Scalable Techniques for Stochastic Power Flow Problems

Final Report: DOE ARPA-E

Uday V. Shanbhag*(PI)
Gary and Sheila Bello Chair and Professor
Department of Industrial and Manufacturing Engineering
Pennsylvania State University, University Park, PA, 16802.

December 20, 2021

*This report is based on research carried out with Shisheng Cui (doctoral student and postdoctoral fellow), Afrooz Jalilzadeh (doctoral student), and Wendian Wan (doctoral student).

Contents

1 Introduction

In this report, we review our recent efforts in developing tools and techniques for contending with the next generation of operational problems in power systems when plagued by risk, uncertainty, nonlinearity, discreteness, and in some instances, the presence of equilibrium constraints.

The optimal power flow (OPF) problem is amongst the most fundamental decision-making problems in power systems. There are many variations and generalizations of this problem including unit commitment, reserve scheduling, economic dispatch, security-constrained, DC approximations, and full AC power-flow formulations [1–5]. A key complication arising from the presence of uncertainty, possibly arising from stochasticity in the availability and demand. One approach to address the presence of uncertainty lies in developing robust optimization models [6–10] where uncertainty sets are assumed and feasibility is ensured for every realization from such uncertainty sets. For instance, a robust formulation for AC power flow problems was provided in [11]. Uncertainty in power systems also can be dealt by adding chance constraints [5, 12] which are approximated by finite sampling of uncertain parameters from an assumed statistical model. Simulation based studies on real time dispatch are also been conducted. For instance, in [13], a simulation based framework is used in a power system with renewable resources. Our interest lies in adaptive two-stage models for such problems possibly complicated by the presence of risk measures. Such avenues have been considered in [14] and revisited in [15], where computable closed-form expressions were derived. In [16], a model for risk-limiting dispatch with generation limitation and network constraint was provided with networked variants were examined in [17].

Overlaying of discreteness emerges through the need to commit generation capacity prior to the clearing of the real-time market. The unit commitment problem considers the determination of the optimal production schedule of power generating units, so that in a certain amount of time the operational cost may be minimized while meeting demand requirements and physical constraints. Basically, binary variables represent the status of unit. An overview of unit commitment problem in literature was provided in [18]. In [19, 20], branch and bound schemes were used to solve the unit commitment problem while the Lagrangian relaxation is also widely used in solving the unit commitment problem [21–25].

We draw inspiration from the `Grid Optimization (GO) Competition`[1] where comprehensive formulations for two-stage optimal power flow problem have been provided in Challenges 1 and 2. The framework truly represents the next generation of such problems, capturing an alternating current formulation, incorporating security constraints, pre-contingency and post-contingency operational constraints, preventive and corrective actions, contingency modeling of generators and transmission (line/transformer), amongst others. While there has been a significant amount of activity in resolving this class of problems as part of the competition, our focus is to make inroads into developing algorithmic tools supported by rigorous convergence theory for addressing subclasses of such problems. To this end, we consider the following complexities that have received less attention from the community.

We emphasize that the key distinctions in the developments in Gates 2 and 3 are that Gate 3 tends to emphasize applications to power systems operation and large-scale testing. We also introduce further enhancements and extensions in Gate 3 (such as through the introduction of risk and the addition of structured nonconvexity in the form of equilibrium constraints).

[1] `https://gocompetition.energy.gov/challenges/challenge-1/formulation`,

Gate 2. (I) *Uncertainty and nonlinearity.* When contending with two-stage models for decision-making in settings complicated by nonlinearity (under suitable convexity assumptions), cutting-plane schemes appear to be the de-facto standard. Stochastic approximation schemes have to contend with nonsmoothness in the recourse function, limiting the performance of the schemes in practical regimes. However, by combining Nesterov acceleration, smoothing, and variance-reduction, we demonstrate that smoothed variance-reduced accelerated schemes display optimal rates and far better performance. In Section 2, we present the smoothed accelerated variance-reduced schemes and provide numerical evidence to support the theoretical claims.

Gate 3. (II)*Risk-based extensions to economic dispatch.* In Section 3, the schemes from (I) can be extended to risk-afflicted regimes with a subtle modification to the nature of the subproblems. The techniques in (I) are applied on an IEEE system as well as on ARPA-E systems.

Gate 2. (III) *Uncertainty, nonlinearity, and discreteness.* Significant inroads have been made in developing cutting-plane schemes for mixed-integer generalizations of two-stage stochastic linear programming (with some generalizations). Our focus is on overlaying discreteness in both stages when contending with problems of the form seen in (a). Notably, we consider the development of stochastic branch-and-bound schemes where the continuous relaxations are two-stage stochastic/risk-averse programs. This avenues allows for developing upper and lower bounds and can accommodate the introduction of various types of cuts which improve performance. This framework is developed in Section 4.

Gate 3. (IV) *Complementarity constraints, uncertainty, and discreteness.* Several aspects complicate the resolution of the problems presented in the GO competition. One particular intricacy that proves debilitating comes in the form of complementarity constraints as seen in `https://gocompetition.energy.gov/challenges/challenge-2/formulation`. When overlaid by discreteness, such problems can be viewed as mixed-integer stochastic mathematical programs with complementarity constraints (mi-SMPCCs) have seen little or no research. In Section 5, we propose a framework for computing efficient solutions for a relatively broad subclass of two-stage stochastic MPCCs. This framework can then be combined with the branching framework presented in (III) to resolve mixed-integer variants. Note that in this setting, we test our scheme on problems with over 100,000 variables.

The remainder of this report is organized into five sections. Sections 2 – 5 correspond to (a) – (c) in the discussion above while Sections 6– 7 discuss commercial development and concluding remarks.

2 Variance-reduced accelerated schemes for 2-stage risk-averse optimization

We consider the following stochastic nonsmooth convex optimization problem[2]

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

$f(x) \triangleq \mathbb{E}[\tilde{f}(x,\xi(\omega))]$, $\xi : \Omega \to \mathbb{R}^o$, $\tilde{f} : \mathbb{R}^n \times \mathbb{R}^o \to \mathbb{R}$, 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 $\tilde{f}(x,\omega)$, while $\tilde{F}(x,\omega) \triangleq \tilde{f}(x,\omega) + g(x)$. We consider settings where $\tilde{f}(\cdot,\omega)$ is nonsmooth strongly convex/convex in x for every ω, **generalizing the focus** beyond the *structured nonsmooth* setting where the "stochastic part" is smooth. Specifically, structured nonsmooth problems require minimizing $f(x)+g(x)$ where f is smooth while g is nonsmooth with an efficient prox evaluation (allows for capturing constrained problems over closed and convex sets).

