sequences that converge a.s. to an optimal solution.

3.1. Smoothing Techniques

In this section, we consider minimizing F where F is defined as $F(x) \triangleq \mathbb{E}[\tilde{F}(x,\omega)]$, where $\tilde{f}(x,\omega) = \tilde{f}(x,\omega) + g(x)$ such that f and g are convex and may be nonsmooth, whereas g has an efficient prox evaluation (or "proximable") but f is not proximable. Note that this setting is more general than structured nonsmooth problems, in which the function f is considered to be convex and smooth. In contrast to the previous section, we assume that $\nabla_x \tilde{f}_{\eta_k}(x_k,\omega_k)$ is generated from the stochastic oracle, in which η_k is a smoothing parameter at iteration k such that its sequence is diminishing. Beck and Teboulle (2012) define an (α,β)-smoothable function as follows.

Definition 1 ((α,β)-Smoothable; Beck 2017). A convex function $h : \mathbb{R}^n \to \mathbb{R}$ is referred to as (α,β)-smoothable if, for any $\eta > 0$, there exists a convex differentiable function $h_\eta : \mathbb{R}^n \to \mathbb{R}$ that satisfies the following: (i) $h_\eta(x) \le h(x) \le h_\eta(x) + \eta\beta$ for all x, and (ii) h_η is α/η smooth.

There are a host of smoothing functions based on the nature of h. For instance, when $h(x) = \|x\|_2$, then $h_\eta(x) = \sqrt{\|x\|_2^2 + \eta^2} - \eta$, implying that h is a (1, 1)-smoothable function. If $h(x) = \max\{x_1, x_2, \ldots, x_n\}$, then h is $(1, \log(n))$-smoothable and $h_\eta(x) = \eta\log(\sum_{i=1}^n e^{x_i/\eta}) - \eta\log(n)$. (see Beck and Teboulle 2012 for more examples). Recall that, when h is a proper, closed, and convex function, the Moreau envelope is defined as $h_\eta(x) \triangleq \min_u \left\{h(u) + \frac{1}{2\eta}\|u-x\|^2\right\}$. In fact, h is $(1, B^2)$-smoothable when h_η is given by the Moreau envelope (see Beck and Teboulle 2012) and B denotes a uniform bound on $\|s\|$ in x, where $s \in \partial h(x)$. There are a range of other smoothing techniques, including Nesterov smoothing (see Nesterov 2005b) and inf-conv smoothing (see Beck 2017); our approach is agnostic to the choice of smoothing. In particular, if $\tilde{f}(\cdot,\omega)$ is a proper, closed, and convex function in x for every ω, then $\tilde{f}(\cdot,\omega)$ is $(1, B^2)$-smoothable for every ω for which $\tilde{f}_\eta(\cdot,\omega)$ is a suitable smoothing. In fact, if $\tilde{f}(\cdot,\omega)$ satisfies the following smoothability assumption, then smoothability of f follows as shown by Lemma 4. It is worth emphasizing that the smoothing of f, denoted by f_η, is defined as

$$f_\eta(x) \triangleq \mathbb{E}[\tilde{f}_\eta(x,\omega)], \tag{22}$$

where $\tilde{f}_\eta(\cdot,\omega)$ is a smoothing of $\tilde{f}(\cdot,\omega)$.

Assumption 6. *The function $\tilde{f}(\cdot,\omega)$ is an $(\alpha(\omega),\beta(\omega))$-smoothable function for every $\omega \in \Omega$, where $\mathbb{E}[\alpha(\omega)] \le \tilde{\alpha}$ and $\mathbb{E}[\beta(\omega)] \le \tilde{\beta}$ with $\tilde{\alpha}, \tilde{\beta} > 0$; that is, for any $\eta > 0$, there exists a convex differentiable function $\tilde{f}_\eta(\cdot,\omega)$ for every $\omega \in \Omega$*

such that

$$\tilde{f}_\eta(x,\omega) \le \tilde{f}(x,\omega) \le \tilde{f}_\eta(x,\omega) + \eta\beta(\omega), \quad \textit{for all } x$$

$$\textit{and } \|\nabla_x \tilde{f}_\eta(x,\omega) \text{-} \nabla_x \tilde{f}_\eta(y,\omega)\| \le \frac{\alpha(\omega)}{\eta}\|x-y\|, \qquad \textit{for all } x,y,$$

where $\mathbb{E}[\alpha(\omega)] \le \tilde{\alpha}$ *and* $\mathbb{E}[\beta(\omega)] \le \tilde{\beta}$.

Based on the following lemma, we observe that f is $(\tilde{\alpha},\tilde{\beta})$-smoothable if $\tilde{f}(\cdot,\omega)$ satisfies suitable smoothability requirements for almost every $\omega \in \Omega$.

Lemma 4. *Suppose Assumption 6 holds. Then, there exist* $\tilde{\alpha},\tilde{\beta} > 0$ *such that* f *is* $(\tilde{\alpha},\tilde{\beta})$*-smoothable, where* $f(x) \triangleq \mathbb{E}[\tilde{f}(x,\omega)]$.

We proceed to develop a smoothed variant of VS-APM, referred to as sVS-APM, in which $\nabla_x \tilde{f}_{\eta_k}(x_k,\omega_k)$ is generated from the stochastic oracle and η_k is driven to zero at a sufficient rate (See Algorithm 2).

Algorithm 2 (Iteratively Smoothed VS-APM (sVS-APM))

```
(0) Given budget $M$, $x_0 \in X$, $y_0 = x_0$ and positive sequences $\{\gamma_k, N_k\}$. Set $\lambda_0 = 0$, $\lambda_1 = 1$; $k := 1$.
(1) $y_{k+1} = \mathbf{P}_{\gamma_k,g}(x_k - \gamma_k(\nabla_x f_{\eta_k}(x_k) + \bar{w}_{k,N_k}))$;
(2) $\lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}$;
(3) $x_{k+1} = y_{k+1} + \frac{(\lambda_k - 1)}{\lambda_{k+1}}(y_{k+1} - y_k)$;
(4) If $\sum_{j=1}^k N_j > M$, then stop; else $k := k+1$; return to (1).
```

3.2. Rate and Complexity Analysis

In this section, we develop rate and oracle complexity statements for Algorithm 2 when f is $(1,B^2)$ smoothable and then specialize these results to both the deterministic nonsmooth and stochastic smooth regimes. We begin with a modified assumption.

Assumption 7. *(i) The function* g *is lower semicontinuous and convex with effective domain denoted by* $\operatorname{dom}(g)$*; (ii)* f *is proper, closed, convex, and* $(1,B^2)$*-smoothable on an open set containing* $\operatorname{dom}(g)$*; (iii) there exists* $C > 0$ *such that* $\mathbb{E}[\|x_0 - x^*\|] \le C$ *for all* $x^* \in X^*$.

Note that Assumption 6 represents a set of sufficiency conditions for f to be smoothable; here, we directly assume that f is smoothable to ease exposition.

Lemma 5. *Suppose Assumption 7 holds. Consider the iterates generated by sVS-APM on* $F(x)$*. Suppose Assumption 3 holds for* $f_{\eta_k}(x)$*. If* $\{\gamma_k\}$ *is a decreasing sequence and* $\gamma_k \le \eta_k/2$*, then the following holds for all* $K \ge 2$:

$$\mathbb{E}[F_{\eta_k}(y_K) - F_{\eta_k}(x^*)] \le \frac{2}{\gamma_{K-1}(K-1)^2}\sum_{k=1}^{K-1}\gamma_k^2 k^2 \frac{\nu^2}{N_k} + \frac{2C^2}{\gamma_{K-1}(K-1)^2}.$$

Proof. By the update rule in Algorithm 2, we have

$$y_{k+1} = \arg\min_x \; g(x) + \frac{1}{2\gamma_k}\|x - x_k\|^2 + (\nabla_x f_{\eta_k}(x_k) + \bar{w}_k)^T x. \tag{23}$$

From the optimality condition for (23), $0 \in \partial g(y_{k+1}) + \frac{1}{\gamma_k}(y_{k+1} - x_k) + \nabla_x f_{\eta_k}(x) + \bar{w}_k$. By convexity of $g(x)$, we have that $g(x) \ge g(y_k) + s^T(x - y_{k+1})$ for all $s \in \partial g(y_k)$. Hence, we obtain the following.

$$g(x) + (\nabla_x f_{\eta_k}(x_k) + \bar{w}_k)^T x \ge g(y_{k+1}) + (\nabla_x f_{\eta_k}(x_k) + \bar{w}_k)^T y_{k+1} - \frac{1}{\gamma_k}(x - y_{k+1})^T(y_{k+1} - x_k).$$

Now, by using Lemma A.2, we obtain that

$$\begin{aligned} & g(x) + (\nabla_x f_{\eta_k}(x_k) + \bar{w}_k)^T x + \frac{1}{2\gamma_k}\|x - x_k\|^2 \\ & \ge g(y_{k+1}) + (\nabla_x f_{\eta_k}(x_k) + \bar{w}_k)^T y_{k+1} + \frac{1}{2\gamma_k}\|x_k - y_{k+1}\|^2 + \frac{1}{2\gamma_k}\|x - y_{k+1}\|^2. \end{aligned} \tag{24}$$

By invoking the convexity of f_{η_k} and by using the Lipschitz continuity of $\nabla_x f_{\eta_k}$, we obtain

$$\begin{aligned} f_{\eta_k}(x) &\geq f_{\eta_k}(x_k) + \nabla_x f_{\eta_k}(x_k)^T(x - x_k) \\ &\geq f_{\eta_k}(y_{k+1}) + \nabla_x f_{\eta_k}(x_k)^T(x - y_{k+1}) - \frac{1}{2\eta_k}\|x_k - y_{k+1}\|^2 \\ &= f_{\eta_k}(y_{k+1}) + (\nabla_x f_{\eta_k}(x_k) + \bar{w}_k)^T(x - y_{k+1}) - \frac{1}{2\eta_k}\|x_k - y_{k+1}\|^2 - \bar{w}_k^T(x - y_{k+1}), \end{aligned} \tag{25}$$

where the last equality follows from adding and subtracting $\bar{w}_k$. By adding (24) and (25), we obtain

$$\begin{aligned} F_{\eta_k}(y_{k+1}) - F_{\eta_k}(x) &\leq \frac{1}{2\gamma_k}\|x - x_k\|^2 - \frac{1}{2\gamma_k}\|x - y_{k+1}\|^2 + \frac{1}{2}\left(\frac{1}{\eta_k} - \frac{1}{\gamma_k}\right)\|x_k - y_{k+1}\|^2 - \bar{w}_k^T(y_{k+1} - x) \\ &= \left(\frac{1}{2\eta_k} - \frac{1}{\gamma_k}\right)\|x_k - y_{k+1}\|^2 + \frac{1}{\gamma_k}(x_k - y_{k+1})^T(x_k - x) - \bar{w}_k^T(y_{k+1} - x), \end{aligned} \tag{26}$$

where the last inequality follows from Lemma A.2 by choosing $Q = I$, $v_1 = x_k$, $v_2 = x$, and $v_3 = y_k$. By setting $x = y_k$ in (26), we have

$$\begin{aligned} F_{\eta_k}(y_{k+1}) - F_{\eta_k}(y_k) \leq \left(\frac{1}{2\eta_k} - \frac{1}{\gamma_k}\right)\|x_k - y_{k+1}\|^2 + \frac{1}{\gamma_k}(x_k - y_{k+1})^T(x_k - y_k) \\ - \bar{w}_{k,N_k}^T(y_{k+1} - y_k). \end{aligned} \tag{27}$$

Similarly, by letting $x = x^*$, we can obtain

$$\begin{aligned} F_{\eta_k}(y_{k+1}) - F_{\eta_k}(x^*) \quad \leq \left(\frac{1}{2\eta_k} - \frac{1}{\gamma_k}\right)\|x_k - y_{k+1}\|^2 + \frac{1}{\gamma_k}(x_k - y_{k+1})^T(x_k - x^*) \\ - \bar{w}_{k,N_k}^T(y_{k+1} - x^*). \end{aligned} \tag{28}$$

By invoking Lemma A.2 in which $v_1 = x_k$, $v_2 = y_{k+1}$ and $v_3 = y_k$, we obtain

$$\frac{1}{\gamma_k}(y_{k+1} - x_k)^T(y_k - x_k) = \frac{1}{2\gamma_k}(\|y_k - x_k\|^2 + \|y_{k+1} - x_k\|^2 - \|y_{k+1} - y_k\|^2).$$

Consequently, (27) can further bounded as follows:

$$\begin{aligned} F_{\eta_k}(y_{k+1}) - F_{\eta_k}(y_k) &\leq \left(\frac{1}{2\eta_k} - \frac{1}{\gamma_k}\right)\|x_k - y_{k+1}\|^2 + \frac{1}{\gamma_k}(x_k - y_{k+1})^T(x_k - y_k) - \bar{w}_{k,N_k}^T(y_{k+1} - y_k) \\ &= \left(\frac{1}{2\eta_k} - \frac{1}{\gamma_k}\right)\|x_k - y_{k+1}\|^2 + \frac{1}{2\gamma_k}(\|x_k - y_k\|^2 + \|y_{k+1} - x_k\|^2 - \|y_{k+1} - y_k\|^2) - \bar{w}_{k,N_k}^T(y_{k+1} - y_k) \\ &= \left(\frac{1}{2\eta_k} - \frac{1}{2\gamma_k}\right)\|x_k - y_{k+1}\|^2 + \frac{1}{2\gamma_k}(\|x_k - y_k\|^2 - \|y_{k+1} - y_k\|^2) - \bar{w}_{k,N_k}^T(y_{k+1} - y_k). \end{aligned} \tag{29}$$

Similarly, we have that

$$\begin{aligned} F_{\eta_k}(y_{k+1}) - F_{\eta_k}(x^*) \leq \left(\frac{1}{2\eta_k} - \frac{1}{2\gamma_k}\right)\|x_k - y_{k+1}\|^2 + \frac{1}{2\gamma_k}(\|x_k - x^*\|^2 - \|y_{k+1} - x^*\|^2) \\ - \bar{w}_{k,N_k}^T(y_{k+1} - x^*). \end{aligned} \tag{30}$$

By multiplying (29) by $(\lambda_k - 1)$ and adding to (30), where $\delta_k \triangleq F_{\eta_k}(y_k) - F_{\eta_k}(x^*)$, we have

$$\lambda_k\delta_{k+1} - (\lambda_k - 1)\delta_k \leq \left(\frac{1}{2\eta_k} - \frac{1}{2\gamma_k}\right)\lambda_k\|y_{k+1} - x_k\|^2 \tag{31}$$

$$+ \frac{1}{2\gamma_k}(\lambda_k - 1)(\|x_k - y_k\|^2 - \|y_{k+1} - y_k\|^2) + \frac{1}{2\gamma_k}(\|x_k - x^*\|^2 - \|y_{k+1} - x^*\|^2) \tag{32}$$

$$+ \bar{w}_{k,N_k}^T((\lambda_k - 1)y_k + x^* - \lambda_k y_{k+1}). \tag{33}$$

Again, by using Lemma A.2, we may express the terms in (32) as follows:

$$\begin{aligned}
&\frac{1}{2\gamma_k}(\lambda_k-1)(\|x_k-y_k\|^2-\|y_{k+1}-y_k\|^2)+\frac{1}{2\gamma_k}(\|x_k-x^*\|^2-\|y_{k+1}-x^*\|^2)\\
&=\frac{1}{2\gamma_k}(\lambda_k\|x_k-y_k\|^2-\lambda_k\|y_{k+1}-y_k\|^2-\|x_k-y_k\|^2+\|y_{k+1}-y_k\|^2+\|x_k-x^*\|^2-\|y_{k+1}-x^*\|^2)\\
&=\frac{1}{2\gamma_k}\Big(-\lambda_k\|y_{k+1}-x_k\|^2+2\lambda_k(y_{k+1}-x_k)^T(y_k-x_k)+\|y_{k+1}-x_k\|^2-2(y_{k+1}-x_k)^T(y_k-x_k)\\
&-\|y_{k+1}-x_k\|^2+2(y_{k+1}-x_k)^T(x^*-x_k)\Big)\\
&=\frac{1}{2\gamma_k}\Big(-\lambda_k\|y_{k+1}-x_k\|^2+2(y_{k+1}-x_k)^T((\lambda_k-1)y_k-\lambda_kx_k+x^*)\Big).
\end{aligned}$$

In addition,

$$\bar{w}_{k,N_k}^T((\lambda_k-1)y_k+x^*-\lambda_ky_{k+1})=\bar{w}_{k,N_k}^T((\lambda_k-1)y_k+x^*-\lambda_kx_k)+\bar{w}_{k,N_k}^T(\lambda_kx_k-\lambda_ky_{k+1}).$$

From the update rule, $\lambda_{k-1}^2=\lambda_k(\lambda_k-1)=\lambda_k^2-\lambda_k$. Now, by multiplying (31) by λ_k, we obtain the following, in which $u_k=(\lambda_k-1)y_k-\lambda_kx_k+x^*$:

$$\begin{aligned}
&\lambda_k^2\delta_{k+1}-\lambda_{k-1}^2\delta_k\le\lambda_k^2\left(\frac{1}{2\eta_k}-\frac{1}{2\gamma_k}\right)\|y_{k+1}-x_k\|^2\\
&+\frac{1}{2\gamma_k}\Big(-\|\lambda_ky_{k+1}-\lambda_kx_k\|^2+2(\lambda_ky_{k+1}-\lambda_kx_k)^T((\lambda_k-1)y_k+x^*-\lambda_kx_k)\Big)\\
&-\lambda_k^2\bar{w}_{k,N_k}^T(x_k-y_{k+1})-\lambda_kw_k^Tu_k=\lambda_k^2\left(\frac{1}{2\eta_k}-\frac{1}{2\gamma_k}\right)\|y_{k+1}-x_k\|^2-\lambda_k^2\bar{w}_{k,N_k}^T(x_k-y_{k+1})\\
&+\frac{1}{2\gamma_k}(\|\lambda_kx_k-(\lambda_k-1)y_k-x^*\|^2-\|\lambda_ky_{k+1}-(\lambda_k-1)y_k-x^*\|^2)-\lambda_kw_k^Tu_k\\
&\le\frac{\lambda_k^2}{\dfrac{2}{\gamma_k}-\dfrac{2}{\eta_k}}\|\bar{w}_{k,N_k}\|^2+\frac{1}{2\gamma_k}(\|u_k\|^2-\|u_{k+1}\|^2)-\lambda_kw_k^Tu_k,
\end{aligned}\tag{34}$$

where, in the last inequality, we use the update rule of algorithm, $x_{k+1}=y_{k+1}+\frac{\lambda_k-1}{\lambda_{k+1}}(y_{k+1}-y_k)$, to obtain the following:

$$u_{k+1}=(\lambda_{k+1}-1)y_{k+1}-\lambda_{k+1}x_{k+1}+x^*=(\lambda_k-1)y_k-\lambda_ky_{k+1}+x^*.$$

By multiplying both sides by γ_k and assuming $\gamma_k\le\gamma_{k-1}$, we obtain

$$\gamma_k\lambda_k^2\delta_{k+1}-\gamma_{k-1}\lambda_{k-1}^2\delta_k\le\frac{\gamma_k\lambda_k^2}{\dfrac{2}{\gamma_k}-\dfrac{2}{\eta_k}}\|\bar{w}_{k,N_k}\|^2+\frac{1}{2}(\|u_k\|^2-\|u_{k+1}\|^2)-\gamma_k\lambda_kw_k^Tu_k.\tag{35}$$

By assuming $\gamma_k\le\frac{\eta_k}{2}$, we obtain $\frac{1}{\gamma_k}-\frac{1}{\eta_k}\ge\frac{1}{2\gamma_k}$, implying that

$$\gamma_k\lambda_k^2\delta_{k+1}-\gamma_{k-1}\lambda_{k-1}^2\delta_k\le\gamma_k^2\lambda_k^2\|\bar{w}_{k,N_k}\|^2+\frac{1}{2}(\|u_k\|^2-\|u_{k+1}\|^2)-\gamma_k\lambda_kw_k^Tu_k.\tag{36}$$

Summing (36) from $k=1$ to $K-1$, we have the following:

$$\begin{aligned}
&\gamma_{K-1}\lambda_{K-1}^2\delta_K\le\sum_{k=1}^{K-1}\gamma_k^2\lambda_k^2\|\bar{w}_{k,N_k}\|^2+\frac{1}{2}\|u_1\|^2-\sum_{k=1}^{K-1}\gamma_k\lambda_kw_k^Tu_k\\
&\Rightarrow\delta_K\le\frac{1}{\gamma_{K-1}\lambda_{K-1}^2}\sum_{k=1}^{K-1}\gamma_k^2\lambda_k^2\|\bar{w}_{k,N_k}\|^2+\frac{1}{2\gamma_{K-1}\lambda_{K-1}^2}\|u_1\|^2-\frac{1}{\gamma_{K-1}\lambda_{K-1}^2}\sum_{k=1}^{K-1}\gamma_k\lambda_kw_k^Tu_k.
\end{aligned}$$

Taking expectations, we note that the last term on the right is zero (under a zero-bias assumption), leading to the following:

$$\mathbb{E}[\delta_K] \le \frac{1}{\gamma_{K-1}\lambda_{K-1}^2}\sum_{k=1}^{K-1}\gamma_k^2\lambda_k^2\frac{\nu^2}{N_k} + \frac{1}{2\gamma_{K-1}\lambda_{K-1}^2}\mathbb{E}[\|u_1\|^2\|] \le \frac{2}{\gamma_{K-1}(K-1)^2}\sum_{k=1}^{K-1}\gamma_k^2k^2\frac{\nu^2}{N_k} + \frac{2C^2}{\gamma_{K-1}(K-1)^2},$$

where, in the last inequality, we use the fact that $\|y - x^*\| \le C$ for all $y \in \mathrm{dom}(g)$ and $\frac{k}{2} \le \lambda_k \le k$, which may be shown inductively. □

We are now ready to prove our main rate result and oracle complexity bound for sVS-APM.

Theorem 4 (Rate Statement and Oracle Complexity Bound for sVS-APM). *Suppose Assumption 7 holds. Consider the iterates generated by sVS-APM on $F(x)$. Suppose Assumption 3 holds for f_{η_k}. Suppose $\{\lambda_k\}$ is specified in sVS-APM, $\eta_k = 1/k$, $\gamma_k = 1/2k$, and $N_k = \lfloor k^a \rfloor$.*

i. *The following holds for any $K \ge 1$:*

$$\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \begin{cases} \dfrac{\left(\dfrac{2\nu^2 a}{a-1} + 4C^2 + B^2\right)}{K}, & a = 1+\delta, \delta \in [\delta_L, \delta_U] \\[2ex] \dfrac{2\nu^2(1+\log(K)) + 4C^2 + B^2}{K}, & a = 1 \end{cases}$$

ii. *Let $\epsilon \le \tilde{C}/2$, and K is such that $\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \epsilon$. Then, the following holds.*

$$\sum_{k=1}^{K} N_k \le \begin{cases} \mathcal{O}\left(\dfrac{1}{\epsilon^{2+\delta_L}}\right), & a = 1+\delta, \delta \in [\delta_L, \delta_U] \\[2ex] \mathcal{O}\left(\dfrac{1}{\epsilon^2}\log^2(1/\epsilon)\right). & a = 1 \end{cases}$$

Proof.

(i) If $N_k = \lfloor k^a \rfloor \ge \frac{1}{2}k^a$ and $\gamma_k = 1/(2k)$ is utilized in Lemma 5, we obtain the following:

$$\mathbb{E}[\delta_{K+1}] \le \frac{2\nu^2}{K}\sum_{k=1}^{K}\frac{1}{k^a} + \frac{4C^2}{K}. \tag{37}$$

a. $a = 1+\delta$, where $\delta \in [\delta_L, \delta_U]$. Consequently, we may derive the next bound.

$$\sum_{k=1}^{K} k^{-a} = 1 + \sum_{k=2}^{K} k^{-a} \le 1 + \int_1^K k^{-a}dk = 1 + \frac{1-K^{1-a}}{a-1} \le \frac{1+\delta_U}{\delta_L}.$$

By invoking $(1, B^2)$-smoothability of f and $\eta_K = 1/K$, we have that $F_{\eta_K}(y_{K+1}) \le F(y_{K+1})$ and $-F_{\eta_K}(x^*) \le -F(x^*) + \eta B^2$. Hence, the required bound follows from (37)

$$\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \frac{2\nu^2 a}{(a-1)K} + \frac{4C^2 + B^2}{K} \le \frac{\bar{C}}{K}, \quad \text{where } \bar{C} \triangleq \frac{2\nu^2 a}{(a-1)} + 4C^2 + B^2.$$

b. $a = 1$. Recall that the convergence rate is given by the following:

$$\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \frac{\frac{2\nu^2(a-K^{1-a})}{(a-1)} + 4C^2 + B^2}{K}.$$

Taking limits, we obtain that

$$\lim_{a\to 1}\frac{a - K^{1-a}}{a-1} = \lim_{a\to 1}\frac{1 + K^{1-a}\log(K)}{1} = 1 + \log(K).$$

Therefore, we have that

$$\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \frac{2\nu^2 \log(K) + 4C^2 + B^2}{K} \triangleq \frac{a + b\log(K)}{K}.$$

(ii) Consider y_{K+1} satisfying $\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \epsilon$. We again consider two cases. a. $a = 1+\delta$, where $\delta \in [\delta_L, \delta_U]$. We have $\frac{\bar{C}}{K} \le \epsilon$, which implies that $K = \lceil \bar{C}/\epsilon \rceil$. To obtain the optimal oracle complexity, we require $\sum_{k=1}^{K} N_k$ gradients. Hence, the following holds for sufficiently small ϵ such that $2 \le \bar{C}/\epsilon$:

$$\sum_{k=1}^{K} N_k \le \sum_{k=1}^{K} k^a = \sum_{k=1}^{1+\bar{C}/\epsilon} k^a \le \int_0^{2+\bar{C}/\epsilon} k^a da = \frac{(2+\bar{C}/\epsilon)^{1+a}}{1+a} \le \left(\frac{\bar{C}}{\epsilon}\right)^{1+a} \le \mathcal{O}\left(\frac{1}{\epsilon^{1+a}}\right) \le \mathcal{O}\left(\frac{1}{\epsilon^{2+\delta_L}}\right).$$

b. $a = 1$. To compute K such that $\frac{a+b\log(K)}{K} \le \epsilon$ is not immediately obvious but may be obtained via the Lambert function[2] (Chatzigeorgiou 2013). For purposes of simplicity, suppose $a = 0$ and $b = 1$. Then, we have the following.

$$\frac{\log(K)}{K} \le \epsilon \Leftrightarrow \frac{-\log(K)}{K} \ge -\epsilon$$

$$\Leftrightarrow W_{-1}\left(\frac{-\log(K)}{K}\right) \le W_{-1}(-\epsilon), \text{ since } W_{-1}(\cdot) \text{ is decreasing.}$$

But $W_{-1}(-\frac{\log(x)}{x}) = -\log(x)$ for $x > e$. Consequently, we have that

$$-\log(K) \le W_{-1}(-\epsilon) \Leftrightarrow K \ge e^{-W_{-1}(-\epsilon)}.$$

By definition of the Lambert function, we have that $e^{W(x)} = \frac{x}{W(x)}$, implying that

$$K \ge e^{-W_{-1}(-\epsilon)} = \frac{W_{-1}(-\epsilon)}{\epsilon} \ge \mathcal{O}\left(\frac{\log(\epsilon)}{-\epsilon}\right) = \mathcal{O}\left(\frac{1}{\epsilon}\log(1/\epsilon)\right).$$

Here, the first inequality follows from (3) in (Chatzigeorgiou 2013). Hence, the oracle complexity for $a = 1$ is $\mathcal{O}\left(\frac{\log^2(1/\epsilon)}{\epsilon^2}\right)$, which is near optimal (optimal is $\mathcal{O}(1/\epsilon^2)$). □

We now consider two cases of Theorem 4 for which similar rate statements are available.

Case 1 (Structured Stochastic Nonsmooth Optimization with *f* Smooth). Now, consider Problem (1), in which $f(x)$ is a smooth function. Recall that we consider such a problem in Section 2 for strongly convex f, and in this case, we consider the merely convex case. When f is deterministic, accelerated gradient methods first proposed by Nesterov (1983) and their proximal generalizations suggested by Beck and Teboulle (2009) are characterized by the optimal rate of convergence of $\mathcal{O}(1/K^2)$. When f is expectation-valued, Ghadimi and Lan (2016) present the first known accelerated scheme for stochastic convex optimization for which the optimal rate of $1/k^2$ is shown for the expected suboptimality error. This rate required choosing the simulation length K and choosing $N_k = \lfloor k^2 K \rfloor$, which led to the optimal oracle complexity of $\mathcal{O}(1/\epsilon^2)$. However, this method is somewhat different from VS-APM. In particular, every step requires two prox evaluations (rather than one for VS-APM).[3] Jofré and Thompson (2017) develop an accelerated proximal scheme for convex problems with a similar algorithm but allow for state-dependent noise. The weakening of the noise requirement still allows for deriving the optimal rate of $\mathcal{O}(1/K^2)$ but necessitates choosing $N_k = \lfloor k^3(\ln k) \rfloor$. As a consequence, the oracle complexity is slightly poorer than the optimal level and is given by $\mathcal{O}(\epsilon^{-2}\ln^2(\epsilon^{-0.5}))$. We note that VS-APM displays the optimal oracle complexity $\mathcal{O}(\epsilon^{-2})$ by choosing $N_k = \lfloor k^2 K \rfloor$, whereas by choosing $N_k = \lfloor k^a \rfloor$ for $a = 3+\delta$, then the oracle complexity can be made arbitrarily close to optimal and is given by $\mathcal{O}(\epsilon^{-2-\delta/2})$. However, VS-APM imposes a stronger assumption on noise as formalized next.

Corollary 2 (Rate and Oracle Complexity Bounds with Smooth *f* for VS-APM). *Suppose Assumptions 2, 3, and 7 hold. Suppose $\gamma_k = \gamma \le 1/2L$ for all k.*

i. *Let $N_k = \lfloor k^a \rfloor$, where $a = 3 + \delta$ and $\hat{C} \triangleq \frac{2\nu^2\gamma(a-2)}{a-3} + \frac{4C^2}{\gamma}$. Then, the following holds.*

$$\mathbb{E}[F(y_{K+1} - F(x^*))] \le \frac{\hat{C}}{K^2} \text{ for all } K \text{ and } \sum_{k=1}^{K(\epsilon)} N_k \le \mathcal{O}\left(\frac{1}{\epsilon^{2+\delta/2}}\right),$$

where $\mathbb{E}[F(y_{K(\epsilon)+1}) - F(x^*)] \le \epsilon$.

ii. *Given a $K > 0$, let $N_k = \lfloor k^2 K \rfloor$, where $a > 3$ and $\tilde{C} \triangleq 2\nu^2\gamma + \frac{4C^2}{\gamma}$. Then, the following holds.*

$$\mathbb{E}[F(y_{K+1} - F(x^*))] \le \frac{\tilde{C}}{K^2} \text{ and } \sum_{k=1}^{K} N_k \le \mathcal{O}\left(\frac{1}{\epsilon^2}\right), \text{ where } \mathbb{E}[F(y_{K+1}) - F(x^*)] \le \epsilon.$$

Proof.

(i) Similar to the proof of Lemma 5, by defining $\delta_k = F(y_k) - F(x^*)$ we can prove

$$\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \frac{2\nu^2\gamma}{K^2}\sum_{k=1}^{K}\frac{k^2}{k^a} + \frac{4C^2}{\gamma K^2}.$$

Let $N_k = \lfloor k^a \rfloor \ge \frac{1}{2}k^a$ and $\gamma_k = \gamma$. Then, we have that the following holds in which $\hat{C} \triangleq \frac{2\nu^2\gamma(a-2)}{a-3} + \frac{4C^2}{\gamma}$.

$$\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \frac{2\nu^2\gamma}{K^2}\sum_{k=1}^{K}\frac{k^2}{k^a} + \frac{4C^2}{\gamma K^2} \le \frac{2\nu^2\gamma(a-2)}{(a-3)K^2} + \frac{4C^2}{\gamma K^2} = \frac{\hat{C}}{K^2}, \tag{38}$$

where the first inequality follows from bounding the summation as follows:

$$\sum_{k=1}^{K} k^{2-a} = 1 + \sum_{k=2}^{K} k^{2-a} \le 1 + \int_1^K x^{2-a}dx = \frac{1}{a-3} - \frac{K^{3-a}}{a-3} + 1 \le \frac{1}{a-3} + 1 = \frac{a-2}{a-3}.$$

Suppose y_{K+1} satisfies $\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \epsilon$, implying that $\frac{\hat{C}}{K^2} \le \epsilon$ or $K = \lceil \hat{C}^{1/2}/\epsilon^{1/2} \rceil$. If $\epsilon \le \hat{C}/2$, then the oracle complexity can be bounded as follows:

$$\sum_{k=1}^{K} N_k \le \sum_{k=1}^{K} k^a = \sum_{k=1}^{1+\sqrt{\hat{C}/\epsilon}} k^a \le \int_0^{2+\sqrt{\hat{C}/\epsilon}} k^a da = \frac{\left(2 + \sqrt{\hat{C}/\epsilon}\right)^{1+a}}{1+a} \le \left(\frac{\sqrt{\hat{C}}}{2\sqrt{\epsilon}}\right)^{1+a} = \mathcal{O}\left(\frac{1}{\epsilon^{2+\delta/2}}\right).$$

(ii) Let $N_k = \lfloor k^2 K \rfloor \ge \frac{1}{2}k^2 K$. Then, similar to part (i), we may bound the expected suboptimality as follows in which $\tilde{C} \triangleq 2\nu^2\gamma + \frac{4C^2}{\gamma}$.

$$\mathbb{E}[F(y_{K+1}) - F(x^*)] \le \frac{2\nu^2\gamma}{K^2}\sum_{k=1}^{K}\frac{k^2}{k^2 K} + \frac{4C^2}{\gamma K^2} = \frac{2\nu^2\gamma}{K^2} + \frac{4C^2}{\gamma K^2} \le \frac{\tilde{C}}{K^2}.$$

Because $K = \lceil \tilde{C}^{1/2} / \epsilon^{1/2} \rceil$, the oracle complexity may be bounded as follows:

$$\sum_{k=1}^{K} N_k \le \sum_{k=1}^{K} k^2 K = \frac{1}{6}K^2(K+1)(2K+1) = \frac{1}{6}K^2(2K^2 + 3K + 1) \le K^4 \le \mathcal{O}\left(\frac{1}{\epsilon^2}\right). \quad \square$$

Case 2 (Deterministic Nonsmooth Convex Optimization). When the function *f* in (1) is deterministic but possibly nonsmooth, Nesterov (2005b) shows that, applying an accelerated scheme to a suitably smoothed problem (with a fixed smoothing parameter) leads to a convergence rate of $\mathcal{O}(1/K)$. In contrast with Theorem 4, utilizing a fixed

Table 2. Bounding the Second Moments for Certain Smoothings

$\tilde f(x,\omega)$	$\tilde f_\eta(x,\omega)$	$\nabla \tilde f_\eta(x,\omega)$	$\mathbb{E}[\|\nabla_x \tilde f_\eta(x,\omega) - \nabla_x f_\eta(x)\|^2]$
$\tilde f_1(x,\omega) = \lambda(\omega)\|x\|_1$	$\sum_{i=1}^n h_\eta(x_i,\omega)$, where $h_\eta(x_i,\omega) =$ $\begin{Bmatrix} \lambda^2(\omega)\frac{x_i^2}{2\eta}, & \lambda(\omega)\lvert x_i\rvert < \eta \\ \lambda(\omega)\lvert x_i\rvert - \eta/2, & o.w. \end{Bmatrix}$	$[\nabla_{x_i} h_\eta(x_i,\omega)]_{i=1}^n$, where $\nabla_{x_i} h_\eta(x_i,\omega) =$ $\begin{Bmatrix} \lambda^2(\omega)\frac{x_i}{\eta}, & \lambda(\omega)\lvert x_i\rvert < \eta \\ \lambda(\omega)x_i/\lvert x_i\rvert, & o.w. \end{Bmatrix}$	$4n\mathbb{E}[\lambda^2(\omega)]$
$\tilde f_2(x,\omega) = \lambda(\omega)\|x\|_2$	$\sqrt{\lambda^2(\omega)\|x\|^2 + \eta^2} - \eta$	$\frac{\lambda^2(\omega)x}{\sqrt{\lambda^2(\omega)\|x\|^2+\eta^2}}$	$4\mathbb{E}[\lambda^2(\omega)]$
$\tilde f_3(x,\omega) = \max_{1\le i\le n}\{h_i(x,\omega)\}$ where $h_i(x,\omega) = v_i + s_i c(\omega)^T x$	$\eta \log\left(\sum_{i=1}^n \exp(h_i(x,\omega)/\eta)\right)$	$\frac{\sum_{i=1}^n \nabla_x h_i(x,\omega)\exp(h_i(x,\omega)/\eta)}{\sum_{i=1}^n \exp(h_i(x,\omega)/\eta)}$	$4\mathbb{E}[(\max_{1\le i\le n}\|s_i c(\omega)\|)^2]$,

smoothing parameter leads to an approximate solution at best, and such a scheme is not characterized by asymptotic convergence guarantees. In addition, we observe that the rate statement for the deterministic counterpart of sVS-APM, denoted by s-APM, is global (valid for all k), whereas any statement with constant smoothing holds for the prescribed K. We observe that the rate statements by using an appropriately chosen smoothing and step length parameter matches that by using a selecting a suitable smoothing and step length sequence.

Corollary 3 (Iterative vs. Constant Smoothing for Deterministic Nonsmooth Convex Optimization). *Consider* (1) *and assume f(x) is a deterministic function. Suppose Assumption* 7 *holds. (i) Iterative smoothing: suppose* $\gamma_k = 1/2k$ *and* $\eta_k = 1/k$. *Then,* $F(y_{k+1}) - F(x^*) \le \frac{4C^2+B^2}{k}$, *for all* $k > 0$. *(ii) Fixed smoothing: for a given* $K > 0$, *suppose* $\eta_k = 1/K$ *and* $\gamma_k = 1/2K$. *Then,* $F(y_{K+1}) - F(x^*) \le \frac{4C^2+B^2}{K}$.