Amongst the earliest avenues for resolving (1) is stochastic approximation [27, 28] and has proven to be effective on a breadth of stochastic computational problems including convex optimization problems. [29] developed an averaging scheme in convex differentiable settings, deriving the optimal convergence rate of $\mathcal{O}(1/\sqrt{k})$ under classical assumptions, where k is the number of iterations. Amongst 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 were given as $\left(\max\left\{\frac{M^2}{\mu^2}, \|x_0 - x^*\|^2\right\}\frac{1}{\epsilon}\right)$ (to ensure that $\mathbb{E}[\|x_k - x^*\|^2] \leq \epsilon$) and $\mathcal{O}(\frac{MD_X}{\epsilon^2})$ (to ensure that the expected optimality gap is less than ϵ), respectively where $S(x,\omega)$ denotes a measurable selection from $\partial_x \tilde{f}(x,\omega)$, $\sup_{x\in X}\mathbb{E}[\|S(x,\omega)\|^2] \leq M^2$ and $D_X \triangleq \max_{x\in X}\|x_0 - x\|$. Of these, the former was presented by [30] while the latter is the result of an optimal robust constant steplength SA scheme suggested by [31]. When f is both L-smooth and μ-strongly convex, an improved complexity requirement (from a constant factor standpoint) of $\mathcal{O}\left(\sqrt{\frac{L\|x_0-x^*\|^2}{\epsilon}} + \frac{\nu^2}{\mu\epsilon}\right)$ was provided by [32]. This contrasts sharply with the deterministic regime where $\mathcal{O}(\log(1/\epsilon))$ and $\mathcal{O}(1/\sqrt{\epsilon})$ steps are required in smooth strongly convex and smooth convex regimes to compute an ϵ-solution in terms of mean-squared error and expected sub-optimality, respectively. In structured nonsmooth regimes, there has been 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_1, our proposed variable sample-size accelerated proximal gradient scheme (**VS-APM**) (also see [33] and [34]) is stated as follows where the true gradient is replaced by a sample average $(\nabla_x f(x_k) + \bar{w}_{k,N_k})$ with batch size N_k.

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

where $\bar{w}_{k,N_k} \triangleq \frac{\sum_{j=1}^{N_k}\left(\nabla_x \tilde{f}(x_k,\omega_{j,k}) - \nabla_x f(x_k)\right)}{N_k}$, $\mathbf{P}_{\eta g}(y) \triangleq \arg\min_x\{\frac{1}{2}\|x-y\|^2 + \frac{1}{2\eta}g(x)\}$, γ_k, and β_k are suitably defined steplengths. Our approach produces linearly convergent iterates in strongly convex regime and achieves iteration complexity of $\mathcal{O}(1/k^2)$ in merely convex and smooth regimes, matching the deterministic results seen in the work by [35] and [36]. The avenue represented

[2]This section has been adapted from [26].

by (2) has two key distinctions: (i) *Increasingly exact gradients* through increasing batch-sizes N_k of sampled gradients, allowing for progressive variance reduction; (ii) *Larger (non-diminishing) 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. [37]).

2.1 Prior research.

(**a**) *Stochastic gradient schemes.* In nonsmooth convex stochastic optimization problems, [31] derived an optimal rate of $\mathcal{O}(1/\sqrt{k})$ via an optimal constant steplength (also see [38]) while in strongly convex regimes, they derived a rate of $\mathcal{O}(1/k)$. Structured nonsmooth problems (or composite problems) as defined by (1)) have been examined extensively (cf. [39], [40]) and rates of $\mathcal{O}(L/k^2+1/\sqrt{k})$ and $\mathcal{O}(L/k + 1/\sqrt{k})$ were developed by [41] via a mirror-descent framework for strongly convex and convex problems with L-smooth objectives, respectively. In related work, [42] derive oracle complexities with a deterministic oracle of fixed inexactness, which was extended to a stochastic oracle by [43]. Randomized smoothing techniques have also been employed by [44] together with recursive steplengths (see [45] for a review).

(**b**) *Variance reduction.* In strongly convex regimes (without acceleration), a linear rate of convergence in expected error was first shown for variance-reduced gradient methods by [46] and revisited by [34], while similar rates were provided for extragradient methods by [47]; the accelerated counterpart (**VS-APM**) improves the dependence on κ, improving the bound to $\mathcal{O}(\sqrt{L/\mu}\log(1/\epsilon))$. In smooth regimes, an accelerated scheme was first presented by [33] where 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. [34] extended this scheme to allow for state-dependent noise. An extragradient-based variable sample-size framework was suggested by [47] with a rate of $\mathcal{O}(1/k)$.
(**c**) *Smoothing techniques for nonsmooth problems.* For a subclass of deterministic nonsmooth problems, [48] proved that an ϵ-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, [49] considered primal-dual smoothing in deterministic regimes (extended to composite problems by [50]) 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 [51], [52]). Adaptive smoothing, considered by [53], was shown to have an iteration complexity of $\mathcal{O}(1/\epsilon)$ while [54] showed 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 a.e. ω while $\tilde{f}(\cdot,\omega)$ is either strongly convex or merely convex for a.e. ω (extended by [55])[3].

2.2 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, since they utilize diminishing steplengths 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.

[3]We would like to thank P. Dvurechensky for alerting us to [50] and [56].

(I) (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 can contend with smooth 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 $F_\eta(x)$, where $F_\eta(x)$ is $\frac{1}{\eta}$-smooth and retains the minimizers of $F(x)$. **(a) 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 in inexact gradient steps $\nabla_x F_\eta(x_k)$ of $\mathcal{O}(\log(1/\epsilon))$, where increasingly exact gradients $\nabla_x F_\eta(x)$ are obtained by employing an (**prox-SSG**) scheme. In particular, our variance-reduced scheme endeavors to get increasingly exact gradients by progressively reducing the bias in the gradients (since 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. **(b) Unbounded domains.** When domains are possibly unbounded, assuming that $\mathbb{E}[\|S(x,\omega)\|^2] \leq \bar{M}^2\|x\|^2 + M^2$, where $S(x,\omega) \in \partial_F(x,\omega)$, the proposed (unaccelerated) variable sample-size proximal method (**mVS-PM**) achieves an iteration complexity of $\mathcal{O}(\log(1/\epsilon))$ (in $\nabla_x F_\eta$) and overall complexity in subgradient steps of $\mathcal{O}(1/\epsilon)$.

(II) (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 steplength 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 [48] which produces approximate solutions via a fixed smoothing parameter) and is characterized by the optimal oracle complexity of $\mathcal{O}(1/\epsilon^2)$. We may specialize these results to obtain an optimal rate of $\mathcal{O}(1/k^2)$ when f is convex and smooth and displays an optimal sample complexity of $\mathcal{O}(1/\epsilon^2)$. When f is deterministic but nonsmooth, (**s-APM**) matches the rate by [48] but produces asymptotically exact solutions. Additionally, we prove that for suitable (but distinct) choices of steplength and smoothing sequences, (**sVS-APM**) and (**VS-APM**) produce sequences that converge a.s. to a solution of (1), a convergence statement that was unavailable thus far, matching deterministic results by [57] and [58] which leverage Moreau smoothing; we provide a result for (α,β)-smoothable functions (see [59]). Additionally, we prove a.s. convergence of iterates which does not follow immediately.