Remark 4. By recalling that $f_\eta(x) \triangleq \mathbb{E}[\tilde f_\eta(x,\omega)]$, by using theorem 7.47 in Shapiro et al. (2009) (interchangeability of the derivative and the expectation), and noting that $\tilde f_\eta(\cdot,\omega)$ is differentiable in x for every ω, we have $\nabla f_\eta(x) = \nabla\mathbb{E}[\tilde f_\eta(x,\omega)] = \mathbb{E}[\nabla \tilde f_\eta(x,\omega)] \Rightarrow \mathbb{E}[\nabla f_\eta(x) - \nabla \tilde f_\eta(x,\omega)] = 0$. Therefore, such a gradient estimator is unbiased, and our assumption holds. We now derive bounds on the second moments for some common smoothings in Table 2.

3.3. Almost-Sure Convergence

Whereas the previous section focuses on providing rate statements for expected suboptimality, we now consider the open question of whether the sequence of iterates produced by sVS-APM converges almost sure to a solution. Schemes employing a constant smoothing parameter preclude such guarantees. Proving almost sure convergence requires using the following lemma.

Lemma 6 (Supermartingale Convergence Lemma; Polyak 1987). *Let* $\{v_k\}$ *be a sequence of nonnegative random variables, in which* $\mathbb{E}[v_0] < \infty$, *and let* $\{\alpha_k\}$ *and* $\{\eta_k\}$ *be deterministic scalar sequences such that* $0 \le \alpha_k \le 1$ *and* $\eta_k \ge 0$ *for all* $k \ge 0$, $\sum_{k=0}^\infty \alpha_k = \infty$, $\sum_{k=0}^\infty \eta_k < \infty$, *and* $\lim_{k\to\infty} \frac{\eta_k}{\alpha_k} = 0$, *and* $\mathbb{E}[v_{k+1} \mid \mathcal{H}_k] \le (1-\alpha_k)v_k + \eta_k$ *a.s. for all* $k \ge 0$. *Then,* $v_k \to 0$ *a.s. as* $k \to \infty$.

Proposition 2 (almost sure Convergence of sVS-APM). *Suppose Assumptions* 3 *and* 7 *hold and* $\{y_k\}$ *is a sequence generated by sVS-APM. Suppose* $\gamma_k = k^{-b} < \eta_k$, *where* $b \in (0,1/2]$, $\{\eta_k\}$ *is a decreasing sequence, and* $N_k = \lfloor k^a \rfloor$ *such that* $(a+b) > 1$. *Then,* $\{y_k\}$ *converges to a solution of* (1) *a.s.*

Proof. From Inequality (34), we have that the following holds:

$$\gamma_k \delta_{k+1} \le \frac{\lambda_{k-1}^2}{\lambda_k^2}\gamma_k\delta_k + \frac{1}{2\lambda_k^2}(\|u_k\|^2 - \|u_{k+1}\|^2) + \left(\frac{\gamma_k}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\right)\|\bar w_{k,N_k}\|^2 - \frac{1}{\lambda_k}\bar w_{k,N_k}^T u_k$$

$$\le \frac{\lambda_{k-1}^2}{\lambda_k^2}\gamma_{k-1}\delta_k + \frac{1}{2\lambda_k^2}(\|u_k\|^2 - \|u_{k+1}\|^2) + \left(\frac{\gamma_k}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\right)\|\bar w_{k,N_k}\|^2 - \frac{1}{\lambda_k}\bar w_{k,N_k}^T u_k.$$

Dividing both sides of the previous inequality by γ_k, we obtain the following relationship:

$$\delta_{k+1} + \frac{1}{2\gamma_k\lambda_k^2}\|u_{k+1}\|^2 \leq \frac{\lambda_{k-1}^2}{\lambda_k^2\gamma_k}\gamma_{k-1}\delta_k + \frac{1}{2\gamma_k\lambda_k^2}\|u_k\|^2 + \left(\frac{1}{\dfrac{2}{\gamma_k} - \dfrac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\gamma_k\lambda_k}\bar{w}_{k,N_k}^T u_k$$

$$= \frac{\lambda_{k-1}^2\gamma_{k-1}}{\lambda_k^2\gamma_k}\left(\delta_k + \frac{\|u_k\|^2}{2\gamma_{k-1}\lambda_{k-1}^2}\right) + \left(\frac{1}{\dfrac{2}{\gamma_k} - \dfrac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\gamma_k\lambda_k}\bar{w}_{k,N_k}^T u_k.$$

By defining $v_{k+1} \triangleq \delta_{k+1} + \frac{1}{2\gamma_k\delta_k^2}\|u_{k+1}\|^2$ and $\alpha_k \triangleq 1 - \frac{\lambda_{k-1}^2\gamma_{k-1}}{\lambda_k^2\gamma_k}$, we have the following recursion.

$$v_{k+1} \leq (1-\alpha_k)v_k + \left(\frac{1}{\dfrac{2}{\gamma_k} - \dfrac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\gamma_k\lambda_k}\bar{w}_{k,N_k}^T u_k \iff$$

$$v_{k+1} + \eta_k B^2 \leq (1-\alpha_k)(v_k + \eta_{k-1}B^2) + \eta_k B^2 - (1-\alpha_k)\eta_{k-1}B^2 + \left(\frac{1}{\dfrac{2}{\gamma_k} - \dfrac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\gamma_k\lambda_k}\bar{w}_{k,N_k}^T u_k. \tag{39}$$

Let $\bar{v}_{k+1} \triangleq v_{k+1} + \eta_k B^2$. From $(1, B^2)$ smoothability and the decreasing nature of $\{\eta_k\}$,

$$0 \leq F(y_{k+1}) - F(x^*) \leq F_{\eta_{k+1}}(y_{k+1}) - F_{\eta_{k+1}}(x^*) + \eta_{k+1}B^2 \leq F_{\eta_{k+1}}(y_{k+1}) - F_{\eta_{k+1}}(x^*) + \eta_k B^2.$$

Then, (39) can be rewritten as follows:

$$\bar{v}_{k+1} \leq (1-\alpha_k)\bar{v}_k + \eta_k B^2 - (1-\alpha_k)\eta_{k-1}B^2 + \left(\frac{1}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\gamma_k\lambda_k}\bar{w}_{k,N_k}^T u_k.$$

Recall, by the definition of λ_k, we have $\lambda_{k-1}^2 = \frac{(2\lambda_k - 1)^2 - 1}{4}$ and $\frac{k}{2} \leq \lambda_k \leq k$ if $\gamma_k = k^{-b}$, $b \in (0, 1/2]$, we obtain the following relationship:

$$\alpha_k = 1 - \frac{\lambda_{k-1}^2\gamma_{k-1}}{\lambda_k^2\gamma_k} = 1 - \frac{\gamma_{k-1}(4\lambda_k^2 - 4\lambda_k)}{4\lambda_k^2\gamma_k} = \frac{\lambda_k^2\gamma_k - \gamma_{k-1}\lambda_k^2 + \gamma_{k-1}\lambda_k}{\lambda_k^2\gamma_k} = \frac{\gamma_k - \gamma_{k-1}}{\gamma_k} + \frac{\gamma_{k-1}}{\lambda_k\gamma_k}$$

$$\geq \frac{k^{-b} - (k-1)^{-b}}{k^{-b}} + \frac{(k-1)^{-b}}{k^{1-b}} = \frac{k^{1-b} - (k-1)^{1-b}}{k^{1-b}} \geq \frac{(1-b)}{k}, \quad b \in (0, 1/2], \tag{40}$$

where in the last inequality we use $b \in (0, 1/2]$:

$$k\left(\frac{k^{1-b} - (k-1)^{1-b}}{k^{1-b}}\right) = k - k\left(\frac{k-1}{k}\right)^{1-b} = k - k^b(k-1)^{1-b} = k - (k-1)\left(\frac{k}{k-1}\right)^b$$

$$= k - (k-1)\left(1 + \frac{1}{k-1}\right)^b = k - (k-1) - b - \frac{b(b-1)}{2!(k-1)^2} - \frac{b(b-1)(b-2)}{3!(k-1)^3} - \ldots$$

$$= (1-b) + \frac{b(1-b)}{2!(k-1)^2}\left(1 - \frac{(2-b)}{3(k-1)}\right) + + \frac{b(1-b)(2-b)(3-b)}{4!(k-1)^4}\left(1 - \frac{(4-b)}{5(k-1)}\right) + \ldots$$

$$\geq (1-b), \quad \text{since} \quad k \geq 2 \geq 1 + \max\left\{\frac{2}{3}, \frac{4}{5}, \frac{6}{7}, \ldots\right\}.$$

Table 3. Example 1: mVS-APM vs. SSG (L), mVS-PM vs. SSG (R)

	SSG	$\|y_k - x^*\|$ for mVS-APM					SSG	mVS-PM
μ	$\|y_k - x^*\|$	$\eta = \eta^*$	$\eta = 0.1$	$\eta = 1$	$\eta = 10$	μ	$\|y_k - x^*\|$	$\|y_k - x^*\|$
1	7.8609e-4	2.8078e-1	2.2150e-2	4.7893e-3	1.9443e-2	1	2.0847e-1	3.0971e-2
1e-1	9.9114e-1	3.3207e-3	3.7247e-2	5.8973e-3	1.8865e-2	1e-1	2.4283	9.5149e-2
1e-2	3.0611	3.7218e-2	8.3083e-2	7.3432e-3	3.6886e-2	1e-2	4.2409	1.5115e-1
1e-3	4.0682	1.3893	1.7692e-1	4.7901e-3	5.2147e-2	1e-3	4.4784	1.8033e-1
1e-4	6.3783	2.7269	4.7065e-1	5.5248e-3	6.3872e-2	1e-4	4.5028	1.7261e-1

By taking conditional expectations and recalling that $\eta_k = c\gamma^k$, where $c > 1$, we obtain the following:

$$\begin{aligned}\mathbb{E}[\bar{v}_{k+1} \mid \mathcal{H}_k] &\le (1-\alpha_k)\bar{v}_k + \eta_k B^2 - (1-\alpha_k)\eta_{k-1}B^2 + \left(\frac{1}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\right)\frac{\nu^2}{N_k} \\ &\le (1-\alpha_k)v_k + \eta_k B^2 - (1-\alpha_k)\eta_{k-1}B^2 + \left(\frac{c}{2(c-1)}\right)\frac{\gamma_k \nu^2}{N_k}.\end{aligned}$$

If $\gamma_k = k^{-b}$, where $b \in (0, 1/2]$ and $N_k = \lfloor k^a \rfloor$, where $a + b > 1$, by Lemma A.1, we have that $\sum_{k=1}^{\infty} \frac{\gamma_k \nu^2}{N_k} < \infty$, and the following holds for $\eta_k = ck^{-b}$, $c > 1$, and $b \in (0, 1/2]$:

$$\begin{aligned}\eta_k - (1-\alpha_k)\eta_{k-1} &= \eta_k - \frac{\lambda_{k-1}^2 \gamma_{k-1}}{\lambda_k^2 \gamma_k}\eta_{k-1} = ck^{-b} - \left(1 - \frac{1}{\lambda_k}\right)\frac{c(k-1)^{-2b}}{k^{-b}} \\ &\le ck^{-b} - \left(1 - \frac{1}{\lambda_k}\right)ck^{-b} \le \frac{2c}{k^{1+b}} \Rightarrow \sum_{k=1}^{\infty}(\eta_k B^2 - (1-\alpha_k)\eta_{k-1}B^2) < \infty.\end{aligned}$$

Furthermore, from (40), it follows that $\sum_{k=1}^{\infty} \alpha_k = \infty$ and

$$\lim_{k\to\infty}\left(\frac{1}{\alpha_k}\right)\left(\frac{c}{2(c-1)}\right)\left(\frac{\nu^2}{k^{a+b}}\right) \le \lim_{k\to\infty}\left(\frac{c}{2(c-1)}\right)\left(\frac{\nu^2}{(1-b)k^{a+b-1}}\right) = 0$$

Figure 2. Example 1: (mVS-APM) vs. (SSG) for $\mu = 0.1$

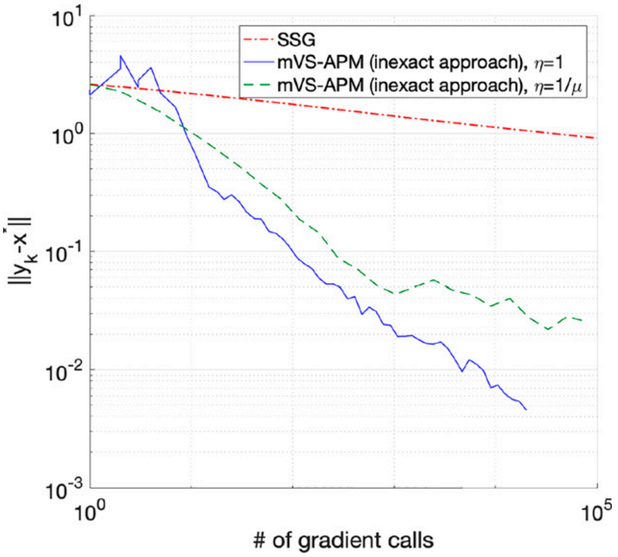


Table 4. Example 1: Comparing mVS-APM vs. SSG: Different Std (L), Different n (R)

	SSG		mVS-APM				SSG		mVS-APM		
Std.	$\|y_k - x^*\|$	Time	η	$\|y_k - x^*\|$	Time	n	$\|y_k - x^*\|$	Time	η	$\|y_k - x^*\|$	Time
1e+1	1.6691	5.8269	1	5.6007e-1	2.9858	20	9.1148e-1	5.9096	1	5.8973e-3	3.8961
1	9.4759e-1	5.9375	1	5.1574e-2	2.9925	30	1.5326	6.117	1	5.9034e-3	3.2213
1e-1	9.1148e-1	5.9096	1	5.8973e-3	3.8961	40	8.5934e-1	6.2494	1	6.0096e-3	3.6658
1e-2	9.1285e-1	5.9444	1	5.7294e-4	3.0362	50	3.6236	6.4209	1	6.3496e-3	3.3903

for $b \in (0, 1/2]$ and $a + b > 1$. Additionally, we have the following:

$$\lim_{k\to\infty} \frac{\eta_k B^2 - (1-\alpha_k)\eta_{k-1}B^2}{\alpha_k} = \lim_{k\to\infty} \frac{ck^{-b}B^2 - c(1-\alpha_k)(k-1)^{-b}B^2}{\alpha_k}$$
$$\leq \lim_{k\to\infty} \frac{ck^{-b}B^2 - c(1-\alpha_k)k^{-b}B^2}{\alpha_k} = \lim_{k\to\infty} \frac{cB^2}{k^b} = 0,$$

where $\eta_k B^2 - (1-\alpha_k)\eta_{k-1}B^2 \geq 0$ can be concluded as follows. For any $b \in (0, 1/2]$, we have

$$\frac{\lambda_{k-1}^2}{\lambda_k^2} = \left(1 - \frac{1}{\lambda_k}\right) \leq \frac{k-1}{k} \leq \frac{(k-1)^{2b}}{k^{2b}} \Rightarrow \frac{\lambda_{k-1}^2}{\lambda_k^2} \frac{k^b}{(k-1)^b} \leq \frac{(k-1)^b}{k^b} \Rightarrow \frac{\lambda_{k-1}^2 \gamma_{k-1}}{\lambda_k^2 \gamma_k} \leq \frac{\eta_k}{\eta_{k-1}}$$
$$\Rightarrow (1-\alpha_k) \leq \frac{\eta_k}{\eta_{k-1}} \Rightarrow \eta_k - (1-\alpha_k)\eta_{k-1} \geq 0.$$

Therefore, Lemma 5 can be applied, and $\bar{v}_k = F_{\eta_k}(x_k) - F_{\eta_k}(x^*) + \eta_k B^2 \to 0$ a.s. By $(1, B^2)$ smoothness of f, $0 \leq F(x_k) - F(x^*) \leq F_{\eta_k}(x_k) - F_{\eta_k}(x^*) + \eta_k B^2$, implying that $F(x_k) \to F(x^*)$ a.s. □

The next proposition provides a similar a.s. convergence for VS-APM that can accommodate structured nonsmooth optimization in which $f(x)$ is a smooth merely convex function. The proof of this result is similar to Proposition 2, but δ_k in this case is defined as $\delta_k = F(y_k) - F(x^*)$.

Proposition 3 (Almost Sure Convergence Theory for VS-APM). *Suppose Assumptions 2, 3, and 7 hold. Suppose $\{y_k\}$ defines a sequence generated by VS-APM. Suppose $\gamma_k = \gamma \leq 1/(2L)$ and $N_k = \lfloor k^a \rfloor$ for $a > 1$. Then, $\{y_k\}$ converges to a solution of (1) almost surely.*

4. Numerical Results

We now compare the performance of (mVS-APM) and sVS-APM with existing solvers on Matlab running on a 64-bit MacOS 10.13.3 with Intel i7-7Y75 @1.4 GHz with 16 GB RAM.

4.1. (mVS-APM): Strongly Convex and Nonsmooth *f*

Example 1. Consider the following constrained problem:

$$\min_{x\in[-1,1]} f(x), \text{ where } f(x) \triangleq \mathbb{E}\left[\frac{1}{2}x^T A(\omega)x + \beta(\omega)^T x + \lambda(\omega)\|x\|_1\right], \tag{41}$$

$A(\omega) = \bar{A} + W \in \mathbb{R}^{n\times n}$ and the elements of W have an independent and identically distributed (i.i.d.) normal distribution with mean zero and standard deviation (std) 0.1. Similarly, $\beta(\omega) = \bar{\beta} + w \in \mathbb{R}^n$, where w is a random vector.

Table 5. Example 2: Comparing (mVS-APM) vs. (SSG)

	SSG		mVS-APM		
μ	$\|y_k - x^*\|$	time	η	$\|y_k - x^*\|$	Time
1	4.4908e-3	4.3883	$1/\mu = 1$	5.8314e-3	1.5191
1e-1	2.7134e-1	3.8794	1	1.0102e-2	1.1964
1e-2	8.7266e-1	3.9742	1	1.8236e-2	1.2065
1e-3	9.8723e-1	4.0129	1	3.8619e-2	1.1510
1e-4	9.9872e-1	4.0684	1	7.1652e-2	1.1490

Table 6. Example 2: Comparing mVS-APM vs. SSG: Different Std (L), Different n (R)

	SSG			mVS-APM			SSG			mVS-APM	
Std.	$\|y_k - x^*\|$	Time	η	$\|y_k - x^*\|$	Time	n	$\|y_k - x^*\|$	Time	η	$\|y_k - x^*\|$	Time
1e+1	9.8253e-1	3.8733	1	9.6709e-1	1.1661	20	2.7134e-1	3.8794	1	1.0102e-2	1.1964
1	2.7134e-1	3.8794	1	1.0102e-2	1.1964	30	3.5948e-1	4.0277	1	1.2010e-2	1.2594
1e-1	2.1394e-1	3.9304	1	8.6589e-3	1.1083	40	5.3537e-1	4.0418	1	7.4431e-3	1.3467
1e-2	2.1813e-1	3.9134	1	1.1027e-1	1.1270	50	2.6880e-1	4.1198	1	8.2670e-3	1.3452

Because tractable prox evaluations are not available for (41), we compute approximate gradients $\nabla_x f_\eta$ using (SSG). We set $N_k = \lfloor \rho^{-k} \rfloor$, where $\rho \triangleq \left(1 - \frac{1}{2a\sqrt{\kappa}}\right)$ and $a = 2.01$. Using a budget of 1e5 and 10 replications, we provide results in Table 3 (L), whereas Figure 2 shows the behavior of (mVS-APM) with different smoothing parameters η versus (SSG). When the strong convexity modulus μ is small, (mVS-APM) performs significantly better than (SSG) and is far more stable. For instance, when $\eta = 1$, (mVS-APM) terminates with an empirical error of approximately 4.8e-3 and 5.5e-3 for $\mu = 1$ and $\mu = 1e\text{-}4$, whereas corresponding errors for (SSG) are 7.8e-3 to 6.3. As one can see, $\eta = 1$ for (mVS-APM) seems to be a reasonable practical choice for different problem settings. Note that, in this table, η^* is chosen according to Lemma 3, and we note that, as $\mu \ll 1$, the benefit of utilizing η^* is muted. Next, we consider the unconstrained variant (41), in which $x \in \mathbb{R}^n$. Because the subgradient is unbounded, we use the unaccelerated method mVS-PM. In Table 3 (R), the behavior of mVS-PM is compared with (SSG) for different choices of μ. As suggested after Theorem 3, we set $\eta = \frac{1}{\mu} + 1e\text{-}3 > \frac{1}{\mu}$.

In Table 4, we compare (mVS-APM) with (SSG) for different choices of standard deviation of noise and dimension (n). In Table 4 (L), we set $\mu = 0.1$ and $n = 20$, whereas in Table 4 (R), we set $\mu = 0.1$ and standard deviation is 0.1. We run both schemes with a total budget in subgradient evaluations of 1e5 and 10 replications and observe that (mVS-APM) outperforms (SSG).

Example 2. We revisit this comparison using a stochastic utility problem.

$$\min_{\|x\| \le 1} \mathbb{E}\left[\phi\left(\sum_{i=1}^{n}\left(\frac{i}{n} + \omega_i\right)x_i\right)\right] + \frac{\mu}{2}\|x\|^2,$$

where $\phi(t) \triangleq \max_{1 \le j \le m}(v_i + s_i t)$, ω_i are i.i.d. normal random variables with mean zero and variance one and $v_i, s_i \in (0,1)$. Table 5 shows similar behavior as in Example 1. In Table 6, we compare (mVS-APM) with (SSG) for different choices of standard deviation and dimension (n). In Table 6 (L), we set $\mu = 0.1$, whereas $n = 20$, and in Table 6 (R), we set $\mu = 0.1$ and standard deviation is one. Similar to Example 1, (mVS-APM) outperforms (SSG) in all cases.

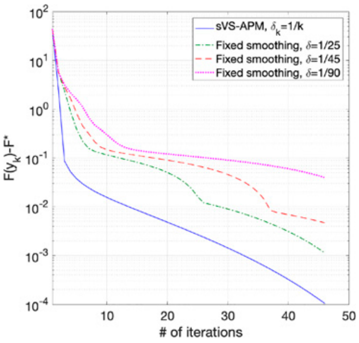
4.2. sVS-APM: Convex and Smoothable *f*

Example 3. In this setting, we compare the performance of sVS-APM for merely convex problems on Example 2 with $\mu = 0$. The δ-smoothed approximation of $\phi(t)$ provided by Beck and Teboulle (2012) is given by $\phi_\delta(t) = \delta \log\left(\sum_{i=1}^{m} e^{(v_i + s_i t)/\delta}\right)$. In Table 7, we generate 20 replications for sVS-APM with fixed and diminishing smoothing

Table 7. Example 3: Comparing (sVS-APM) with Fixed Smoothing

			sVS-APM		Fixed smooth.
n	m	δ_k	$\mathbb{E}[f(y_k) - f^*]$	δ	$\mathbb{E}[f(y_k) - f^*]$
20	10	$1/k$	1.832e-4	$1/K$	3.455e-3
		$1/(2k)$	3.014e-3	$1/(2K)$	2.157e-2
		$1/(3k)$	1.269e-2	$1/(3K)$	6.079e-2
100	25	$1/k$	1.944e-3	$1/K$	3.126e-2
		$1/2k$	1.181e-2	$1/2K$	5.130e-2
		$1/3k$	2.411e-2	$1/3K$	5.817e-2
200	10	$1/k$	1.067e-4	$1/K$	4.695e-3
		$1/2k$	5.173e-3	$1/2K$	3.957e-2
		$1/3k$	1.594e-2	$1/3K$	6.929e-2

Figure 3. Example 3: sVS-APM vs. Fixed Smoothing; $n = 200$



sequences with $\eta_k = \delta_k/2$, $N_k = \lfloor k^{3.001} \rfloor$, and sampling budget is 1e6. In Figure 3, we compare trajectories for sVS-APM with those for constant smoothing for $n = 200$.

4.3. Key Observations

The empirical behavior of sVS-APM appears to be better on this test problem. One rationale for this may be drawn from noting that sVS-APM allows for larger step lengths early (because $\eta_k \le \delta_k$), whereas in the fixed smoothing technique, $\eta_k \le \delta_k$ (δ_k may be quite small). This can be seen in the trajectories in which early progress by the iterative smoothing scheme can be observed. A larger δ_k allows for larger step lengths but leads to a coarser approximation of the original problem, whereas smaller δ_k leads to poorer progress but better approximations (see Table 7 and Figure 3).

4.4. Almost Sure Convergence

Next, we implement sVS-APM on the stochastic utility problem with $n = 20$ and $m = 10$ for different choices of the smoothing sequences. Specifically, we allow δ_k to be $\delta_k \in \{1/k, 1/\sqrt{k}, 1/k^{0.25}\}$ ($\delta_k = 1/k$ is required for convergence in mean and $\delta_k = 1/k^b$ with $b \in (0, 1/2]$ for a.s. convergence). We employ $N_k = \lfloor k^{3.001} \rfloor$. For each experiment, the mean of 20 replications and their 95% confidence intervals are plotted in Figures 4 and 5. It can be seen that, when $\delta_k \to 0$ at a slower rate as mandated by the requirement of the a.s. convergence result, the confidence bands are tighter, becoming more apparent in Figure 4 in which the variance is five. Furthermore, our numerical studies reveal that, even for less aggressive choices of N_k such as when $N_k = k^a$ and $a > 1$, the trajectories show the desired behavior in accordance with Proposition 2.

5. Concluding Remarks

Drawing motivation from the often poor behavior of (SSG) schemes on general (rather than structured) nonsmooth stochastic convex optimization problems, we develop two sets of accelerated proximal variance-reduced schemes, both of which rely on a variable sample-size accelerated proximal method (VS-APM) for smooth convex problems. In nonsmooth strongly convex regimes, we present three sets of schemes, each of which produces linearly convergent sequences and is characterized by an overall complexity in subgradients (or proximal

Figure 4. Almost Sure Convergence for sVS-APM, $N_k = \lfloor k^{3.001} \rfloor$, $\nu^2 = 5$

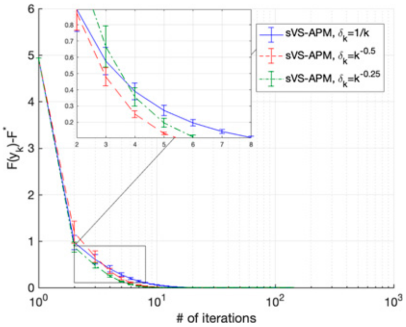
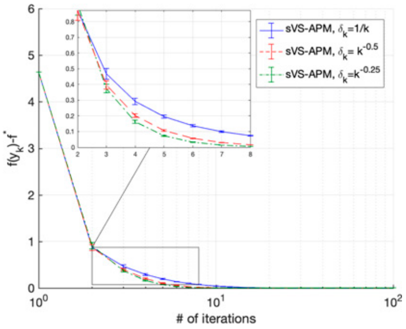


Figure 5. Almost Sure Convergence for sVS-APM, $N_k = \lfloor k^{3.001} \rfloor$, $\nu^2 = 2$



evaluations in the third case) that is optimal (or near optimal). First, in compact domains, we propose (mVS-APM), an avenue that requires applying VS-APM on the Moreau envelope of F, where increasingly exact gradients are computed via an inner (SSG) scheme. Second, in unbounded domains, we apply an unaccelerated variable sample-size proximal method (VS-PM), which also relies on (SSG) for approximating gradients to increasing accuracy. When $\tilde{f}(\cdot,\omega)$ is smoothable and convex, our smoothed VS-APM scheme (or sVS-APM) admits optimal rate and oracle complexity. Our findings, when specialized to the smooth and convex f, provide an optimal accelerated rate of $\mathcal{O}(1/K^2)$ with optimal oracle complexity matching findings by Ghadimi and Lan (2016) and Jofré and Thompson (2017). When f is deterministic, our rate matches that obtained by Nesterov (2005b) but does so while providing asymptotically convergent schemes. Preliminary numerics suggest that the schemes compare well with existing techniques in terms of both complexity as well as sensitivity to problem parameters.

Appendix

Lemma A.1. *For any real number $y \ge 1$, we have that $\lfloor y \rfloor \ge \lceil \frac{1}{2} y \rceil$.*

Proof. Let $T = \lfloor y \rfloor$. If T is an even number, then we have $\lceil \frac{1}{2} y \rceil = \lceil \frac{1}{2}(T+\epsilon) \rceil = \frac{T}{2} + 1$, where $\epsilon \in (0,1)$. Because $T \ge \frac{T}{2} + 1$, $\lfloor y \rfloor \ge \lceil \frac{1}{2} y \rceil$. If T is an odd number, we have $\lceil \frac{1}{2} y \rceil = \lceil \frac{T-1}{2} + \frac{\epsilon+1}{2} \rceil = \frac{T-1}{2} + 1 = \frac{T+1}{2}$. Again, because $T \ge \frac{T+1}{2}$, we have that $\lfloor y \rfloor \ge \lceil \frac{1}{2} y \rceil$. □

Lemma A.2. *Given a symmetric positive definite matrix Q, we have the following for any v_1, v_2, v_3: $(v_2 - v_1)^T Q (v_3 - v_1) = \frac{1}{2}(\|v_2 - v_1\|_Q^2 + \|v_3 - v_1\|_Q^2 - \|v_2 - v_3\|_Q^2)$, where $\|v\|_Q \triangleq \sqrt{v^T Q v}$.*

Lemma A.3. *Suppose Assumptions 1 and 3(i) hold. Furthermore, $\gamma_k = 1/(2L)$ for all k. If $h(x_k) \triangleq 2L(x_k - y_{k+1})$, $F(x) - \frac{\mu}{4}\|x - x_k\|^2 \ge F(y_{k+1}) + \frac{1}{4L}\|h(x_k)\|^2 + h(x_k)^T(x - x_k) - \left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k,N_k}\|^2$.*

Proof. Because $y_{k+1} \triangleq \operatorname{argmin}_x \ \frac{1}{2L} g(x) + \frac{1}{2}\|x - [x_k - \frac{1}{2L}(\nabla_x f(x_k) + \bar{w}_{k,N_k})]\|^2$, we have that

$$\begin{aligned} y_{k+1} &= \arg\min_x \ \frac{1}{2L} g(x) + \frac{1}{2}\left[\|x - x_k\|^2 + \frac{1}{L}(x - x_k)^T(\nabla_x f(x_k) + \bar{w}_{k,N_k}) + \frac{1}{4L^2}\|\nabla_x f(x_k) + \bar{w}_{k,N_k}\|^2\right] \\ &= \arg\min_x \ g(x) + \left[L\|x - x_k\|^2 + f(x_k) + (x - x_k)^T(\nabla_x f(x_k) + \bar{w}_{k,N_k})\right]. \end{aligned}$$

Let $\psi_k(x) \triangleq f(x_k) + \nabla_x f(x_k)^T(x - x_k) + L\|x - x_k\|^2 + \bar{w}_{k,N_k}^T(x - x_k)$, implying that

$$y_{k+1} = \arg\min_x \ \psi_k(x) + g(x). \tag{A.1}$$

Then, $\nabla_x \psi_k(x)$ may be expressed as $\nabla_x \psi_k(x) = \nabla_x f(x_k) + 2L(x - x_k) + \bar{w}_{k,N_k}$. By the optimality condition of (A.1), we have $0 \in \partial g(y_{k+1}) + \nabla \psi_k(y_{k+1})$. Hence, by convexity of function $g(x)$, we obtain

$$g(x) \ge g(y_{k+1}) - \nabla \psi_k(y_{k+1})^T(x - y_{k+1}) \Rightarrow \nabla \psi_k(y_{k+1})^T(x - y_{k+1}) \ge g(y_{k+1}) - g(x). \tag{A.2}$$

Consequently, by using the definition of $\psi_k(x)$ and $h(x)$, we have that

$$\nabla_x f(x_k)^T(x - y_{k+1}) \ge g(y_{k+1}) - g(x) + (h(x_k) - \bar{w}_{k,N_k})^T(x - y_{k+1}), \quad \forall x. \tag{A.3}$$

Because f is a μ-strongly convex function,

$$\begin{aligned}
f(x) - \frac{\mu}{2}\|x - x_k\|^2 &\geq f(x_k) + \nabla_x f(x_k)^T(x - x_k) = f(x_k) + \nabla_x f(x_k)^T(x - x_k + y_{k+1} - y_{k+1}) \\
&\overset{\text{(From (44))}}{\geq} f(x_k) + \nabla_x f(x_k)^T(y_{k+1} - x_k) + (h(x_k) - \bar{w}_{k,N_k})^T(x - y_{k+1}) + g(y_{k+1}) - g(x) \\
&= \psi_k(y_{k+1}) - L\|y_{k+1} - x_k\|^2 - \bar{w}_{k,N_k}^T(y_{k+1} - x_k) + (h(x_k) - \bar{w}_{k,N_k})^T(x - y_{k+1}) + g(y_{k+1}) - g(x) \\
&= \psi_k(y_{k+1}) - L\|y_{k+1} - x_k\|^2 + \bar{w}_{k,N_k}^T(x_k - x) + h(x_k)^T(x - y_{k+1}) + g(y_{k+1}) - g(x).
\end{aligned}$$

From the definition of $h(x_k)$, $L\|y_{k+1} - x_k\|^2 = \frac{1}{4L}\|h(x_k)\|^2$ and Inequality (A.2), we have the following:

$$\begin{aligned}
F(x) - \frac{\mu}{2}\|x - x_k\|^2 &\geq \psi_k(y_{k+1}) - \frac{1}{4L}\|h(x_k)\|^2 + h(x_k)^T(x - y_{k+1}) + \bar{w}_{k,N_k}^T(x_k - x) + g(y_{k+1}) \\
&= \psi_k(y_{k+1}) - \frac{1}{4L}\|h(x_k)\|^2 + h(x_k)^T(x - y_{k+1} + x_k - x_k) + \bar{w}_{k,N_k}^T(x_k - x) + g(y_{k+1}) \\
&= \psi_k(y_{k+1}) + \frac{1}{4L}\|h(x_k)\|^2 + h(x_k)^T(x - x_k) + \bar{w}_{k,N_k}^T(x_k - x) + g(y_{k+1}), \qquad \text{(A.4)} \\
&\geq \psi_k(y_{k+1}) + \frac{1}{4L}\|h(x_k)\|^2 + h(x_k)^T(x - x_k) - \frac{1}{\mu}\|\bar{w}_{k,N_k}\| - \frac{\mu}{4}\|x_k - x\|^2 + g(y_{k+1}), \qquad \text{(A.5)}
\end{aligned}$$

where (A.4) follows from the definition of $h(x_k)$ and (A.5) follows by using the fact that $a^Tb \geq -\frac{1}{2\alpha}\|a\|^2 - \frac{\alpha}{2}\|b\|^2$ with $\alpha = 2$. From the L-smoothness of f,

$$\begin{aligned}
\psi_k(y_{k+1}) &= f(x_k) + \nabla_x f(x_k)^T(y_{k+1} - x_k) + L\|x_k - y_{k+1}\|^2 + \bar{w}_{k,N_k}^T(y_{k+1} - x_k) \\
&\geq f(y_{k+1}) + \bar{w}_{k,N_k}^T(y_{k+1} - x_k) + \frac{L}{2}\|x_k - y_{k+1}\|^2 \geq f(y_{k+1}) - \frac{2}{L}\|\bar{w}_{k,N_k}\|^2, \qquad \text{(A.6)}
\end{aligned}$$

where (A.6) follows from $2a^Tb + \|a\|^2 \geq -\|b\|^2$. By substituting (A.6) in (A.5), the result follows. □

It is worth emphasizing that in the proof of Lemma A.3, we employ a simple bound to ensure that the term $\bar{w}_{k,N_k}^T(y_{k+1} - x_k)$ does not appear in the final bound. Instead, the term $\|\bar{w}_{k,N_k}\|^2$ emerges, and this allows for deriving the optimal (rather than suboptimal) oracle complexity. Next, we define a set of parameter sequences that form the basis for updating the iterates.

Definition A.1 (v_k, α_k, τ_k). Given v_0, τ_0, sequences $\{v_k, \tau_k, \alpha_k\}$ are defined as follows:

$$v_{k+1} := \frac{1}{\tau_{k+1}}\left[(1 - \alpha_k)\tau_k v_k + \frac{1}{2}\alpha_k \mu x_k - \alpha_k(h(x_k))\right], \qquad \text{(A.7)}$$

$$\alpha_k \ \text{ solves } \ (1 - \alpha_k)\tau_k + \frac{1}{2}\alpha_k\mu = 2\alpha_k^2 L, \qquad \text{(A.8)}$$

$$\tau_{k+1} := (1 - \alpha_k)\tau_k + \frac{1}{2}\alpha_k\mu. \qquad \text{(A.9)}$$

We employ this set of parameters in showing that the update rule (3) in Algorithm 1 can be recast using the parameters τ_k, α_k, and v_k. This observation is crucial as we analyze the update.