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

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)** Non-diminishing 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)** Non-dminishing outer steps; Approx. $\nabla_x F_\eta(x)$ by **(SSG)** with increasing exactness;

2.3 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 5.2, we review a linearly convergent variance-reduced accelerated proximal scheme (**VS-APM**) for smooth stochastic convex optimization; this scheme will serve as our subproblem solver. In Section 5.2.1, we present a Moreau-smoothed variant of (**VS-APM**), referred to as (**mVS-APM**), which relies on minimizing the Moreau envelope $F_\eta(x)$ of the strongly convex nonsmooth function $F(x)$ by **(VS-APM)**. In Section 5.2.2, we then derive rate and complexity guarantees for (**mVS-APM**) $\nabla_x F_\eta(x)$ is approximated with increasing accuracy by stochastic subgradient (**SSG**) scheme. Finally, in Section 2.3.4, 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.

2.3.1 Background on (VS-APM)

Consider (1) where f, g, and the initial point x_1 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 $\mathbb{E}[\|x_1 - x^*\|^2] \leq C$ and $\mathbb{E}[\|F(x_1) - F(x^*)\|] \leq D$, 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. *Consider f is continuously differentiable function with 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 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 (VS-APM)

(0) Given x_1, $y_1 = x_1$, κ, and positive sequences $\{\gamma_k, N_k\}$; Set $\lambda_1 \in (1, \sqrt{\kappa}\,]$; $k := 1$;

(1) $y_{k+1} := \mathbf{P}_{\gamma_k g}\left(x_k - \gamma_k\left(\nabla_x f(x_k) + \bar{w}_{k,N_k}\right)\right)$;

(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 (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 $\mathbb{E}[\|\bar{w}_k\|^2 \mid \mathcal{H}_k] \leq \frac{\nu^2}{N_k}$ 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 $w_k \triangleq \nabla_x f(x_k) - \nabla_x \tilde{f}(x_k, \omega_k)$.*

(VS-APM) can be shown to achieve linear convergence akin to that by [37] by combining inexact gradients where 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-solution. These differences lead to a slightly modified set of update rules in contrast with that developed by [37] and requires 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 [37] and is in the Appendix. Importantly, this scheme allows for a possibly **biased** estimate of the gradient.

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

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

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

Theorem 2 (**Rate and oracle complexity of (VS-APM) under biased oracles).** *Suppose Assumptions 1, 2, and 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* $k \geq 0$ *and* $a > 2$.

(i) For all K, *we have that* $\mathbb{E}[F(y_K) - F^*] \leq \tilde{C}\rho^{K-1}$ *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*-solution, i.e.* $\mathbb{E}[F(y_{K+1}) - F^*] \leq \epsilon$.
(ii) To compute an ϵ*-solution,* $\sum_{k=1}^{K} N_k \leq \left(\left(D + \frac{\mu C^2}{2}\right) + \frac{4\nu^2}{\mu} + \frac{2\nu^2\sqrt{\kappa}}{\mu}\right)\mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$.

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, [60] impose unbiasedness in strongly convex regimes. Next, we show that by adding the unbiasedness requirement, i.e. $\mathbb{E}[w_k \mid \mathcal{H}_k] = 0$ a.s. for all k, improves the constants in these bounds.

Corollary 3 (**Rate and oracle complexity of (VS-APM) under unbiased oracles).** *Suppose Assumptions 1, 2, and 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* $k \geq 0$ *and* $a > 2$.

(i) For all K, *we have that* $\mathbb{E}[F(y_K) - F^*] \leq \tilde{C}\rho^{K-1}$ *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*-solution.*
(ii) To compute an ϵ*-solution,* $\sum_{k=1}^{K} N_k \leq \left(\left(D + \frac{\mu C^2}{2}\right) + \frac{4\nu^2}{\mu}\right)\mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$.

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**), which can always be implemented if μ, strong convexity modulus, is known. Later, in Section 2.4, we introduce iteratively smoothed VS-APM (sVS-APM) method which does not require the knowledge of Lipschitz constant L and the strong convexity modulus μ.

2.3.2 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)** where convergence relies on choosing square-summable but non-summable steplength sequences. The choice of the parameters in such sequences can have debilitating impact on performance in some settings (cf. [30]). Specifically, while choosing γ_k as $\frac{1}{\mu k}$ minimizes the mean-squared error but over-estimating μ can have catastophic impact as seen in [30, Sec 5.9, Ex. 5.36]. More generally, such choices are often characterized by poor asymptotic behavior, a consequence that arises in part from the diminishing nature of steplength 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. [61]), denoted by $F_\eta(x)$ 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 $F(x)$ when F is strongly convex.

Lemma 4. *[62, 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 $F_\eta(x)$; (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 $F_\eta(x)$, which is not necessarily an easy task since computing $\nabla_x F_\eta(x)$ necessitates solving nonsmooth stochastic optimization problems. We adopt an inexact accelerated proximal scheme for minimizing $F_\eta(x)$. But in contrast with (**SSG**) schemes applied to minimizing $F(x)$, we control the smoothness of the outer problem by choosing η and utilize **(i) larger non-diminishing steplengths**, **(ii) acceleration**, and **(iii) increasingly exact gradients**, all of which are distinct from (**SSG**), as shown next.

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

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 $F_\eta(x)$ is defined as $\nabla_x F_\eta(x) = \frac{1}{\eta}(x - \text{prox}_{\eta F}(x))$, where

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

But $\text{prox}_{\eta F}(x)$ cannot be computed in finite time since F is a nonsmooth expectation-valued convex function. Instead, via stochastic approximation, we compute an approximate solution of $\text{prox}_{\eta F}(x)$, denoted by $\widehat{\text{prox}}_{\eta F}(x)$, implying the inexact gradient of $F_\eta(x)$ is given by $\frac{1}{\eta}(x - \widehat{\text{prox}}_{\eta F}(x))$. In Algorithm 1, the inexact gradient $\nabla_x F_\eta(x_k) + \bar{w}_{k,N_k}$ is defined as

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

We now proceed to develop **(mVS-APM)** for compact domains in Section 5.2.2 and then weaken compactness requirements in Section 2.3.4 for an unaccelerated variant.

2.3.3 Linear convergence of (mVS-APM): compact domains

When $F(x) = \mathbb{E}[\tilde{f}(x,\omega)] + g(x)$, $\text{prox}_{\eta F}(x)$, 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**) scheme [30]. In particular, we propose (**mVS-APM**) with the following update rules for $k \geq 1$,

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

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

where $\widehat{\text{prox}}_{\eta F}(x_k)$ is obtained by taking finite number of steps of (**prox-SSG**) with single sampling having the following update rule for $j = 0, \ldots, N_k - 1$,

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

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, i.e., there exists $\Delta > 0$ such that $\|x\| \leq \Delta$ for any $x \in \text{dom}(g)$.*

Theorem 5 (**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 $Q \triangleq \max\left\{\eta^2 M^2, 4\Delta^2\right\}$.*
(i) **(Rate).** *For all $K \geq 1$, we have that*

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

(ii) **(Outer iteration complexity).** *The iteration complexity of* **(mVS-APM)** *in gradient steps (of $\nabla_x f_\eta(x_k)$) to obtain an ϵ-solution is $\mathcal{O}(\sqrt{\tilde{\kappa}}\log(\widehat{C}/\epsilon))$.*
(iii) **(Oracle complexity)**. *To compute y_{K+1} such that $\mathbb{E}[\|y_K - x^*\|^2] \leq \epsilon$, the complexity of SSG steps is bounded as follows: $\sum_{k=1}^K N_k \leq \frac{2a^2\sqrt{\tilde{\kappa}}\widehat{C}}{(a-1)\epsilon} = \mathcal{O}(1/\epsilon)$.*