Lemma A.4 (Equivalence of Update Rules). *Suppose Assumptions 1 and 3(i) hold. Suppose the sequences $\{v_k\}, \{\alpha_k\}$, and $\{\tau_k\}$ are prescribed by Definition A.1. Consider the sequence $\{x_k\}$ generated by the algorithm. Then, the following hold:*

i. $\left[x_{k+1} := y_{k+1} + \frac{\alpha_{K+1}\tau_{k+1}(1-\alpha_k)}{\tau_{k+2}+\alpha_{k+1}\tau_{k+1}}(y_{k+1} - y_k)\right] \equiv \left[x_{k+1} := \frac{1}{\tau_{k+1}+\frac{1}{2}\alpha_{k+1}\mu}(\alpha_{k+1}\tau_{k+1}v_{k+1} + \tau_{k+2}y_{k+1})\right]$.

ii. *Suppose $\alpha_k = \frac{1}{\lambda_k}$ for all k. Then, the update rule* (1b) *in* Algorithm 1 *with $\sigma_k \triangleq \frac{(\lambda_k - 1)\left(1 - \frac{\lambda_{k+1}}{4\kappa}\right)}{\left(1 - \frac{1}{4\kappa}\right)\lambda_{k+1}}$ for all k is equivalent to the following:*

$$[x_{k+1} := y_{k+1} + \sigma_k(y_{k+1} - y_k)] \equiv \left[x_{k+1} := \frac{1}{\tau_{k+1} + \frac{1}{2}\alpha_{k+1}\mu}(\alpha_{k+1}\tau_{k+1}v_{k+1} + \tau_{k+2}y_k)\right].$$

Proof.

i. The update rule on the right in (i) can be recast as follows:

$$x_k = \frac{1}{\tau_k + \alpha_k\mu}(\alpha_k\tau_k v_k + \tau_{k+1}y_k) \Longleftrightarrow v_k = \frac{\left(\tau_k + \frac{1}{2}\alpha_k\mu\right)x_k - \tau_{k+1}y_k}{\alpha_k\tau_k}. \qquad \text{(A.10)}$$

Now, by substituting the expression for v_k from (A.10) in (A.7) and recalling that $\tau_{k+1} = (1-\alpha_k)\tau_k + \frac{1}{2}\alpha_k\mu = 2L\alpha_k^2$ and $h(x_k) = 2L(x_k - y_{k+1})$, we obtain the following sequence of equalities:

$$
\begin{aligned}
v_{k+1} &= \frac{1}{\tau_{k+1}}\left[(1-\alpha_k)\tau_k v_k + \frac{1}{2}\alpha_k\mu x_k - \alpha_k(h(x_k)\right] \\
&= \frac{1}{\tau_{k+1}}\left[(1-\alpha_k)\tau_k \frac{\left(\tau_k + \frac{1}{2}\alpha_k\mu\right)x_k - \tau_{k+1}y_k}{\alpha_k\tau_k} + \frac{1}{2}\alpha_k\mu x_k - \alpha_k(h(x_k)\right] \\
&= \frac{(1-\alpha_k)\tau_k + \frac{1}{2}\alpha_k\mu - \frac{1}{2}\alpha_k^2\mu}{\tau_{k+1}\alpha_k}x_k - \frac{1-\alpha_k}{\alpha_k}y_k + \frac{\alpha_k\mu}{2\tau_{k+1}}x_k - \frac{\alpha_k}{\tau_{k+1}}(h(x_k)) \\
&= \frac{\tau_{k+1} - \frac{1}{2}\alpha_k^2\mu}{\tau_{k+1}\alpha_k}x_k - \frac{1-\alpha_k}{\alpha_k}y_k + \frac{\alpha_k\mu}{2\tau_{k+1}}x_k - \frac{\alpha_k}{\tau_{k+1}}h(x_k) \\
&= y_k + \frac{1}{\alpha_k}(x_k - y_k) - \frac{\alpha_k}{2L\alpha_k^2}(2L(x_k - y_{k+1})) = y_k + \frac{1}{\alpha_k}(y_{k+1} - y_k).
\end{aligned}
\tag{A.11}
$$

We now show that the update rule for x_{k+1} on the left is equivalent to that on the right in (i).

$$
\begin{aligned}
x_{k+1} &= \frac{1}{\tau_{k+1} + \frac{1}{2}\alpha_{k+1}\mu}(\alpha_{k+1}\tau_{k+1}v_{k+1} + \tau_{k+2}y_{k+1}) \\
&\overset{(52)}{=} \frac{1}{\tau_{k+1} + \frac{1}{2}\alpha_{k+1}\mu}\left(\alpha_{k+1}\tau_{k+1}y_k + \frac{\alpha_{k+1}\tau_{k+1}}{\alpha_k}(y_{k+1} - y_k) + \tau_{k+2}y_{k+1}\right) \\
&= \left(\frac{\tau_{k+2} + \alpha_{k+1}\tau_{k+1}}{\tau_{k+1} + \frac{1}{2}\alpha_{k+1}\mu)}\right)y_{k+1} + \left(\frac{1}{\alpha_k} - 1\right)\left(\frac{\alpha_{k+1}\tau_{k+1}}{\tau_{k+1} + \frac{1}{2}\alpha_{k+1}\mu}\right)(y_{k+1} - y_k) \\
&= y_{k+1} + \left(\frac{1}{\alpha_k} - 1\right)\left(\frac{\alpha_{k+1}\tau_{k+1}}{\tau_{k+1} + \frac{1}{2}\alpha_{k+1}\mu}\right)(y_{k+1} - y_k) \\
&= y_{k+1} + \frac{\alpha_{k+1}\tau_{k+1}(1-\alpha_k)}{\alpha_k(\tau_{k+1} + \frac{1}{2}\alpha_{k+1}\mu)}(y_{k+1} - y_k) = y_{k+1} + \frac{\alpha_{k+1}\tau_{k+1}(1-\alpha_k)}{\alpha_k(\tau_{k+2} + \alpha_{k+1}\tau_{k+1})}(y_{k+1} - y_k),
\end{aligned}
$$

because $\tau_{k+1} = (1-\alpha_k)\tau_k + \frac{1}{2}\alpha_k\mu$.

ii. By choosing $\tau_{k+1} = 2\alpha_k^2 L$ for $k \geq 0$, satisfying (A.8) and (A.9),

$$
\begin{aligned}
x_{k+1} &= y_{k+1} + \frac{\alpha_{k+1}\tau_{k+1}(1-\alpha_k)}{\alpha_k(\tau_{k+2} + \alpha_{k+1}\tau_{k+1})}(y_{k+1} - y_k) = y_{k+1} + \frac{\alpha_{k+1}\alpha_k(1-\alpha_k)}{\alpha_{k+1}^2 + \alpha_{k+1}\alpha_k^2}(y_{k+1} - y_k) \\
&= y_{k+1} + \frac{\alpha_k(1-\alpha_k)}{\alpha_{k+1} + \alpha_k^2}(y_{k+1} - y_k).
\end{aligned}
\tag{A.12}
$$

Now, by choosing $\alpha_k = \frac{1}{\lambda_k}$, we have the following:

$$
\frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}} = \frac{\frac{1}{\lambda_k}\left(1 - \frac{1}{\lambda_k}\right)}{\left(\frac{1}{\lambda_k}\right)^2 + \frac{1}{\lambda_{k+1}}} = \frac{\lambda_{k+1}(\lambda_k - 1)}{\lambda_{k+1} + \lambda_k^2}.
\tag{A.13}
$$

From the update rule for λ_k, we can obtain

$$
\lambda_{k+1} = \frac{1 - \frac{\lambda_k^2}{4\kappa} + \sqrt{\left(1 - \frac{\lambda_k^2}{4\kappa}\right)^2 + 4\lambda_k^2}}{2} \Rightarrow \lambda_k^2 = \frac{\lambda_{k+1}(\lambda_{k+1} - 1)}{1 - \frac{\lambda_{k+1}}{4\kappa}}.
\tag{A.14}
$$

By substituting (A.14) in (A.13), we obtain $\frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}} = \frac{(\lambda_k-1)(1-\frac{\lambda_{k+1}}{4\kappa})}{(1-\frac{1}{4\kappa})\lambda_{k+1}}$. Hence, (A.12) can be written as

$$x_{k+1} = y_{k+1} + \sigma_k(y_{k+1} - y_k), \quad \sigma_k = \frac{(\lambda_k - 1)\left(1 - \frac{\lambda_{k+1}}{4\kappa}\right)}{\left(1 - \frac{1}{4\kappa}\right)\lambda_{k+1}}. \quad \square$$

We now utilize the previous lemma in defining an auxiliary function sequence $\{\phi_{k+1}(x)\}$ and a sequence $\{p_k\}$. These sequences form the basis for carrying out the final rate analysis.

Lemma A.5. *Suppose Assumptions 1 and 3(i) hold. Consider the iterates generated by Algorithm 1, where $\gamma_k = 1/(2L)$, whereas $\{v_k\}, \{\tau_k\}$, and $\{\alpha_k\}$ are defined in (A.7)–(A.9). Suppose $\phi_1(x) \triangleq F(x_0) + \frac{\tau_1}{2}\|x - x_0\|^2$ and $p_1 = 0$. If $\phi_k(x)$ and p_k are defined as follows for $k \geq 1$,*

$$\phi_{k+1}(x) := (1-\alpha_k)\phi_k(x) + \alpha_k\left[F(y_{k+1}) + \frac{1}{4L}\|h(x_k)\|^2 + \frac{\mu}{4}\|x - x_k\|^2 + h(x_k)^T(x - x_k)\right] \tag{A.15}$$

$$p_{k+1} := (1-\alpha_k)\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k,N_k}\|^2 + (1-\alpha_k)p_k, \tag{A.16}$$

where $h(x_k) = 2L(x_k - y_{k+1})$. If $\phi_k^* \triangleq \min_x \phi_k(x)$, then $\phi_k^* \geq F(y_k) - p_k$, for all $k \geq 1$.

Proof. We begin by showing that $\nabla^2\phi_k(x) = \tau_k I$, where I denotes the identity matrix. For $k = 1$, $\nabla^2\phi_1(x) = \tau_1 I$. Suppose this holds for k, and we proceed to show that this holds for $k := k+1$:

$$\nabla^2\phi_{k+1}(x) = (1-\alpha_k)\nabla^2\phi_k(x) + \frac{1}{2}\alpha_k\mu I = (1-\alpha_k)\tau_k I + \frac{1}{2}\alpha_k\mu I. \tag{A.17}$$

By choosing $\tau_{k+1} = (1-\alpha_k)\tau_k + \frac{1}{2}\alpha_k\mu$, the required claim follows. Next, we show that the sequence $\phi_k(x)$ can be written as follows:

$$\phi_k(x) = \phi_k^* + \frac{\tau_k}{2}\|x - v_k\|^2, \tag{A.18}$$

where $\phi_k^* = \min_x \phi_k(x)$ and $v_k = \operatorname{argmin}_x \phi_k(x)$. Because $\phi_{k+1}(x)$ is a convex quadratic function by definition, we may represent it as $\phi_{k+1}(x) = a + b^T x + \frac{1}{2}x^T Q x$. First, we note that $\nabla^2\phi_{k+1}(x) = Q = \tau_{k+1}I$. By noting that $\nabla_x\phi_{k+1}(v_{k+1}) = 0$, implying that $b + \tau_{k+1}v_{k+1} = 0 \Rightarrow b = -\tau_{k+1}v_{k+1}$. Consequently, we have that $\phi_{k+1}(v_{k+1}) = \phi_{k+1}^* = a - \tau_{k+1}v_{k+1}^T v_{k+1} + \frac{1}{2}\tau_{k+1}\|v_{k+1}\|^2 \Rightarrow a = \phi_{k+1}^* + \frac{\tau_{k+1}}{2}\|v_{k+1}\|^2$. This implies that $\phi_{k+1}(x) = \phi_{k+1}^* + \frac{\tau_{k+1}}{2}\|x - v_{k+1}\|^2$ and (A.18) is shown to be true for all k. Next, we proceed to obtain the recursive rule for v_{k+1} and ϕ_{k+1}^*. By using the optimality conditions for the unconstrained strongly convex problem $\min_x \phi_k(x)$, we obtain the following:

$$\begin{aligned}
0 &= \nabla_x\phi_{k+1}(x) = (1-\alpha_k)\nabla_x\phi_k(x) + \alpha_k\left[\frac{1}{2}\mu(x - x_k) + h(x_k)\right] \\
&\overset{(59)}{=} (1-\alpha_k)\tau_k(x - v_k) + \alpha_k\left[\frac{1}{2}\mu(x - x_k) + h(x_k)\right] \\
&\Rightarrow \nabla_x\phi_{k+1}(x) = \tau_{k+1}(x - v_{k+1}) \text{ implying } v_{k+1} = \frac{1}{\tau_{k+1}}\left[(1-\alpha_k)\tau_k v_k + \frac{1}{2}\alpha_k\mu x_k - \alpha_k h(x_k)\right].
\end{aligned} \tag{A.19}$$

By using Equations (A.15) and (A.18), we obtain the following:

$$\begin{aligned}
\phi_{k+1}^* &= \phi_{k+1}(x_k) - \frac{\tau_{k+1}}{2}\|x_k - v_{k+1}\|^2 \\
&= (1-\alpha_k)\left[\phi_k^* + \frac{\tau_k}{2}\|x_k - v_k\|^2\right] + \alpha_k\left[F(y_{k+1}) + \frac{1}{4L}\|h(x_k)\|^2\right] - \frac{\tau_{k+1}}{2}\|x_k - v_{k+1}\|^2 \\
&= (1-\alpha_k)\left[\phi_k^* + \frac{\tau_k}{2}\|x_k - v_k\|^2\right] + \alpha_k\left[F(y_{k+1}) + \frac{1}{4L}\|h(x_k)\|^2\right] \\
&\quad - \frac{\tau_{k+1}}{2}\left\|x_k - \frac{1}{\tau_{k+1}}\left[(1-\alpha_k)\tau_k v_k + \frac{1}{2}\alpha_k\mu x_k - \alpha_k h(x_k)\right]\right\|^2 \\
&= (1-\alpha_k)\,\phi_k^* + \alpha_k F(y_{k+1}) + (1-\alpha_k)\frac{\tau_k}{2}\|x_k - v_k\|^2 + \alpha_k\left[\frac{1}{4L}\|h(x_k)\|^2\right] \\
&\quad - \frac{\tau_{k+1}}{2}\left\|x_k - \frac{1}{\tau_{k+1}}\left[(1-\alpha_k)\tau_k(v_k - x_k + x_k) + \frac{1}{2}\alpha_k\mu x_k - \alpha_k h(x_k)\right]\right\|^2.
\end{aligned}$$

The expression on the right can be further simplified as follows:

$$\phi_{k+1}^* = (1-\alpha_k)\,\phi_k^* + \alpha_k F(y_{k+1}) + (1-\alpha_k)\,\frac{\tau_k}{2}\|x_k - v_k\|^2 + \alpha_k\left[\frac{1}{4L}\|h(x_k)\|^2\right]$$

$$-\frac{\tau_{k+1}}{2}\left\|\frac{1}{\tau_{k+1}}[-(1-\alpha_k)\tau_k(v_k - x_k) + \alpha_k h(x_k)]\right\|^2$$

$$= (1-\alpha_k)\,\phi_k^* + \alpha_k F(y_{k+1}) + (1-\alpha_k)\,\frac{\tau_k}{2}\|x_k - v_k\|^2 + \alpha_k\left[\frac{1}{4L}\|h(x_k)\|^2\right] - \frac{(1-\alpha_k)^2\tau_k^2}{2\tau_{k+1}}\|v_k - x_k\|^2$$

$$-\frac{\alpha_k^2}{2\tau_{k+1}}\|h(x_k)\|^2 + \frac{(1-\alpha_k)\alpha_k\tau_k}{\tau_{k+1}}h(x_k)^T(v_k - x_k)$$

$$= (1-\alpha_k)\,\phi_k^* + \alpha_k F(y_{k+1}) + (1-\alpha_k)\,\frac{\tau_k}{2}\|x_k - v_k\|^2 + \left(\frac{\alpha_k}{4L} - \frac{\alpha_k^2}{2\tau_{k+1}}\right)\|h(x_k)\|^2$$

$$-\frac{(1-\alpha_k)^2\tau_k^2}{2\tau_{k+1}}\|v_k - x_k\|^2 + \frac{(1-\alpha_k)\alpha_k\tau_k}{\tau_{k+1}}h(x_k)^T(v_k - x_k)$$

$$\Rightarrow \phi_{k+1}^* = (1-\alpha_k)\,\phi_k^* + \alpha_k F(y_{k+1}) + (1-\alpha_k)\,\frac{\tau_k}{2}\left(1 - \frac{(1-\alpha_k)\tau_k)}{\tau_{k+1}}\right)\|x_k - v_k\|^2$$

$$+\left(\frac{\alpha_k}{4L} - \frac{\alpha_k^2}{2\tau_{k+1}}\right)\|h(x_k)\|^2 + \frac{(1-\alpha_k)\alpha_k\tau_k}{\tau_{k+1}}h(x_k)^T(v_k - x_k)$$

$$= (1-\alpha_k)\,\phi_k^* + \alpha_k F(y_{k+1}) + \frac{(1-\alpha_k)\alpha_k\tau_k(\mu/2)}{2\tau_{k+1}}\|x_k - v_k\|^2 + \left(\frac{\alpha_k}{4L} - \frac{\alpha_k^2}{2\tau_{k+1}}\right)\|h(x_k)\|^2$$

$$+\frac{(1-\alpha_k)\alpha_k\tau_k}{\tau_{k+1}}h(x_k)^T(v_k - x_k)$$

$$= (1-\alpha_k)\,\phi_k^* + \alpha_k F(y_{k+1}) + \frac{(1-\alpha_k)\alpha_k}{\tau_{k+1}}\tau_k\left(\frac{\mu}{4}\|x_k - v_k\|^2 + h(x_k)^T(v_k - x_k)\right) + \left(\frac{\alpha_k}{4L} - \frac{\alpha_k^2}{2\tau_{k+1}}\right)\|h(x_k)\|^2.$$

Next, we inductively prove that $\phi_k^* \ge F(y_k) - p_k$, where p_k is defined in (A.16). This holds for $k = 1$, where $p_1 = 0$. Assuming, it is true for k, we prove it holds for $k+1$ by invoking Lemma A.3 for $x = y_k$:

$$\phi_{k+1}^* \ge (1-\alpha_k)(F(y_k) - p_k) + \alpha_k F(y_{k+1}) + \left(\frac{\alpha_k}{4L} - \frac{\alpha_k^2}{2\tau_{k+1}}\right)\|h(x_k)\|^2$$

$$+\frac{\alpha_k(1-\alpha_k)\tau_k}{\tau_{k+1}}\left(\frac{\mu}{4}\|x_k - v_k\|^2 + h(x_k)^T(v_k - x_k)\right) \qquad (\text{Since } \phi_k^* \ge F(y_k) - p_k)$$

$$\ge (1-\alpha_k)\left(F(y_{k+1}) + h(x_k)^T(y_k - x_k) + \frac{1}{4L}\|h(x_k)\|^2 + \frac{\mu}{4}\|y_k - x_k\|^2\right.$$

$$\left.-\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k,N_k}\|^2\right) - (1-\alpha_k)p_k + \alpha_k F(y_{k+1}) + \left(\frac{\alpha_k}{4L} - \frac{\alpha_k^2}{2\tau_{k+1}}\right)\|h(x_k)\|^2 + \frac{\alpha_k(1-\alpha_k)\tau_k}{\tau_{k+1}}$$

$$\times\left(\frac{\mu}{4}\|x_k - v_k\|^2 + h(x_k)^T(v_k - x_k)\right)$$

$$= F(y_{k+1}) + \left(\frac{1}{4L} - \frac{\alpha_k^2}{2\tau_{k+1}}\right)\|h(x_k)\|^2 + (1-\alpha_k)h(x_k)^T\left(\frac{\alpha_k\tau_k}{\tau_{k+1}}(v_k - x_k) + (y_k - x_k)\right)$$

$$-(1-\alpha_k)p_k - (1-\alpha_k)\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k,N_k}\|^2) + (1-\alpha_k)\frac{\mu}{4}\|y_k - x_k\|^2 + \frac{\alpha_k(1-\alpha_k)\tau_k}{\tau_{k+1}}\frac{\mu}{4}\|x_k - v_k\|^2$$

$$\ge F(y_{k+1}) + (1-\alpha_k)h(x_k)^T\overbrace{\left(\frac{\alpha_k\tau_k}{\tau_{k+1}}(v_k - x_k) + (y_k - x_k)\right)}^{\text{Term (a)}} + \overbrace{\left(\frac{1}{4L} - \frac{\alpha_k^2}{2\tau_{k+1}}\right)}^{\text{Term (b)}}\|h(x_k)\|^2$$

$$-(1-\alpha_k)\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k,N_k}\|^2 - (1-\alpha_k)p_k = F(y_{k+1}) - (1-\alpha_k)\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k,N_k}\|^2 - (1-\alpha_k)p_k,$$

where the last inequality follows noting that terms (a) and (b) are zero from recalling that $2L\alpha_k^2 = \tau_{k+1}$ and $x_k = \frac{1}{\tau_k + \frac{1}{2}\alpha_k\mu}(\alpha_k\tau_k v_k + \tau_{k+1}y_k)$ (by Lemma A.4). By choosing $p_{k+1} = (1-\alpha_k)\left(\frac{2}{L}+\frac{1}{\mu}\right)\|\bar{w}_{k,N_k}\|^2 + (1-\alpha_k)p_k$, we have that[4] $\phi^*_{k+1} \geq F(y_{k+1}) - \overbrace{p_{k+1}}^{\text{Term (c)}}$. □

Before analyzing the rate of convergence, we proceed to examine the limiting behavior of the sequence $\{\lambda_k\}$ and show that $\lambda_k \to \sqrt{\kappa}$, where κ denotes the condition number of the problem.

Lemma A.6 (Properties of $\{\lambda_k\}$). *Suppose sequence $\{\lambda_k\}_{k\geq 1}$ is defined by the recursion*

$$\lambda_{k+1} := \frac{1 - \frac{\lambda_k^2}{4\kappa} + \sqrt{\left(1 - \frac{\lambda_k^2}{4\kappa}\right)^2 + 4\lambda_k^2}}{2}, \tag{A.20}$$

where $\lambda_1 \in (1, 2\sqrt{\kappa}]$. Then, $\{\lambda_k\}$ is an increasing and bounded sequence such that $\lim_{k\to\infty}\lambda_k = 2\sqrt{\kappa}$.

Proof. First, by induction, we show that sequence $\{\lambda_k\}$ is bounded above by $2\sqrt{\kappa}$. By assumption, $\lambda_1 \leq 2\sqrt{\kappa}$, we assume $\lambda_k \leq 2\sqrt{\kappa}$ and proceed to show that $\lambda_{k+1} \leq 2\sqrt{\kappa}$:

$$\lambda_{k+1} = \frac{1 - \frac{\lambda_k^2}{4\kappa} + \sqrt{\left(1 - \frac{\lambda_k^2}{4\kappa}\right)^2 + 4\lambda_k^2}}{2} \Leftrightarrow \lambda_k^2 = \frac{\lambda_{k+1}(\lambda_{k+1}-1)}{1 - \frac{\lambda_{k+1}}{4\kappa}}$$

$$\Rightarrow \lambda_k \leq 2\sqrt{\kappa} \Leftrightarrow \frac{\lambda_{k+1}(\lambda_{k+1}-1)}{1 - \frac{\lambda_{k+1}}{4\kappa}} \leq 4\kappa \Leftrightarrow \lambda_{k+1}^2 \leq 4\kappa \Leftrightarrow \lambda_{k+1} \leq 2\sqrt{\kappa}.$$

Because the sequence is increasing and bounded above, its limit exists. Suppose $\lim_{k\to\infty}\lambda_{k+1} = \lambda$, implying $\lambda = \frac{1 - \frac{\lambda^2}{4\kappa} + \sqrt{\left(1 - \frac{\lambda^2}{4\kappa}\right)^2 + 4\lambda^2}}{2} \Rightarrow \lambda = 2\sqrt{\kappa}$. Second, we show that sequence $\{\lambda_k\}$ is increasing, that is, $\lambda_{k+1} \geq \lambda_k$, which can be written equivalently by replacing the recursive rule λ_{k+1} as follows

$$\frac{1 - \frac{\lambda_k^2}{4\kappa} + \sqrt{\left(1 - \frac{\lambda_k^2}{4\kappa}\right)^2 + 4\lambda_k^2}}{2} \geq \lambda_k \Leftrightarrow \left(1 - \frac{\lambda_k^2}{4\kappa}\right)^2 + 4\lambda_k^2 \geq \left(\frac{\lambda_k^2}{4\kappa} - 1 + 2\lambda_k\right)^2 \Leftrightarrow 4\lambda_k\left(1 - \frac{\lambda_k^2}{4\kappa}\right) \leq 0 \Leftrightarrow \lambda_k \leq 2\sqrt{\kappa}. \quad \square$$

We are now in a position to provide our main proposition that provides a bridge toward deriving rate statements and oracle complexity bounds.

Proof of Lemma 1. We have that

$$\mathbb{E}[\phi_{k+1}(x)] \overset{(56)}{=} (1-\alpha_k)\mathbb{E}[\phi_k(x)] + \alpha_k\mathbb{E}\left[F(y_{k+1}) + \frac{1}{4L}\|h(x_k)\|^2 + \frac{\mu}{4}\|x - x_k\|^2 + h(x_k)^T(x - x_k)\right]$$
$$\leq (1-\alpha_k)\mathbb{E}[\phi_k(x)] + \alpha_k\mathbb{E}[F(x)] + \alpha_k\left(\frac{2}{L}+\frac{1}{\mu}\right)\mathbb{E}[\|\bar{w}_{k,N_k}\|^2].$$

By rearranging terms and setting $x = x^*$ in the preceding inequality, we obtain

$$\mathbb{E}[\phi_{k+1}(x^*) - F(x^*)] \leq (1-\alpha_k)\mathbb{E}[\phi_k(x^*) - F(x^*)] + \left(\frac{2}{L}+\frac{1}{\mu}\right)\mathbb{E}\left[\|\bar{w}_{k,N_k}\|^2\right]$$
$$\leq (1-\alpha_k)(1-\alpha_{k-1})\mathbb{E}[\phi_{k-1}(x^*) - F(x^*)] + \alpha_k\left(\frac{2}{L}+\frac{1}{\mu}\right)\mathbb{E}\left[\|\bar{w}_{k,N_k}\|^2\right] + \alpha_k(1-\alpha_{k-1})\left(\frac{2}{L}+\frac{1}{\mu}\right)\mathbb{E}\left[\|\bar{w}_{k-1,N_{k-1}}\|^2\right]\Bigg]$$
$$\leq \left(\prod_{i=1}^{k}(1-\alpha_i)\right)\mathbb{E}[\phi_1(x^*) - F(x^*)] + \alpha_k\sum_{i=0}^{k-1}\left(\prod_{j=0}^{i-1}(1-\alpha_{k-j})\right)\left(\frac{2}{L}+\frac{1}{\mu}\right)\mathbb{E}\left[\|\bar{w}_{k-i,N_{k-i}}\|^2\right].$$

From Lemma A.6, $\alpha_k = \frac{1}{\lambda_k} \in [\bar{\alpha}, 1)$, where $\bar{\alpha} = \frac{1}{2\sqrt{\kappa}}$, and by recalling that $\mathbb{E}[\|\bar{w}_{k-i,N_{k-i}}\|^2 \,|\, \mathcal{H}_{k-i}] \le \nu^2/N_{k-i}$, we obtain the following sequence of inequalities:

$$\begin{aligned}\mathbb{E}[\phi_{k+1}(x^*) - F(x^*)] &\le \left(\prod_{i=1}^{k}(1-\alpha_i)\right)\mathbb{E}[\phi_1(x^*) - F(x^*)] + \sum_{i=0}^{k-1}((1-\bar{\alpha})^i)\left(\frac{2}{L} + \frac{1}{\mu}\right)\mathbb{E}[\mathbb{E}[\|\bar{w}_{k-i,N_{k-i}}\|^2 \,|\, \mathcal{H}_{k-i}]] \\ &\le \left(\prod_{i=1}^{k}(1-\alpha_i)\right)\mathbb{E}[\phi_1(x^*) - F(x^*)] + \sum_{i=0}^{k-1}\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{\nu^2(1-\bar{\alpha})^i}{N_{k-i}}. \end{aligned} \tag{A.21}$$

By using Lemma A.5 and (A.21), we may obtain

$$\begin{aligned} F(y_k) - F(x^*) &\le \mathbb{E}[\phi_k^* + p_k] - F(x^*) \le \mathbb{E}[\phi_k(x^*) - F(x^*)] + \mathbb{E}[p_k] \\ &\le \left(\prod_{i=1}^{k-1}(1-\alpha_i)\right)\mathbb{E}[\phi_1(x^*) - F(x^*)] + \sum_{i=0}^{k-2}\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{\nu^2(1-\bar{\alpha})^i}{N_{k-1-i}} + \mathbb{E}[p_k] \\ &= \left(\prod_{i=1}^{k-1}(1-\alpha_i)\right)\mathbb{E}[F(x_0) - F(x^*) + \frac{\tau_1}{2}\|x^* - x_0\|^2] + \sum_{i=0}^{k-2}\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{\nu^2(1-\bar{\alpha})^i}{N_{k-1-i}} + \mathbb{E}[p_k] \\ &\le (1-\bar{\alpha})^{k-1}\left(D + \frac{\mu}{2}C^2\right) + \sum_{i=0}^{k-2}\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{\nu^2(1-\bar{\alpha})^i}{N_{k-1-i}} + \mathbb{E}[p_k], \end{aligned} \tag{A.22}$$

where we use the fact that $\tau_1 = \mu$ and $\alpha_k \in [\bar{\alpha}, 1)$. Next, we derive a bound on $\mathbb{E}[p_k]$. By definition, we have $p_k = (1-\bar{\alpha})\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k-1,N_{k-1}}\|^2 + (1-\bar{\alpha})p_{k-1}$, implying that

$$\begin{aligned} p_k &= (1-\bar{\alpha})\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k-1,N_{k-1}}\|^2 + (1-\bar{\alpha})^2\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k-2,N_{k-2}}\|^2 + (1-\bar{\alpha})^2 p_{k-2} \\ &= \ldots = \sum_{i=0}^{k-2}(1-\bar{\alpha})^{i+1}\left(\frac{2}{L} + \frac{1}{\mu}\right)\|\bar{w}_{k-i-1,N_{k-i-1}}\|^2. \end{aligned}$$

By taking expectations and invoking Assumptions 1 and 3(i),

$$\mathbb{E}[p_k] \le \sum_{i=0}^{k-2}(1-\bar{\alpha})^{i+1}\left(\frac{2}{L} + \frac{1}{\mu}\right)\mathbb{E}[\mathbb{E}[\|\bar{w}_{k-i-1,N_{k-i-1}}\|^2 \,|\, \mathcal{H}_{k-i-1}]] \le \sum_{i=0}^{k-2}\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{\nu^2(1-\bar{\alpha})^{i+1}}{N_{k-i-1}}. \tag{A.23}$$

By substituting (A.23) in (A.22), we obtain the desired result. □

Proof of Theorem 1.

i. From (3) and by the definition of θ, we may claim the following:

$$\begin{aligned} \mathbb{E}[F(y_K) - F^*] &\le \left(D + \frac{\mu}{2}C^2\right)\theta^{K-1} + \sum_{j=0}^{K-2}\theta^j\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{\nu^2}{N_{K-j-1}} + \sum_{j=0}^{K-2}\theta^{j+1}\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{\nu^2}{N_{K-j-1}} \\ &= \left(D + \frac{\mu}{2}C^2\right)\theta^{K-1} + \left(\frac{2}{L} + \frac{1}{\mu}\right)\theta\sum_{j=0}^{K-2}\theta^j\frac{4\nu^2}{N_{K-j-1}} \le \left(D + \frac{\mu}{2}C^2\right)\theta^{K-1} + \sum_{j=0}^{K-2}\theta^j\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{2\nu^2}{N_{K-j-1}}, \end{aligned} \tag{A.24}$$

Where, in the last inequality, we use the fact that $\bar{\alpha} + 2\theta = 2 - \bar{\alpha} \le 2$. If $N_{K-j-1} = \lfloor \rho^{-(K-j-1)} \rfloor$, by using Lemma A.1, we have the following:

$$\begin{aligned} &\sum_{i=0}^{K-2}\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{2\theta^j\nu^2}{\lfloor \rho^{-(K-j-1)} \rfloor} \le \sum_{i=0}^{K-2}\left(\frac{2}{L} + \frac{1}{\mu}\right)\frac{\theta^i\nu^2}{\rho^{-(K-i-1)}} \le \left(\frac{2}{L} + \frac{1}{\mu}\right)\nu^2\rho^{K-1}\sum_{i=0}^{K-2}\left(\frac{\theta}{\rho}\right)^i \\ &\le \left(\frac{2}{L} + \frac{1}{\mu}\right)\left(\frac{\nu^2\rho}{(\rho-\theta)}\right)\rho^{K-1}. \end{aligned} \tag{A.25}$$

By substituting (A.25) in (A.24), the bound in terms of K is provided next, where $\tilde{C}$ is defined in (4):

$$\mathbb{E}[F(y_K) - F^*] \le \left(D + \frac{\mu}{2}C^2\right)\theta^{K-1} + \left(\frac{2}{L} + \frac{1}{\mu}\right)2\nu^2\sqrt{\kappa}\rho^{K-1} \le \tilde{C}\rho^{K-1}$$

$$\text{where } \tilde{C} = \left(D + \frac{\mu C^2}{2}\right) + \left(\frac{2}{L} + \frac{1}{\mu}\right)2\nu^2\sqrt{\kappa} \le \left(D + \frac{\mu C^2}{2}\right) + \frac{4\nu^2}{\mu} + \frac{2\nu^2\sqrt{\kappa}}{\mu} \qquad . \tag{A.26}$$

Furthermore, we may derive the number of steps K to obtain an ϵ-optimal solution:

$$\frac{1}{\rho} = \frac{1}{\left(1 - \frac{1}{2a\sqrt{\kappa}}\right)} = \frac{2a\sqrt{\kappa}}{(2a\sqrt{\kappa} - 1)} \Rightarrow K \ge \frac{\log(\tilde{C}) - \log(\epsilon)}{\log(1/\rho)} \approx \mathcal{O}(\sqrt{\kappa})\log(\sqrt{\kappa}/\epsilon). \tag{A.27}$$

ii. To compute a vector y_{K+1} satisfying $\mathbb{E}[F(y_{K+1}) - F^*] \le \epsilon$, we have $\tilde{C}\rho^K \le \epsilon$, implying that $K = \lceil \log_{(1/\rho)}(\tilde{C}/\epsilon)\rceil$. To obtain the optimal oracle complexity, we require $\sum_{k=1}^K N_k$ gradients. If $N_k = \lfloor \rho^{-k} \rfloor \le \rho^{-k}$, we obtain the following because $(1-\rho) = (1/(a\sqrt{\kappa}))$.

$$\sum_{k=1}^{K}\rho^{-k} \le \frac{1}{\left(\frac{1}{\rho} - 1\right)}\left(\frac{1}{\rho}\right)^{2+K} \le \frac{1}{\left(\frac{1}{\rho} - 1\right)}\left(\frac{1}{\rho}\right)^{3+\log_{(1/\rho)}(\tilde{C}/\epsilon)} \le \left(\frac{\tilde{C}}{\epsilon}\right)\frac{1}{\rho^2(1-\rho)} = \frac{a\sqrt{\kappa}\tilde{C}}{\rho^2\epsilon}.$$

$$\rho = 1 - \frac{1}{2a\sqrt{\kappa}} \Rightarrow \rho^2 = 1 - 2/(2a\sqrt{\kappa}) + 1/(4a^2\kappa) = \frac{4a^2\kappa - 4a\sqrt{\kappa} + 1}{4a^2\kappa} \ge \frac{4a^2\kappa - 8a\kappa}{4a^2\kappa} = \frac{(a^2 - 2a)\kappa}{a^2\kappa}$$

$$\Rightarrow \frac{\sqrt{\kappa}}{\rho^2} \le \frac{a^2\kappa\sqrt{\kappa}}{(a^2 - 2a)\kappa} = \left(\frac{a}{a-2}\right)\sqrt{\kappa} \Rightarrow \sum_{k=1}^{\log_{(1/\rho)}(\tilde{C}/\epsilon)+1} \rho^{-k} \le \frac{2a^2\sqrt{\kappa}\tilde{C}}{(a-2)\epsilon}$$

$$= \left(\left(D + \frac{\mu C^2}{2}\right) + \frac{4\nu^2}{\mu} + \frac{2\nu^2\sqrt{\kappa}}{\mu}\right)\mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\right). \quad \square$$

Proof of Lemma 3.

i. $\lim_{\eta\to 0}\hat{C}(\eta) = +\infty$ and $\lim_{\eta\to+\infty}\hat{C}(\eta) = +\infty$ because $\lim_{\eta\to 0}\tilde{\kappa}(\eta) = +\infty$ and $\lim_{\eta\to+\infty}\tilde{\kappa}(\eta) = 1$. In other words, $\bar{C}(\eta)$ is a coercive function on the set $\{\eta : \eta \ge 0\}$.

ii. We observe that, for $\eta > 0$,

$$\tilde{\kappa}(\eta) = 1 + \frac{1}{\eta\mu} > 0, \qquad \tilde{\kappa}(\eta)' = -\frac{1}{\eta^2\mu} < 0, \qquad \tilde{\kappa}''(\eta) = \frac{2}{\eta^3\mu} > 0.$$

Furthermore, $Q(\eta) = \max\{\eta^2M^2, 4\Delta^2\}$ and $\bar{\eta} \triangleq \frac{2\Delta}{M}$. Therefore, we have that $Q(\eta)$ is a.e. twice differentiable, and its Clarke generalized gradient and Hessian are defined as follows.

$$\partial_\eta Q(\eta) = \begin{cases} \{2\eta M^2\}, & \eta > \bar{\eta} \\ [0, 2\bar{\eta}M^2], & \eta = \bar{\eta} \\ \{0\}, & \eta < \bar{\eta} \end{cases} \text{ and } \partial^2_\eta\, Q(\eta) = \begin{cases} 2M^2, & \eta > \bar{\eta} \\ \{2\alpha M^2 \mid \alpha \in [0,1]\} & \eta = \bar{\eta}, \\ 0. & \eta < \bar{\eta} \end{cases} \tag{A.28}$$

From Facchinei and Pang (2003, proposition 7.1.9) and by recalling that $\tilde{\kappa}(\eta)$ is continuously differentiable in η, we may define $\partial\hat{C}(\eta)$ as follows.

$$\partial_\eta\hat{C}(\eta) = \partial[2D\eta\tilde{\kappa}] + \partial[8\tilde{\kappa}(\eta)^{5/2}Q(\eta)a] = 2D\eta\tilde{\kappa}' + 2D\tilde{\kappa} + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'Q(\eta)a + 8\tilde{\kappa}^{5/2}a\partial Q(\eta)$$

$$= \begin{cases} \{2D\eta\tilde{\kappa}' + 2D\tilde{\kappa} + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'Q(\eta)a + 8\tilde{\kappa}^{5/2}aQ'(\eta)\}, & \eta > \bar{\eta} \\ \{2D\bar{\eta}\tilde{\kappa}' + 2D\tilde{\kappa} + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'Q(\bar{\eta})a + 8\tilde{\kappa}^{5/2}a(2\alpha\bar{\eta}M^2) \mid \alpha \in [0,1]\}, & \eta = \bar{\eta}\,. \\ \{2D\eta\tilde{\kappa}' + 2D\tilde{\kappa} + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'Q(\eta)a\}, & \eta < \bar{\eta} \end{cases} \tag{A.29}$$

We may then define the Clarke generalized Hessian of $\hat{C}$ as follows.

$$\partial_\eta^2 \hat{C}(\eta) = \begin{cases} \left\{ \begin{array}{l} \{4D\tilde{\kappa}' + 2D\eta\tilde{\kappa}'' + 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2 Q(\eta)a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'' Q(\eta)a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'(2\eta M^2)a \\ \quad + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'(2\eta M^2)a + 8\tilde{\kappa}^{5/2}(2M^2)a\} \end{array} \right\}, & \eta > \bar{\eta} \\ \left\{ \begin{array}{l} \{4D\tilde{\kappa}' + 2D\eta\tilde{\kappa}'' + 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2 Q(\eta)a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'' Q(\eta)a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'(2\alpha\eta M^2)a \\ \quad + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'(2\alpha\bar{\eta}M^2)a + 8\tilde{\kappa}^{5/2}(2\alpha M^2)a \mid \alpha \in [0,1]\} \end{array} \right\}, & \eta = \bar{\eta} \\ \left\{ 4D\tilde{\kappa}' + 2D\eta\tilde{\kappa}'' + 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2 Q(\eta)a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'' Q(\eta)a \right\}. & \eta < \bar{\eta} \end{cases}$$

We now proceed to show that $H > 0$ for all $H \in \partial^2 \hat{C}(\eta)$ and for all $\eta > 0$.