Proof. 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**) [30], $\mathbb{E}[\|\bar{w}_{k,N_k}\|^2] \leq \frac{\bar{Q}_k}{\eta^2 N_k}$, where $\bar{Q}_k \triangleq \max\left\{\eta^2 M^2, \|z_{k,0} - z_k^*\|^2\right\} \leq Q$, since $\|z_{k,0} - z_k^*\|^2 \leq 4\Delta^2$. 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 $\tilde{\kappa} = \frac{\mu\eta+1}{\eta\mu}$:

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

From Lemma 4, x^* is minimizer of function F if and only if x^* is a minimizer of function F_η. Since F_η is $\frac{\mu}{\mu\eta+1}$-strongly convex, $\frac{\mu}{2(\mu\eta+1)}\|y_K - x^*\|^2 \leq F_\eta(y_K) - F_\eta(x^*)$, implying (11) can be written as

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

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

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

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

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

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

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

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

$$\tfrac{1}{\rho} = \frac{1}{(1-\frac{1}{2a\sqrt{\tilde{\kappa}}})} = \frac{2a\sqrt{\tilde{\kappa}}}{(2a\sqrt{\tilde{\kappa}}-1)} \implies \frac{\log(\widehat{C})-\log(\epsilon)}{\log(1/\rho)} \leq \frac{\log(\widehat{C})-\log(\epsilon)}{(1-\rho)} = (2a\sqrt{\tilde{\kappa}})\log(\widehat{C}/\epsilon) \leq K.$$

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

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

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

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

$$\implies \tfrac{\sqrt{\tilde{\kappa}}}{\rho^2} \leq \tfrac{a^2\sqrt{\tilde{\kappa}}}{(a^2-a)} = \tfrac{a}{a-1}\sqrt{\tilde{\kappa}} \implies \text{by (16)}, \quad \sum_{k=1}^{\log_{(1/\rho)}(\widehat{C}/\epsilon)+1} \rho^{-k} \leq \tfrac{2a^2\sqrt{\tilde{\kappa}}\widehat{C}}{(a-1)\epsilon}. \quad \square$$

$\square$

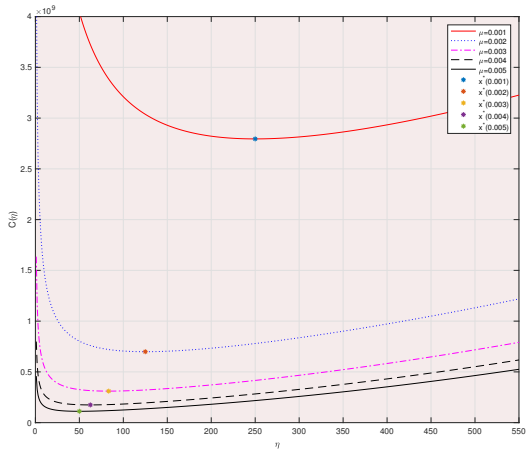

Figure 1: Schematic of $\widehat{C}(\eta)$ when $D = 10, M = 10, C = 100, a = 2.1, \Delta = 1$ for $\mu \in \{0.001, \cdots, 0.005\}$

Remark 1. *In Theorem 5, choosing* $\eta = 1/\mu$ *leads to* $\mathbb{E}[\|y_K - x^*\|^2] \leq \left(\frac{4D}{\mu} + C^2 + 12\sqrt{2}aQ\right)\rho^{K-1}$, *and an oracle complexity of* $\mathcal{O}\left(\frac{\max\{M^2/\mu^2, \|\tilde{x}_1 - \tilde{x}^*\|^2\}}{\epsilon}\right)$, *matching the result by [30].*

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

Lemma 6. *Consider* $\widehat{C}(\eta)$ *defined as* $\widehat{C}(\eta) \triangleq 2D\eta\tilde{\kappa}(\eta) + C^2 + 8\tilde{\kappa}(\eta)^{5/2}Q(\eta)a$, *where* $Q \triangleq \max\{\eta^2 M^2, 4\Delta^2\}$. *Then the following hold.*
(i) $\widehat{C}(\eta)$ *is a coercive function on* $\{\eta \mid \eta \geq 0\}$.
(ii) $\widehat{C}(\eta)$ *is a strictly convex function on* $\{\eta \mid \eta \geq 0\}$.
(iii) The minimizer of $\widehat{C}(\eta)$ *on* $\{\eta \mid \eta \geq 0\}$ *is unique.* □

Remark 2. *Lemma 6 allows for claiming that* $\widehat{C}(\eta)$ *has a unique minimizer* η^**; in fact, such a minimizer can be computed by a standard semismooth Newton method [63]. Fig. 1 provides a schematic of* $\widehat{C}(\eta)$ *for different values of* μ *while* 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.*

2.3.4 Linear convergence of (mVS-PM): non-compact domains

In this subsection, we derive rate and complexity guarantees when **(VS-PM)**, an unaccelerated variant of **(VS-APM)**, is applied on a Moreau-smoothed problem under possibly non-compact 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 due to a recursive relation. Hence, in this section, we consider a more general case, i.e. subgradient of g has a state-dependent bound and by employing an unaccelerated method, we derive a similar oracle complexity as in section 5.2.2. 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 $\min_{u\in\mathbb{R}^n} \left[\mathbb{E}[\tilde{F}(u,\omega)] + \frac{1}{2\eta}\|u - x_k\|^2\right]$ 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.

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

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

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

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, i.e.

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

In other words, z_k^* is defined as

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

Since $\text{prox}_{\eta F}(x_k)$ 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**).

$$z_{k,j+1} = z_{k,j} - \sigma_j G(z_{k,j},\omega_{k,j}), \quad j \geq 0, \text{ where } G(z_{k,j},\omega_{k,j}) \in \partial_F(z_{k,j},\omega_{k,j}) + \tfrac{1}{\eta}(z_{k,j} - x_k). \tag{**SSG**}$$

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

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

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 $S(z_{k,j},\omega_{k,j}) \in \partial_F(z_{k,j},\omega_{k,j})$, $M_1^2 \triangleq 2\bar{M}^2 + \frac{4}{\eta^2}$, $M_2^2 \triangleq \frac{4}{\eta^2}$, and $M_3^2 \triangleq 2M^2$. This requires defining the history upto 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 non-compact 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 $\mathcal{F}_{k,j}$ 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 $S(z_{k,j}, \omega_{k,j}) \in \partial \tilde{F}(z_{k,j}, \omega_{k,j})$. Then the following hold.*
(a) (Unbiasedness). We have that $\mathbb{E}[S(z_{k,j}, \omega_{k,j}) \mid \mathcal{F}_{k,j}] = S(z_{k,j}) \in \partial F(z_{k,j})$ almost surely.
(b) (Subgradient boundedness). There exists $M, \bar{M} > 0$ such that for any x, $\mathbb{E}[\|S(z_{k,j}, \omega_{k,j})\|^2 \mid \mathcal{F}_{k,j}] \leq \bar{M}^2 \|z_{k,j}\|^2 + M^2$ almost surely.

Consequently, we have that

$$\begin{aligned}\|G(z_{k,j}, \omega_{k,j})\|^2 \leq 2\|S(z_{k,j}, \omega_{k,j})\|^2 + \tfrac{2}{\eta^2}\|z_{k,j} - x_k\|^2 \leq 2\|S(z_{k,j}, \omega_{k,j})\|^2 + \tfrac{4}{\eta^2}\|z_{k,j}\|^2 + \tfrac{4}{\eta^2}\|x_k\|^2 \\ \implies \mathbb{E}[\|G(z_{k,j}, \omega_{k,j})\|^2 \mid \mathcal{F}_{k,j}] \overset{\text{Assump. 5}}{\leq} (2\bar{M}^2 + \tfrac{4}{\eta^2})\|z_{k,j}\|^2 + 2M^2 + \tfrac{4}{\eta^2}\|x_k\|^2 \\ =: M_1^2\|z_{k,j}\|^2 + M_2^2\|x_k\|^2 + M_3^2. \qquad (20)\end{aligned}$$

Based on Assumption 5 and inspired by a proof technique from [64] amongst others, we derive a rate statement for (**SSG**) (See Appendix for proof).

Proposition 1. *Consider* (17) *where $F(\cdot, \omega)$ is a μ-strongly convex function and $S(z, \omega) \in \partial_F(z, \omega)$ 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**) *where $\tilde{\mu} = \mu + \frac{1}{\eta}$, $\bar{J} \triangleq \lceil \frac{2M_1^2}{\tilde{\mu}^2} - 1 \rceil$, and*

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

Then the following holds for $j \geq \bar{J}$.

$$\mathbb{E}[\|z_{k,j} - z_k^*\|^2 \mid \mathcal{F}_k] \leq \tfrac{\hat{a}^2\|x_k - x^*\|^2 + \hat{b}^2}{j}. \qquad (21)$$

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

Theorem 7 (**(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 $k \geq 0$, $N_0 > \max\{\frac{2\hat{a}^2}{(1-q/2)}, \bar{J}\}$, $q \triangleq 1 - \frac{1}{\tilde{\kappa}}$, $p_0 \triangleq \frac{q}{2} + \frac{2\hat{a}^2}{N_0}$, and $\bar{J} \triangleq \lceil \frac{2M_1^2}{\tilde{\mu}^2} - 1 \rceil$. Then the following hold.*
(i) **(Rate).** *For all $k \geq 1$, we have that the following holds.*

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

(ii) **(Iteration complexity).** *The iteration complexity of* (**mVS-PM**) *in gradient steps (of $\nabla_x F_\eta(x_k)$) to obtain an ϵ-solution is $\mathcal{O}(\tilde{\kappa}\log(\mathcal{C}/\epsilon))$.*
(iii) **(Oracle complexity in (SSG) steps).** *To compute x_K such that $\mathbb{E}[\|x_K - x^*\|^2] \leq \epsilon$, the complexity in subgradient steps is bounded as $\sum_{k=1}^K N_k \leq \mathcal{O}\left(\tilde{\kappa}\left(\frac{\mathcal{C}}{\epsilon}\right)^{\log_{1/\hat{p}}(1/\rho)}\right)$ for $\hat{p} \in [p_0, 1)$, $\rho \leq p_0$ and $\sum_{k=1}^K N_k \leq \mathcal{O}\left(\tilde{\kappa}\left(\frac{\mathcal{C}}{\epsilon}\right)\right)$ 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 since η is any nonnegative scalar. For instance, if $\eta = \frac{1}{\mu}$, $\tilde{\kappa} = 2$.*

2.4 Iteratively Smoothed VS-APM for Nonsmooth Convex Problems

Thus far, we have considered settings where f is a strongly convex function. However, there are many instances when the function f is neither smooth nor strongly convex. 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. [48] showed that for a subclass of problems, an accelerated gradient scheme may be applied to a suitably *smoothed* problem where 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 $\mathcal{O}(1/\sqrt{k})$. However, since the smoothing parameter is maintained as fixed, Nesterov's approach can provide approximate solutions at best but not asymptotically exact solutions. Subsequently, [49] considered a primal-dual smoothing technique where the smoothing parameter is reduced at every step while extensions and generalizations have been considered more recently by [50] and [56]. 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 [48] where 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 [48]; however, our scheme is blessed with asymptotic guarantees rather than providing approximate solutions. In Section 2.4.1, we derive rate and complexity statements in Section 2.4.2 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 2.4.3, under suitable choices of smoothing sequences, (**sVS-APM**) produces sequences that converge a.s. to an optimal solution.

2.4.1 Smoothing techniques

In this section, we consider minimizing $F(x) \triangleq \mathbb{E}[_F(x,\omega)]$, where $\tilde{f}(x,\omega) = \tilde{f}(x,\omega) + g(x)$ such that f and g are convex and may be nonsmooth while g has an efficient prox evaluation (or "proximable") but f is **not proximable.** Note that this setting is more general than structured nonsmooth problems, where 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, where η_k is a smoothing parameter at iteration k such that its sequence is diminishing. [65] define an (α,β)-smoothable function as follows.

Definition 1 (**(α,β)-smoothable [59]**). *A convex function $h:\mathbb{R}^n \to \mathbb{R}$ is referred to as (α,β)-smoothable if 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) $h_\eta(x)$ 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 $(1,1)$-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 [65] 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 [65]) 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 [48]) and inf-conv smoothing (see [59]); 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 where $\tilde{f}_\eta(\cdot,\omega)$ 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 8. 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. *Consider $\tilde{f}(\cdot,\omega)$ is an $(\alpha(\omega),\beta(\omega))$-smoothable function for every $\omega \in \Omega$ where $\mathbb{E}[\alpha(\omega)] \leq \tilde{\alpha}$ and $\mathbb{E}[\beta(\omega)] \leq \tilde{\beta}$. For any $\omega \in \Omega$, there exists a convex differentiable function $\tilde{f}_\eta(\cdot,\omega)$ with uniform parameter $\eta > 0$ such that*

$$\begin{aligned} \tilde{f}_\eta(x,\omega) \leq \tilde{f}(x,\omega) \leq \tilde{f}_\eta(x,\omega) + \eta\beta(\omega), \quad & \text{for all } x \\ \text{and } \|\nabla_x \tilde{f}_\eta(x,\omega) - \nabla_x \tilde{f}_\eta(y,\omega)\| \leq \tfrac{\alpha(\omega)}{\eta}\|x-y\|, \quad & \text{for all } \ x,y. \end{aligned}$$

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 8. *Suppose $\tilde{f}(\cdot,\omega)$ satisfies Assumption 6 for every $\omega \in \Omega$. 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, $x_1 \in X$, $y_1 = x_1$ and positive sequences $\{\gamma_k, N_k\}$; Set $\lambda_0 = 0$, $\lambda_1 = 1$; $k := 1$.
(1) $y_{k+1} = \mathbf{P}_{\gamma_k,g}\left(x_k - \gamma_k(\nabla_x f_{\eta_k}(x_k) + \bar{w}_{k,N_k})\right)$;
(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).