Case 1: $0 < \eta < \bar{\eta}$. In this setting, $Q'(\eta) = Q''(\eta) = 0$. It follows that $\partial^2 \hat{C}(\eta)$ is a singleton given by the scalar H, and it suffices to show that $H > 0$. This follows as shown next.

$$\begin{aligned} H &= 4D\tilde{\kappa}' + 2D\eta\tilde{\kappa}'' + 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2 Q(\eta)a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'' Q(\eta)a \\ &= 2D\left(\frac{2}{\eta^2\mu} - \frac{2}{\eta^2\mu}\right) + \underbrace{30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2 Q(\eta)a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'' Q(\eta)a}_{>0} > 0. \end{aligned}$$

Case 2: $\eta > \bar{\eta}$. Because $Q'(\eta) = 2\eta M^2$ and $Q''(\eta) = 2M^2$ for $\eta > \bar{\eta}$, we have that $\partial^2 \hat{C}(\eta) = \{H\}$, where it suffices to show that $H > 0$. This follows as shown next.

$$\begin{aligned} H &= 4D\tilde{\kappa}' + 2D\eta\tilde{\kappa}'' + 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2 Q(\eta)a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}'' Q(\eta)a + 40\tilde{\kappa}^{3/2}\tilde{\kappa}' Q'(\eta)a + 8\tilde{\kappa}^{5/2} Q''(\eta)a \\ &= 2D\left(\frac{2}{\eta^2\mu} - \frac{2}{\eta^2\mu}\right) + 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2 Q(\eta)a + 8\tilde{\kappa}^{5/2} Q''(\eta)a + \tilde{\kappa}^{3/2}(20\tilde{\kappa}'' Q(\eta) + 40\tilde{\kappa}' Q'(\eta))a \\ &\geq 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2 Q(\eta)a + 8\tilde{\kappa}^{5/2} Q''(\eta)a + \tilde{\kappa}^{3/2}\left(\frac{40\eta^2 M^2}{\eta^3\mu}\right)a - \tilde{\kappa}^{3/2}\left(\frac{80\eta M^2}{\eta^2\mu}\right)a \\ &\geq 30\tilde{\kappa}^{1/2}\frac{M^2}{\eta^2\mu^2}a + 16\tilde{\kappa}^{5/2}M^2 a + \tilde{\kappa}^{1/2}\left(1 + \frac{1}{\eta\mu}\right)\left(\frac{40M^2}{\eta\mu}\right)a - \tilde{\kappa}^{1/2}\left(\frac{80M^2}{\eta^2\mu^2}\right)a. \end{aligned}$$

Here, the first term follows from $Q(\eta) = 2\eta^2 M^2$ and $\tilde{\kappa}' = -\frac{1}{\eta^2\mu}$, and the last term follows from $-\tilde{\kappa}^{3/2}\left(\frac{80\eta M^2}{\eta^2\mu}\right)a \leq -\tilde{\kappa}^{1/2}\left(\frac{80M^2}{\eta^2\mu^2}\right)a$ because $-\tilde{\kappa}^{3/2} = -\tilde{\kappa}^{1/2}\left(1 + \frac{1}{\eta\mu}\right) \leq -\frac{\tilde{\kappa}^{1/2}}{\eta\mu}$.

$$\begin{aligned} &\tilde{\kappa}^{1/2}\left(\frac{30M^2}{\eta^2\mu^2}\right)a + 16\tilde{\kappa}^{5/2}M^2 a + \tilde{\kappa}^{1/2}\left(1 + \left(1 + \frac{1}{\eta\mu}\right)\right)\left(\frac{40M^2}{\eta\mu}\right)a - \tilde{\kappa}^{1/2}\left(\frac{80M^2}{\eta^2\mu^2}\right)a \\ &\geq \tilde{\kappa}^{1/2}\left(\frac{30M^2}{\eta^2\mu^2}\right)a + 16\tilde{\kappa}^{1/2}\left(1 + \frac{2}{\eta\mu} + \frac{1}{\eta^2\mu^2}\right)M^2 a + \tilde{\kappa}^{1/2}\left(\frac{40M^2}{\eta^2\mu^2}\right)a - \tilde{\kappa}^{1/2}\left(\frac{80M^2}{\eta^2\mu^2}\right)a \\ &\geq \tilde{\kappa}^{1/2}\left(\frac{30M^2}{\eta^2\mu^2}\right)a + \tilde{\kappa}^{1/2}\left(\frac{16M^2}{\eta^2\mu^2}\right)a + \tilde{\kappa}^{1/2}\left(\frac{40M^2}{\eta^2\mu^2}\right)a - \tilde{\kappa}^{1/2}\left(\frac{80M^2}{\eta^2\mu^2}\right)a \\ &= \tilde{\kappa}^{1/2}\left(\frac{6M^2}{\eta^2\mu^2}\right)a > 0. \end{aligned}$$

Case 3: $\eta = \bar{\eta}$. Suppose $Q'(\bar{\eta}) \in \partial\hat{C}(\bar{\eta})$ and $H \in \partial^2\hat{C}(\bar{\eta})$, where $Q'(\bar{\eta}) = 2\alpha\bar{\eta}M^2$ and $H = 2\alpha M^2$ and $\alpha \in [0,1]$. It suffices to show that $H > 0$ for $\alpha \in [0,1]$, as we proceed to do next.

$$\begin{aligned}
H &= 4D\tilde{\kappa}' + 2D\eta\tilde{\kappa}'' + 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2Q(\bar{\eta})a + 20\tilde{\kappa}^{3/2}\tilde{\kappa}''Q(\eta)a + 40\tilde{\kappa}^{3/2}\tilde{\kappa}'Q'(\eta)a + 8\tilde{\kappa}^{5/2}Q''(\bar{\eta})a \\
&= 2D\left(\frac{2}{\bar{\eta}^2\mu} - \frac{2}{\bar{\eta}^2\mu}\right) + 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2Q(\eta)a + 8\tilde{\kappa}^{5/2}Q''(\bar{\eta})a + \tilde{\kappa}^{3/2}(20\tilde{\kappa}''Q(\bar{\eta}) + 40\tilde{\kappa}'Q'(\bar{\eta}))a \\
&\geq 30\tilde{\kappa}^{1/2}(\tilde{\kappa}')^2Q(\bar{\eta})a + 8\tilde{\kappa}^{5/2}Q''(\bar{\eta})a + \tilde{\kappa}^{3/2}\left(\frac{40\bar{\eta}^2M^2}{\eta^3\mu}\right)a - \tilde{\kappa}^{3/2}\left(\frac{80\alpha\eta M^2}{\eta^2\mu}\right)a \\
&\geq \tilde{\kappa}^{1/2}\left(\frac{30M^2}{\bar{\eta}^2\mu^2}\right)a + 16\tilde{\kappa}^{5/2}\alpha M^2a + \tilde{\kappa}^{1/2}\left(1 + \frac{1}{\bar{\eta}\mu}\right)\left(\frac{40M^2}{\bar{\eta}\mu}\right)a - \tilde{\kappa}^{1/2}\left(\frac{80\alpha M^2}{\bar{\eta}^2\mu^2}\right)a \\
&\geq \tilde{\kappa}^{1/2}\left(\frac{30M^2}{\bar{\eta}^2\mu^2}\right)a + 16\tilde{\kappa}^{5/2}\alpha M^2a + \tilde{\kappa}^{1/2}\left(\frac{40M^2}{\bar{\eta}^2\mu^2}\right)a - \tilde{\kappa}^{1/2}\left(\frac{80M^2}{\bar{\eta}^2\mu^2}\right)a \\
&\geq \tilde{\kappa}^{1/2}\left(\frac{30M^2}{\eta^2\mu^2}\right)a + 16\tilde{\kappa}^{1/2}\left(1 + \frac{2}{\eta\mu} + \frac{1}{\eta^2\mu^2}\right)\alpha M^2a + \tilde{\kappa}^{1/2}\left(\frac{40M^2}{\bar{\eta}^2\mu^2}\right)a - \tilde{\kappa}^{1/2}\left(\frac{80\alpha M^2}{\bar{\eta}^2\mu^2}\right)a \\
&\geq \tilde{\kappa}^{1/2}\left(\frac{30M^2}{\bar{\eta}^2\mu^2}\right)a + \tilde{\kappa}^{1/2}\left(\frac{16\alpha M^2}{\bar{\eta}^2\mu^2}\right)a + \tilde{\kappa}^{1/2}\left(\frac{40M^2}{\bar{\eta}^2\mu^2}\right)a - \tilde{\kappa}^{1/2}\left(\frac{80\alpha M^2}{\bar{\eta}^2\mu^2}\right)a \\
&= \tilde{\kappa}^{1/2}\left(\frac{30M^2}{\bar{\eta}^2\mu^2}\right)a + \tilde{\kappa}^{1/2}\left(\frac{40M^2}{\bar{\eta}^2\mu^2}\right)a - \tilde{\kappa}^{1/2}\left(\frac{64\alpha M^2}{\bar{\eta}^2\mu^2}\right)a \\
&\overset{\alpha\leq 1}{\geq} \tilde{\kappa}^{1/2}\left(\frac{30M^2}{\bar{\eta}^2\mu^2}\right)a + \tilde{\kappa}^{1/2}\left(\frac{40M^2}{\bar{\eta}^2\mu^2}\right)a - \tilde{\kappa}^{1/2}\left(\frac{64M^2}{\bar{\eta}^2\mu^2}\right)a \\
&= \tilde{\kappa}^{1/2}\left(\frac{6M^2}{\bar{\eta}^2\mu^2}\right)a > 0.
\end{aligned}$$

Consequently, we have that $H > 0$ for $H \in \partial^2\hat{C}(\eta)$ and $\eta > 0$. It follows that $\hat{C}(\eta)$ is strictly convex for $\eta > 0$ (cf. Hiriart-Urruty et al. 1984, example 2.2). Because $\hat{C}(0) = +\infty$, we may then conclude from the definition of convexity that $\hat{C}$ is a strictly convex function on $\{\eta \mid \eta \geq 0\}$.

iii. By part (i), a minimizer of $\hat{C}$ exists in $\{\eta : \eta \geq 0\}$. By part (ii), this minimizer is necessarily unique because $\hat{C}$ is strictly convex. Therefore, $\hat{C}$ has a unique minimizer on $\{\eta \mid \eta \geq 0\}$. □

Proof of Proposition 1. a. Because $\mathbb{E}[\tilde{F}(\bullet,\omega) + \frac{1}{2\eta}\|x_k - \bullet\|^2]$ is $\tilde{\mu}$-strongly convex, where $\tilde{\mu} = \mu + \frac{1}{\eta}$ and x_k is $\mathcal{F}_k$-measurable, we may utilize the proof technique in Shapiro et al. (2009, section 5.9.1) to obtain the following for $j \geq 0$.

$$\begin{aligned}
\mathbb{E}[\|z_{k,j+1} - z_k^*\|^2 \mid \mathcal{F}_k] &\leq (1 - 2\sigma_j\tilde{\mu})\mathbb{E}[\|z_{k,j} - z_k^*\|^2 \mid \mathcal{F}_k] + \gamma_j^2(M_1^2\mathbb{E}[\|z_{k,j}\|^2 \mid \mathcal{F}_k] + M_2^2\|x_k\|^2 + M_3^2) \\
&\overset{(20)}{\leq} (1 - 2\sigma_j\tilde{\mu} + 2\sigma_j^2M_1^2)\mathbb{E}[\|z_{k,j} - z_k^*\|^2 \mid \mathcal{F}_k] \\
&\quad + \sigma_j^2(2M_1^2\mathbb{E}[\|z_k^*\|^2 \mid \mathcal{F}_k] + M_2^2\|x_k\|^2 + M_3^2).
\end{aligned} \tag{A.30}$$

If $e_j \triangleq E[\|z_{k,j} - z_k^*\|^2 \mid \mathcal{F}_k]$ and $d_k \triangleq 2M_1^2\mathbb{E}[\|z_k^*\|^2 \mid \mathcal{F}_k] + M_2^2\|x_k\|^2 + M_3^2$, for any $t_j > 0$, we have that

$$e_{j+1} \leq (1 - 2\sigma_j\tilde{\mu} + 2\sigma_j^2M_1^2)e_j + \sigma_j^2d_k \Rightarrow t_{j+1}e_{j+1} \leq t_{j+1}(1 - 2\sigma_j\tilde{\mu} + 2\sigma_j^2M_1^2)e_j + t_{j+1}\sigma_j^2d_k. \tag{A.31}$$

We intend to show that $t_{j+1}(1 - 2\sigma_j\tilde{\mu} + 2\sigma_j^2M_1^2)e_j \leq t_je_j$. Let $\bar{J}, t_j$, and σ_j be defined as

$$\bar{J} \triangleq \left\lceil \frac{2M_1^2}{\tilde{\mu}^2} - 1 \right\rceil, t_j \triangleq \begin{Bmatrix} \left(1 - \frac{\tilde{\mu}^2}{2M_1^2}\right)^{-j}, & j < \bar{J} \\ j, & j \geq \bar{J} \end{Bmatrix}, \text{ and } \sigma_j \triangleq \begin{Bmatrix} \min\left\{\frac{1}{(j+1)\log(j+1)}, \frac{\tilde{\mu}}{M_1^2}\right\}, & j < \bar{J} \\ \frac{1}{(j+1)\log(j+1)}, & j \geq \bar{J} \end{Bmatrix}. \tag{A.32}$$

For $j \ge \bar J$, we have the following.

$$t_{j+1}(1-2\sigma_j\tilde\mu+2\sigma_j^2M_1^2)\le t_j \Leftrightarrow (1-2\sigma_j\tilde\mu+2\sigma_j^2M_1^2)\le \frac{t_j}{t_{j+1}}$$

$$\Leftrightarrow \left(1-\frac{t_j}{t_{j+1}}-2\sigma_j\tilde\mu+2\sigma_j^2M_1^2\right)\le 0 \Leftrightarrow \sigma_j \le \frac{\tilde\mu+\sqrt{\tilde\mu^2-2M_1^2\left(1-\frac{t_j}{t_{j+1}}\right)}}{2M_1^2}. \tag{A.33}$$

From (A.32), we have that $\frac{t_j}{t_{j+1}}=\left(1-\frac{1}{j+1}\right)$ for $j\ge\bar J$. Consequently,

$$2M_1^2\left(1-\frac{t_j}{t_{j+1}}\right)=\frac{2M_1^2}{j+1}\le\frac{2M_1^2}{\left\lceil\frac{2M_1^2}{\tilde\mu^2}-1\right\rceil+1}\le\tilde\mu^2\Rightarrow \tilde\mu^2-2M_1^2\left(1-\frac{t_j}{t_{j+1}}\right)\ge 0.$$

Using (A.33), we may show that (A.31) is bounded as follows for $j\ge\bar J$:

$$t_{j+1}e_{j+1}\le t_{j+1}(1-2\sigma_j\tilde\mu+2\sigma_j^2M_1^2)e_j+t_{j+1}\sigma_j^2d_k\le t_je_j+t_{j+1}\sigma_j^2d_k\le t_0e_0+\overbrace{\sum_{\ell=0}^{\bar J-1}\sigma_\ell^2t_{\ell+1}d_k}^{\le c_{\bar J}}\,d_k+\sum_{\ell=\bar J}^{j}\sigma_\ell^2t_{\ell+1}d_k$$

$$\le t_0e_0+c_{\bar J}d_k+\sum_{\ell=\bar J}^{j}\frac{\ell}{(\ell+1)^2\log^2(\ell+1)}d_k\le t_0e_0+c_{\bar J}d_k+\sum_{\ell=\bar J}^{j}\frac{1}{(\ell+1)\log(\ell+1)}d_k$$

$$\le t_0e_0+(c_{\bar J}+3)d_k\triangleq t_0e_0+\bar d_k, \tag{A.34}$$

where (A.34) follows from $\sum_{j=1}^{\infty}\frac{1}{(j+1)\log(j+1)}\le 3$. Next, we derive a bound on $e_0=\mathbb{E}[\|z_{k,0}-z_k^*\|^2\mid\mathcal{F}_k]$.

$$\mathbb{E}[\|z_{k,0}-z_k^*\|^2\mid\mathcal{F}_k]=\mathbb{E}[\|x_k-z_k^*\|^2\mid\mathcal{F}_k]\le 2\|x_k-x^*\|^2+2\mathbb{E}[\|x^*-z_k^*\|^2\mid\mathcal{F}_k]$$

$$=2\|x_k-x^*\|^2+2\mathbb{E}[\|\mathrm{prox}_{\eta F}(x^*)-\mathrm{prox}_{\eta F}(x_k)\|^2\mid\mathcal{F}_k]\le 4\|x_k-x^*\|^2,$$

where the last inequality is a result of x_k being $\mathcal{F}_k$-measurable and nonexpansivity of the prox. operator. Similarly, d_k can be bounded as follows.

$$d_k=(2M_1^2\mathbb{E}[\|z_k^*\|^2\mid\mathcal{F}_k]+M_2^2\|x_k\|^2+M_3^2)$$

$$\le 4M_1^2\mathbb{E}[\|z_k^*-x^*\|^2\mid\mathcal{F}_k]+4M_1^2[\|x^*\|^2]+2M_2^2\|x_k-x^*\|^2+2M_2^2\|x^*\|^2+M_3^2$$

$$\le(4M_1^2+2M_2^2)\|x_k-x^*\|^2+(4M_1^2+2M_2^2)\|x^*\|^2+M_3^2,$$

where the last inequality follows from $\|z_k^*-x^*\|=\|\mathrm{prox}_{\eta F}(x_k)-\mathrm{prox}_{\eta F}(x^*)\|\le\|x_k-x^*\|$. Therefore, using (A.34), we may claim that $\mathbb{E}[\|z_{k,j}-z_k^*\|^2\mid\mathcal{F}_k]\le\frac{\hat a^2\|x_k-x^*\|^2+\hat b^2}{j}$, where $\hat a^2=4+4M_1^2+2M_2^2$ and $\hat b^2=(4M_1^2+2M_2^2)\|x^*\|^2+M_3^2$. □