2.4.2 Rate and Complexity analysis

In this subsection, 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 the 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 dom(g); (ii) f is proper, closed, convex, and $(1, B^2)$-smoothable on an open set containing dom(g); (iii) There exists $C > 0$ such that $\mathbb{E}[\|x_1 - x^*\|] \leq C$ 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 the exposition.

Lemma 9. *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 $\gamma_k \leq \eta_k/2$, then the following holds for all $K \geq 2$:*

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

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

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

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 $g(x) \geq g(y_k) + s^T(x - y_{k+1})$ for all $s \in \partial g(y_k)$. Hence, we obtain the following.

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

We may then obtain that

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

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 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

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

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

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

We may then claim that

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

Consequently, (27) can further bounded as follows:

$$F_{\eta_k}(y_{k+1}) - F_{\eta_k}(y_k) \leq \Big(\frac{1}{2\eta_k} - \frac{1}{\gamma_k}\Big)\|x_k - y_{k+1}\|^2 + \frac{1}{\gamma_k}(x_k - y_{k+1})^T(x_k - y_k) - \bar{w}_{k,N_k}^T(y_{k+1} - y_k)$$

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

Similarly, we have that

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

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

$$\lambda_k\delta_{k+1} - (\lambda_k - 1)\delta_k \leq \Big(\frac{1}{2\eta_k} - \frac{1}{2\gamma_k}\Big)\lambda_k\|y_{k+1} - x_k\|^2 \tag{31}$$
$$+ \frac{1}{2\gamma_k}(\lambda_k - 1)\left(\|x_k - y_k\|^2 - \|y_{k+1} - y_k\|^2\right) + \frac{1}{2\gamma_k}\left(\|x_k - x^*\|^2 - \|y_{k+1} - x^*\|^2\right) \tag{32}$$
$$+ \bar{w}_{k,N_k}^T\left((\lambda_k - 1)y_k + x^* - \lambda_k y_{k+1}\right). \tag{33}$$

Again by using Lemma **??**, we may express the terms in (32) as follows:

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

In addition,

$$\bar{w}_{k,N_k}^T\left((\lambda_k - 1)y_k + x^* - \lambda_k y_{k+1}\right) = \bar{w}_{k,N_k}^T\left((\lambda_k - 1)y_k + x^* - \lambda_k x_k\right) + \bar{w}_{k,N_k}^T\left(\lambda_k x_k - \lambda_k y_{k+1}\right).$$

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, where $u_k = (\lambda_k - 1)y_k - \lambda_k x_k + x^*$:

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \leq \lambda_k^2\left(\frac{1}{2\eta_k} - \frac{1}{2\gamma_k}\right)\|y_{k+1} - x_k\|^2 \tag{34}$$
$$+ \frac{1}{2\gamma_k}\left(-\|\lambda_k y_{k+1} - \lambda_k x_k\|^2 + 2(\lambda_k y_{k+1} - \lambda_k x_k)^T((\lambda_k - 1)y_k + x^* - \lambda_k x_k)\right)$$
$$- \lambda_k^2\bar{w}_{k,N_k}^T(x_k - y_{k+1}) - \lambda_k w_k^T u_k = \lambda_k^2\left(\frac{1}{2\eta_k} - \frac{1}{2\gamma_k}\right)\|y_{k+1} - x_k\|^2 - \lambda_k^2\bar{w}_{k,N_k}^T(x_k - y_{k+1})$$
$$+ \frac{1}{2\gamma_k}\left(\|\lambda_k x_k - (\lambda_k - 1)y_k - x^*\|^2 - \|\lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^*\|^2\right) - \lambda_k w_k^T u_k$$
$$\leq \frac{\lambda_k^2}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\|\bar{w}_{k,N_k}\|^2 + \frac{1}{2\gamma_k}\left(\|u_k\|^2 - \|u_{k+1}\|^2\right) - \lambda_k w_k^T u_k,$$

where in the last inequality we used 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:

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

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

$$\gamma_k\lambda_k^2\delta_{k+1} - \gamma_{k-1}\lambda_{k-1}^2\delta_k \leq \frac{\gamma_k\lambda_k^2}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\|\bar{w}_{k,N_k}\|^2 + \frac{1}{2}\left(\|u_k\|^2 - \|u_{k+1}\|^2\right) - \gamma_k\lambda_k w_k^T u_k. \tag{35}$$

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

$$\gamma_k\lambda_k^2\delta_{k+1} - \gamma_{k-1}\lambda_{k-1}^2\delta_k \leq \gamma_k^2\lambda_k^2\|\bar{w}_{k,N_k}\|^2 + \frac{1}{2}\left(\|u_k\|^2 - \|u_{k+1}\|^2\right) - \gamma_k\lambda_k w_k^T u_k. \tag{36}$$

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

$$\begin{aligned}
&\gamma_{K-1}\lambda_{K-1}^2\delta_K \leq \sum_{k=1}^{K-1}\gamma_k^2\lambda_k^2\|\bar{w}_{k,N_k}\|^2 + \frac{1}{2}\|u_1\|^2 - \sum_{k=1}^{K-1}\gamma_k\lambda_k w_k^T u_k \\
&\implies \delta_K \leq \frac{1}{\gamma_{K-1}\lambda_{K-1}^2}\sum_{k=1}^{K-1}\gamma_k^2\lambda_k^2\|\bar{w}_{k,N_k}\|^2 + \frac{1}{2\gamma_{K-1}\lambda_{K-1}^2}\|u_1\|^2 - \frac{1}{\gamma_{K-1}\lambda_{K-1}^2}\sum_{k=1}^{K-1}\gamma_k\lambda_k w_k^T u_k.
\end{aligned}$$

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

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

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

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

Theorem 10 (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 $K \geq 1$:

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

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

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

Proof. Proof. **(i)** If $N_k = \lfloor k^a \rfloor \geq \frac{1}{2}k^a$ and $\gamma_k = 1/(2k)$ is utilized in Lemma 9, we obtain the following

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

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

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

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

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

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

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

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

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

(ii) Consider y_{K+1} satisfying $\mathbb{E}[F(y_{K+1}) - F(x^*)] \leq \epsilon$. We again consider two cases. (a) $a = 1 + \delta$ *where* $\delta \in [\delta_L, \delta_U]$. Since we have $\frac{\bar{C}}{K} \leq \epsilon$ 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 $2 \leq \bar{C}/\epsilon$:

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

(b) $a = 1$. To compute K such that $\frac{a+b\log(K)}{K} \leq \epsilon$ is not immediately obvious but may be obtained via the Lambert function[4] [67]. For purposes of simplicity, suppose $a = 0$ and $b = 1$. Then we have the following.

$$\begin{aligned}\frac{\log(K)}{K} \leq \epsilon &\Leftrightarrow \frac{-\log(K)}{K} \geq -\epsilon \\ &\Leftrightarrow W_{-1}\left(\frac{-\log(K)}{K}\right) \leq W_{-1}(-\epsilon), \text{ since } W_{-1}(\cdot) \text{ is decreasing.}\end{aligned}$$

[4]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 $x \in [-\frac{1}{e}, 0]$ [66].