Proof of Theorem 3.

i. By using theorem 3.10 in Bubeck (2015) to bound $\|\bar x_{k+1}-x^*\|^2\le q\|x_k-x^*\|^2$, where $\tilde\kappa=\frac{\eta\mu+1}{\eta\mu}$, $q=1-\frac{1}{\tilde\kappa}=\frac{1}{\eta\mu+1}\in(0,1)$ if $\eta>0$, and $\gamma_k=\eta$, we may obtain the following in which $(1+\delta)<\frac{1}{2q}+\frac{1}{2}$.

$$\mathbb{E}[\|x_{k+1}-x^*\|^2\mid\mathcal{F}_k]\le\left(1+\frac{1}{\delta}\right)\mathbb{E}[\|x_{k+1}-\bar x_{k+1}\|^2\mid\mathcal{F}_k]+(1+\delta)\mathbb{E}[\|\bar x_{k+1}-x^*\|^2\mid\mathcal{F}_k]$$

$$\le\left(1+\frac{1}{\delta}\right)\mathbb{E}[\|x_{k+1}-\bar x_{k+1}\|^2]+(1+\delta)q\mathbb{E}[\|x_k-x^*\|^2]$$

$$=\left(1+\frac{1}{\delta}\right)\mathbb{E}\left[\left\|\frac{\gamma_k}{\eta}(x_k-z_{k,N_k})-\frac{\gamma_k}{\eta}(x_k-z_k^*)\right\|^2\mid\mathcal{F}_k\right]+(1+\delta)q\|x_k-x^*\|^2$$

$$=\left(1+\frac{1}{\delta}\right)\frac{\gamma_k^2}{\eta^2}\mathbb{E}[\|(z_{k,N_k}-z_k^*)\|^2\mid\mathcal{F}_k]+(1+\delta)q\|x_k-x^*\|^2$$

$$=\left(1+\frac{1}{\delta}\right)\mathbb{E}[\|(z_{k,N_k}-z_k^*)\|^2\mid\mathcal{F}_k]+(1+\delta)q\|x_k-x^*\|^2, \tag{A.35}$$

where (A.35) follows from $\gamma_k = \eta$. By Proposition 1, the first term on the right can be bounded as

$$\mathbb{E}[\|(z_{k,N_k} - z_k^*)\|^2 \mid \mathcal{F}_k] \le \frac{\hat{a}^2\|x_k - x^*\|^2 + \hat{b}^2}{N_k},$$

where N_k denotes the number of stochastic subgradient steps taken at major iteration k. Then, by taking unconditional expectations, we have

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \le \left((1+\delta)q + \frac{(1+1/\delta)\hat{a}^2}{N_k}\right)\mathbb{E}[\|x_k - x^*\|^2] + \frac{(1+1/\delta)\hat{b}^2}{N_k}.$$

Let $p_k \triangleq (1+\delta)q + \frac{(1+1/\delta)\hat{a}^2}{N_k}$ and $N_k = \lfloor N_0\rho^{-k}\rfloor$ for $k \ge 0$, where $N_0 > \frac{(1+1/\delta)\hat{a}^2}{1-(1+\delta)q}$. Note that $p_0 < 1$ and $\{p_k\}$ is a decreasing sequence based on the choice of N_0 and $\{N_k\}$. We consider two cases.

a. Let $\rho \ne p_0$ and $\rho \in (0,1)$. In this instance, we obtain the following result.

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \le \mathbb{E}[\|x_0 - x^*\|^2]\prod_{i=0}^{k} p_i + \sum_{i=0}^{k}\left(\frac{(1+1/\delta)\hat{b}^2\prod_{j=0}^{i-1}p_{k-j}}{N_{k-i}}\right)$$

$$\le p_0^{k+1}\mathbb{E}[\|x_0 - x^*\|^2] + \frac{\rho^k(1+1/\delta)\hat{b}^2}{N_0}\sum_{i=0}^{k}\left(\frac{p_0}{\rho}\right)^i \le \mathcal{C}(\max\{\rho, p_0\})^{k+1}, \text{ where } \mathcal{C} \triangleq \left(\mathbb{E}[\|x_0 - x^*\|^2] + \frac{(1+1/\delta)\hat{b}^2/N_0}{1 - \frac{\min\{\rho, p_0\}}{\max\{\rho, p_0\}}}\right).$$

b. Let $\rho = p_0$. Consequently, we obtain the following result.

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \le p_0^{k+1}\mathbb{E}[\|x_0 - x^*\|^2] + \frac{p_0^{k+1}(1+1/\delta)\hat{b}^2}{N_0}(k+1) = ap_0^{k+1} + b(k+1)p_0^{k+1}.$$

It can be shown that there exists $\hat{p}$ such that $p_0 < \hat{p} < 1$. By analyzing $\max_{z\ge 0} z\left(\frac{p_0}{\hat{p}}\right)^z$, we may claim that $kp_0^k < D\hat{p}^k$ for $k \ge 0$ and $\widehat{D} > \frac{1}{\ln(p_0/\hat{p})^e}$. Consequently, for $\hat{p} \in (p_0, 1)$ and $\widehat{D} > \frac{1}{\ln(p_0/\hat{p})^e}$,

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \le \mathcal{C}\hat{p}^{k+1}, \text{ where } \mathcal{C} \triangleq \left(\mathbb{E}[\|x_0 - x^*\|^2] + \frac{(1+1/\delta)\hat{b}^2\widehat{D}}{N_0}\right).$$

ii. Suppose $\rho = p_0$ and $\hat{p} \in (p_0, 1)$ and to compute a vector x_K satisfying $\mathbb{E}[\|x_K - x^*\|^2] \le \epsilon$, we have $\mathcal{C}\hat{p}^K \le \epsilon$, where $\mathcal{C}$ depends on $\hat{p}$. This implies that $K = \lceil \log_{(1/\hat{p})}(\mathcal{C}/\epsilon)\rceil$. From the definition of $\hat{p}, p_0, q$ and by choosing $N_0 = \frac{2(1+1/\delta)\,\hat{a}^2}{1-(1+\delta)q}$, we obtain that

$$\frac{1}{\log(1/\hat{p})} = \frac{\log(1/p_0)}{\log(1/\hat{p})}\frac{1}{\log(1/p_0)} \le \frac{\log(1/p_0)}{\log(1/\hat{p})}\frac{1}{(1-p_0)} = \frac{\log(1/p_0)}{\log(1/\hat{p})}\frac{1}{1 - \left((1+\delta)q + \frac{(1+1/\delta)\hat{a}^2}{N_0}\right)}$$

$$\le \frac{\log(1/p_0)}{\log(1/\hat{p})}\left(\frac{1}{1 - \left((1+\delta)q + \frac{1-(1+\delta)q}{2}\right)}\right) = \frac{\log(1/p_0)}{\log(1/\hat{p})}\left(\frac{1}{\frac{1}{2} - \frac{(1+\delta)q}{2}}\right) \le \frac{\log(1/p_0)}{\log(1/\hat{p})}\left(\frac{1}{\frac{1}{4} - \frac{q}{4}}\right) = \frac{4\log(1/p_0)}{\log(1/\hat{p})}\tilde{\kappa},$$

where the last inequality follows from $\frac{(1+\delta)q}{2} \le \frac{1}{4} + \frac{q}{4}$. Therefore, the iteration complexity is bounded as $\log(\mathcal{C}/\epsilon)/\log(1/p_0) \le \left(\frac{4\log(1/p_0)}{\log(1/\hat{p})}\right)\tilde{\kappa}\log(\mathcal{C}/\epsilon)$. Similarly, if $\rho \ne p_0$, because $\mathcal{C}\max\{\rho, p_0\}^k \le \epsilon$, the iteration complexity is $\mathcal{O}(\tilde{\kappa}\log(\mathcal{C}/\epsilon))$.

iii. Suppose $\rho = p_0$ and $\hat{p} \in (p_0, 1)$. To obtain the oracle complexity, we require $\sum_{k=1}^{K} N_k$ gradients in which $K = \lceil \log_{(1/\hat{p})}(\mathcal{C}/\epsilon) \rceil$.

$$N_0 \sum_{k=1}^{K} \rho^{-k} \leq \frac{N_0}{\left(\frac{1}{\rho} - 1\right)} \left(\frac{1}{\rho}\right)^{2+K} \leq \frac{N_0}{\left(\frac{1}{\rho} - 1\right)} \left(\frac{1}{\rho}\right)^{3+\log_{(1/\hat{p})}(\mathcal{C}/\epsilon)} \leq \frac{N_0}{\rho^2(1-\rho)} \left(\frac{1}{\rho}\right)^{\log_{1/\hat{p}}(\mathcal{C}/\epsilon)}$$

$$= \frac{N_0}{\rho^2(1-\rho)} \left(\frac{1}{\rho}\right)^{\log_{1/\rho}(\mathcal{C}/\epsilon)\log_{1/\hat{p}}(1/\rho)} = \frac{N_0}{\rho^2(1-\rho)} \left(\frac{\mathcal{C}}{\epsilon}\right)^{\log_{1/\hat{p}}(1/\rho)} = \left(\frac{p_0^2}{\rho^2}\right) \frac{N_0}{{p_0}^2(1-\rho)} \left(\frac{\mathcal{C}}{\epsilon}\right)^{\log_{1/\hat{p}}(1/\rho)}$$

$$\leq \left(\frac{p_0^2}{\rho^2}\right) \frac{16(1+1/\delta)\hat{a}^2}{(1-q)^2} \left(\frac{\mathcal{C}}{\epsilon}\right)^{\log_{\hat{p}}} (1/\rho) \qquad .$$

It follows that the oracle complexity is $\mathcal{O}\left(\tilde{\kappa}^3 \left(\frac{\mathcal{C}}{\epsilon}\right)^{\log_{1/\hat{p}}(1/\rho)}\right)$. Similarly, it can be shown that, when $\rho > p_0$ (or $\rho < p_0$), the oracle complexity is $\mathcal{O}\left(\frac{\tilde{\kappa}^3 \mathcal{C}}{\epsilon}\right)$ $\left(\text{or } \mathcal{O}\left(\tilde{\kappa}^3 \left(\frac{\mathcal{C}}{\epsilon}\right)^{\log_{1/p_0}(1/\rho)}\right)\right)$. □

Proof of Lemma 4. Because $\tilde{f}_\eta(x,\omega) \leq \tilde{f}(x,\omega) \leq \tilde{f}_\eta(x,\omega) + \eta\beta(\omega)$ for any x, by taking expectations on both sides and recalling that $\mathbb{E}[\beta(\omega)] \leq \tilde{\beta}$, we have that

$$\mathbb{E}[\tilde{f}_\eta(x,\omega)] \leq \mathbb{E}[\tilde{f}(x,\omega)] \leq \mathbb{E}[\tilde{f}_\eta(x,\omega)] + \eta\mathbb{E}[\beta(\omega)] \qquad \forall x.$$

Suppose f_η is defined as

$$f_\eta(x) \triangleq \mathbb{E}[\tilde{f}_\eta(x,\omega)], \tag{A.26}$$

implying that $f_\eta(x) \leq f(x) \leq f_\eta(x) + \eta\tilde{\beta}$. In addition, because $\|\nabla_x \tilde{f}_\eta(x,\omega) - \nabla_x \tilde{f}_\eta(y,\omega)\| \leq \frac{\alpha(\omega)}{\eta}\|x-y\|$, for all x, y, by taking expectations on both sides and invoking Jensen's inequality, we have that

$$\begin{aligned} \|\nabla_x f_\eta(x) - \nabla_x f_\eta(y)\| &= \|\nabla_x \mathbb{E}[\tilde{f}_\eta(x,\omega)] - \nabla_x \mathbb{E}[\tilde{f}_\eta(y,\omega)]\| \\ (\text{Jensen's inequality}) &\leq \mathbb{E}[\|\nabla_x \tilde{f}_\eta(x,\omega) - \nabla_x \tilde{f}_\eta(y,\omega)\|] \\ \left(\tilde{f}_\eta(\cdot,\omega) \text{ is } \frac{\alpha(\omega)}{\eta} \text{-smooth}\right) &\leq \mathbb{E}\left[\frac{\alpha(\omega)}{\eta}\right]\|x-y\| \\ &\leq \frac{\tilde{\alpha}}{\eta}\|x-y\| \quad \forall x, y, \end{aligned}$$

Where, in the first inequality, we use theorem 7.47 in Shapiro et al. (2009) (interchangeability of the derivative and the expectation). It follows that f_η is $\tilde{\alpha}/\eta$-smooth. We may conclude that $(\tilde{\alpha}, \tilde{\beta})$-smoothability of f follows. □

Endnotes

[1] We thank P. Dvurechensky for alerting us to Tran-Dinh et al. (2018) and Van Nguyen et al. (2017).

[2] The Lambert function $W(x)$ is the inverse function of $ye^y = x$ and is denoted by $y = W(x)$. This function has two real branches: an upper branch $W_0(x)$ for $x \in [-\frac{1}{e}, +\infty]$ and a lower branch $W_{-1}(x)$ for $x \in \left[-\frac{1}{e}, 0\right]$ (Veberic 2010).

[3] While pursuing submission of the present work, we were informed of related work by Jofré and Thompson (2017) through a private communication.

[4] The update rule for x_k, according to Lemma A.4, is equivalent to that in the algorithm. Also, compared with the approach by Nesterov, we employ inexact (rather than exact) gradients; the key difference in the proof is term (c).

Acknowledgments

We would like to thank the associate editor and the three referees for all of their comments and suggestions. These have led to a greatly improved manuscript.

References

Beck A (2017) *First-Order Methods in Optimization* (SIAM, Philadelphia).

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* 2(1):183–202.

Beck A, Teboulle M (2012) Smoothing and first order methods: A unified framework. *SIAM J. Optim.* 22(2):557–580.

Boţ RI, Hendrich C (2013) A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems. *Comput. Optim. Appl.* 54(2):239–262.

Boţ RI, Hendrich C (2015) A variable smoothing algorithm for solving convex optimization problems. *TOP* 23(1):124–150.

Bubeck S (2015) *Convex Optimization: Algorithms and Complexity*, Foundations and Trends in Machine Learning (Now Publishers, Inc., Hanover, MA), 8(3–4):231–357.

Chambolle A, Pock T (2011) A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* 40(1):120–145.

Chatzigeorgiou I (2013) Bounds on the Lambert function and their application to the outage analysis of user cooperation. *IEEE Comm. Lett.* 17(8):1505–1508.

Dang CD, Lan G (2015) Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM J. Optim.* 25(2):856–881.

Devolder O, Glineur F, Nesterov Y (2012) Double smoothing technique for large-scale linearly constrained convex optimization. *SIAM J. Optim.* 22(2):702–727.

Devolder O, Glineur F, Nesterov Y (2014) First-order methods of smooth convex optimization with inexact oracle. *Math. Programming* 146(1–2):37–75.

Dvurechensky P, Gasnikov A (2016) Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *J. Optim. Theory Appl.* 171(1):121–145.

Facchinei F, Pang JS (2003) *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer Series in Operations Research, vol. I (Springer-Verlag, New York).

Ghadimi S, Lan G (2012) Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM J. Optim.* 22(4):1469–1492.

Ghadimi S, Lan G (2013) Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: Shrinking procedures and optimal algorithms. *SIAM J. Optim.* 23(4):2061–2089.

Ghadimi S, Lan G (2016) Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Programming* 156(1–2): 59–99.

Hiriart-Urruty JB, Strodiot JJ, Nguyen VH (1984) Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data. *Appl. Math. Optim.* 11(1):43–56.

Jalilzadeh A, Shanbhag UV (2016) eg-VSSA: An extragradient variable sample-size stochastic approximation scheme: Error analysis and complexity trade-offs. *Winter Simulation Conf.*, 690–701.

Jofré A, Thompson P (2017) On variance reduction for stochastic smooth convex optimization with multiplicative noise. Preprint, submitted May 8, https://arxiv.org/abs/1705.02969.

Kushner HJ, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications*, Applications of Mathematics, vol. 35, 2nd ed. (Springer Science & Business Media, New York).

Lan G (2012) An optimal method for stochastic composite optimization. *Math. Programming* (Springer), 133(1):365–397.

Moreau JJ (1965) Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* 93(2):273–299.

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.

Nesterov Y (1983) A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN USSR* 269: 543–547.

Nesterov Y (2005a) Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optim.* 16(1):235–249.

Nesterov Y (2005b) Smooth minimization of non-smooth functions. *Math. Programming* 103(1):127–152.

Nesterov Y (2014) *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed. (Springer Publishing Company, Inc. Norwell, MA).

Newton D, Pasupathy R, Yousefian F (2018) Recent trends in stochastic gradient descent for machine learning and big data. *Proc. 2018 Winter Simulation Conf.* (IEEE Press), 366–380.

Orabona F, Argyriou A, Srebro N (2012) Prisma: Proximal iterative smoothing algorithm. Preprint, submitted June 11, https://arxiv.org/abs/1206.2372.

Ouyang H, Gray A (2012) Stochastic smoothing for nonsmooth minimizations: Accelerating SGD by exploiting structure. Preprint, submitted May 21, https://arxiv.org/abs/1205.4481.

Planiden C, Wang X (2016) Strongly convex functions, Moreau envelopes, and the generic nature of convex functions with strong minimizers. *SIAM J. Optim.* 26(2):1341–1364.

Polyak BT (1987) *Introduction to Optimization* (Optimization Software, Inc., New York).

Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* 30(4):838–855.

Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

Schmidt M, Roux NL, Bach FR (2011) Convergence rates of inexact proximal-gradient methods for convex optimization. *Adv. Neural Inform. Processes Systems* 24:1458–1466.

Shamir O, Zhang T (2013) Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *Internat. Conf. Machine Learn.*, 71–79.

Shanbhag UV, Blanchet JH (2015) Budget-constrained stochastic approximation. *Proc. 2015 Winter Simulation Conf.*, 368–379.

Shapiro A, Dentcheva D, Ruszczýnski A (2009) *Lectures on Stochastic Programming* (SIAM, Philadelphia).

Tran-Dinh Q (2017) Adaptive smoothing algorithms for nonsmooth composite convex minimization. *Comput. Optim. Appl.* 66(3):425–451.

Tran-Dinh Q, Fercoq O, Cevher V (2018) A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM J. Optim.* 28(1):96–134.

Van Nguyen Q, Fercoq O, Cevher V (2017) Smoothing technique for nonsmooth composite minimization with linear operator. Preprint, submitted June 19, https://arxiv.org/abs/1706.05837.

Veberic D (2010) Having fun with Lambert $W(x)$ function. Preprint, submitted March 8, https://arxiv.org/abs/1003.1628.

Yousefian F, Nedić A, Shanbhag UV (2012) On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica J. IFAC* 48(1):56–67.

Zhong W, Kwok J (2014) Accelerated stochastic gradient method for composite regularization. *Artificial Intelligence Statist.*, 1086–1094.