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

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

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

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

where the first inequality follows from (3) in [67]. Hence, the oracle complexity for $a = 1$ will be $\mathcal{O}\left(\frac{\log^2(1/\epsilon)}{\epsilon^2}\right)$, which is near optimal (where optimal is $\mathcal{O}(1/\epsilon^2)$). □ □

We now consider two cases of Theorem 10 for which similar rate statements are available.

Case 1. Structured stochastic nonsmooth optimization with f smooth. Now consider problem (1), where $f(x)$ is a smooth function. Recall that we considered such a problem in Section 2.3 for strongly convex f and in this case, we consider the merely convex case. When f is deterministic, accelerated gradient methods first proposed by [36] and their proximal generalizations suggested by [35] were characterized by the optimal rate of convergence of $\mathcal{O}(1/k^2)$. When f is expectation-valued, [33] presented the first known accelerated scheme for stochastic convex optimization where the optimal rate of $1/k^2$ was shown for the expected sub-optimality 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**).[5] [34] developed 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 $\mathcal{O}\left(\epsilon^{-2}\ln^2(\epsilon^{-0.5})\right)$. We note that (**VS-APM**) displays the optimal oracle complexity $\mathcal{O}(\epsilon^{-2})$ by choosing $N_k = \lfloor k^2 K \rfloor$ while 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 11. (Rate and oracle complexity bounds with smooth f for (VS-APM)) *Suppose Assumptions 2, 3, and 7 hold. Suppose $\gamma_k = \gamma \leq 1/2L$ for all k.*
(i) Let $N_k = \lfloor k^a \rfloor$ where $a = 3 + \delta$ and $\widehat{C} \triangleq \frac{2\nu^2\gamma(a-2)}{a-3} + \frac{4C^2}{\gamma}$. Then the following holds.

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

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

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

[5] While pursuing submission of the present work, we were informed of related work by [34] through a private communication.

Proof. Proof. **(i)** Similar to the proof of Lemma 9, by defining $\delta_k = F(y_k) - F(x^*)$ we can prove:

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

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

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

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

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

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

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

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

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

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

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

$\square$

Case 2: Deterministic nonsmooth convex optimization. When the function f in (1) is deterministic but possibly nonsmooth, [48] showed that by 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 10, utilizing a fixed 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 (i-VS-APM) is global (valid for all k) while constant smoothing holds for the prescribed K. We observe that the rate statements by using an appropriately chosen smoothing and steplength parameter matches that by using a selecting a suitable smoothing and steplength sequence.

Corollary 12. (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, $F(y_{k+1}) - F(x^*) \leq \frac{4C^2+B^2}{k}$, for all $k > 0$.* **(ii)** *Fixed smoothing: For a given $K > 0$, suppose $\eta_k = 1/K$ and $\gamma_k = 1/2K$. Then, $F(y_{K+1}) - F(x^*) \leq \frac{4C^2+B^2}{K}$.*

Remark 4. *By recalling that $f_\eta(x) \triangleq \mathbb{E}[\tilde{f}_\eta(x,\omega)]$, by using Theorem 7.47 in [30] (interchangeability of the derivative and the expectation), and noting that $\tilde{f}_\eta(\cdot,\omega)$ is differentiable in x for every ω, we have $\nabla f_\eta(x) = \nabla \mathbb{E}[\tilde{f}_\eta(x,\omega)] = \mathbb{E}[\nabla \tilde{f}_\eta(x,\omega)] \implies \mathbb{E}[\nabla f_\eta(x) - \nabla \tilde{f}_\eta(x,\omega)] = 0$. 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.*

$\tilde{f}(x,\omega)$	$\tilde{f}_\eta(x,\omega)$	$\nabla \tilde{f}_\eta(x,\omega)$	$\mathbb{E}[\|\nabla_x \tilde{f}_\eta(x,\omega) - \nabla_x f_\eta(x)\|^2]$
$\tilde{f}_1(x,\omega) = \lambda(\omega)\|x\|_1$	$\sum_{i=1}^n h_\eta(x_i,\omega)$, where $h_\eta(x_i,\omega) = \left\{ \begin{array}{ll} \lambda^2(\omega)\frac{x_i^2}{2\eta}, & \lambda(\omega)\|x_i\| < \eta \\ \lambda(\omega)\|x_i\| - \eta/2, & o.w. \end{array} \right\}$	$[\nabla_{x_i} h_\eta(x_i,\omega)]_{i=1}^n$, where $\nabla_{x_i} h_\eta(x_i,\omega) = \left\{ \begin{array}{ll} \lambda^2(\omega)\frac{x_i}{\eta}, & \lambda(\omega)\|x_i\| < \eta \\ \lambda(\omega)x_i/\|x_i\|, & o.w. \end{array} \right\}$	$4n\mathbb{E}[\lambda^2(\omega)]$
$\tilde{f}_2(x,\omega) = \lambda(\omega)\|x\|_2$	$\sqrt{\lambda^2(\omega)\|x\|^2 + \eta^2} - \eta$	$\frac{\lambda^2(\omega)x}{\sqrt{\lambda^2(\omega)\|x\|^2+\eta^2}}$	$4\mathbb{E}[\lambda^2(\omega)]$
$\tilde{f}_3(x,\omega) = \max_{1\le i\le n}\{h_i(x,\omega)\}$ where $h_i(x,\omega) = v_i + s_i c(\omega)^T x$	$\eta \log\left(\sum_{i=1}^n \exp(h_i(x,\omega)/\eta)\right)$	$\frac{\sum_{i=1}^n \nabla_x h_i(x,\omega)\exp(h_i(x,\omega)/\eta)}{\sum_{i=1}^n \exp(h_i(x,\omega)/\eta)}$	$4\mathbb{E}\left[\left(\max_{1\le i\le n}\|s_i c(\omega)\|\right)^2\right]$,

Table 2: Bounding the second moments for certain smoothings

2.4.3 Almost sure convergence

While the previous subsection focused on providing rate statements for expected sub-optimality, we now consider the open question of whether the sequence of iterates produced by (**sVS-APM**) converge a.s. to a solution. Schemes employing a constant smoothing parameter preclude such guarantees. Proving a.s. convergence requires using the following lemma.

Lemma 13 (Supermartingale convergence lemma ([68]))**.** *Let $\{v_k\}$ be a sequence of nonnegative random variables, where $\mathbb{E}[v_0] < \infty$ 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. (a.s. 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. Proof. From inequality (34), we have that the following holds.

$$\begin{aligned}\gamma_k\delta_{k+1} &\le \frac{\lambda_{k-1}^2}{\lambda_k^2}\gamma_k\delta_k + \frac{1}{2\lambda_k^2}\left(\|u_k\|^2 - \|u_{k+1}\|^2\right) + \left(\frac{\gamma_k}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\lambda_k}\bar{w}_{k,N_k}^T u_k \\ &\le \frac{\lambda_{k-1}^2}{\lambda_k^2}\gamma_{k-1}\delta_k + \frac{1}{2\lambda_k^2}\left(\|u_k\|^2 - \|u_{k+1}\|^2\right) + \left(\frac{\gamma_k}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\lambda_k}\bar{w}_{k,N_k}^T u_k.\end{aligned}$$

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

$$\begin{aligned}\delta_{k+1} + \frac{1}{2\gamma_k\lambda_k^2}\|u_{k+1}\|^2 &\le \frac{\lambda_{k-1}^2}{\lambda_k^2\gamma_k}\gamma_{k-1}\delta_k + \frac{1}{2\gamma_k\lambda_k^2}\|u_k\|^2 + \left(\frac{1}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\gamma_k\lambda_k}\bar{w}_{k,N_k}^T u_k \\ &= \frac{\lambda_{k-1}^2\gamma_{k-1}}{\lambda_k^2\gamma_k}\left(\delta_k + \frac{\|u_k\|^2}{2\gamma_{k-1}\lambda_{k-1}^2}\right) + \left(\frac{1}{\frac{2}{\gamma_k} - \frac{2}{\eta_k}}\right)\|\bar{w}_{k,N_k}\|^2 - \frac{1}{\gamma_k\lambda_k}\bar{w}_{k,N_k}^T u_k.\end{aligned}$$

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.

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

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

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.

$$\begin{aligned}
\alpha_k &= 1 - \frac{\lambda_{k-1}^2 \gamma_{k-1}}{\lambda_k^2 \gamma_k} = 1 - \frac{\gamma_{k-1}(4\lambda_k^2 - 4\lambda_k)}{4\lambda_k^2 \gamma_k} = \frac{\lambda_k^2 \gamma_k - \gamma_{k-1}\lambda_k^2 + \gamma_{k-1}\lambda_k}{\lambda_k^2 \gamma_k} = \frac{\gamma_k - \gamma_{k-1}}{\gamma_k} + \frac{\gamma_{k-1}}{\lambda_k \gamma_k} \\
&\geq \frac{k^{-b} - (k-1)^{-b}}{k^{-b}} + \frac{(k-1)^{-b}}{k^{1-b}} = \frac{k^{1-b} - (k-1)^{1-b}}{k^{1-b}} \geq \frac{(1-b)}{k}, \quad b \in (0, 1/2],
\end{aligned} \tag{40}$$

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

$$\begin{aligned}
&k\left(\frac{k^{1-b} - (k-1)^{1-b}}{k^{1-b}}\right) = k - k\left(\frac{k-1}{k}\right)^{1-b} = k - k^b(k-1)^{1-b} = k - (k-1)\left(\frac{k}{k-1}\right)^b \\
&= k - (k-1)\left(1 + \frac{1}{k-1}\right)^b = k - (k-1) - b - \frac{b(b-1)}{2!(k-1)^2} - \frac{b(b-1)(b-2)}{3!(k-1)^3} - \dots \\
&= (1-b) + \frac{b(1-b)}{2!(k-1)^2}\left(1 - \frac{(2-b)}{3(k-1)}\right) + + \frac{b(1-b)(2-b)(3-b)}{4!(k-1)^4}\left(1 - \frac{(4-b)}{5(k-1)}\right) + \dots \\
&\geq (1-b), \text{ since } k \geq 2 \geq 1 + \max\left\{\frac{2}{3}, \frac{4}{5}, \frac{6}{7}, \dots\right\}.
\end{aligned}$$

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

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

If $\gamma_k = k^{-b}$ where $b \in (0, 1/2]$ and $N_k = \lfloor k^a \rfloor$ where $a + b > 1$, 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 $b \in (0, 1/2]$:

$$\eta_k - (1-\alpha_k)\eta_{k-1} = \eta_k - \frac{\lambda_{k-1}^2 \gamma_{k-1}}{\lambda_k^2 \gamma_k}\eta_{k-1} = ck^{-b} - \left(1 - \frac{1}{\lambda_k}\right)\frac{c(k-1)^{-2b}}{k^{-b}}$$

$$\leq ck^{-b} - \left(1 - \frac{1}{\lambda_k}\right) ck^{-b} \leq \frac{2c}{k^{1+b}} \implies \sum_{k=1}^{\infty} (\eta_k B^2 - (1-\alpha_k)\eta_{k-1}B^2) < \infty.$$

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

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

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:

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

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 where $f(x)$ 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) *a.s.* .

2.5 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 @1.4GHz with 16GB RAM.

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 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. Since, tractable prox. evals 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) while 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 $\mu = 1e$-4 while 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 6 where we note that as $\mu \ll 1$, the benefit of utilizing η^* is muted. Next, we consider the unconstrained variant (41), where $x \in \mathbb{R}^n$. Since the subgradient is unbounded, we use 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 7, we set $\eta = \frac{1}{\mu} + 1e\text{-}3 > \frac{1}{\mu}$.

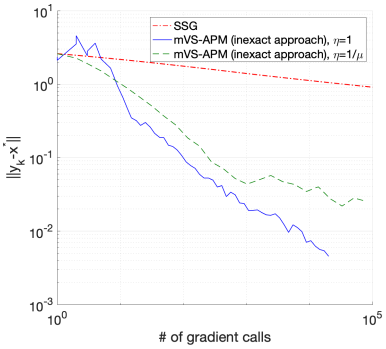

Figure 2: Example 1: (**mVS-APM**) vs SSG for $\mu = 0.1$

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

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 $n = 20$ while in Table 4 (R), we set $\mu = 0.1$ and std. dev. is 0.1. We run both schemes with total budget in subgradient evaluations of 1e5 and 10 replications and observe that **(mVS-APM)** outperforms **(SSG)** .

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

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 iid 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 std. dev. and dimension (n). In Table 6 (L),

we set $\mu = 0.1$ while $n = 20$ and in Table 6 (R), we set $\mu = 0.1$ and std. dev. is 1. Similar to **Example 1**, **(mVS-APM)** outperforms **(SSG)** in all cases.

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	9.9872-1	4.0684	1	7.1652-2	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	1.1027-1	1.1270	50	2.6880e-1	4.1198	1	8.2670e-3	1.3452

2. (sVS-APM). Convex and smoothable f.

Example 4. 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 [65] 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 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 $n = 200$.

Table 7: Example 4: 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

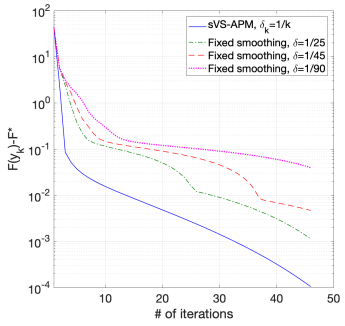

Figure 3: Example 4: **(sVS-APM)** vs fixed smoothing; $n = 200$

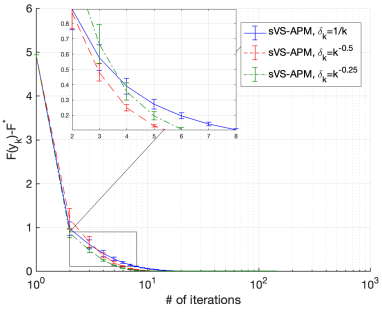

Figure 4: a.s. convergence for **(sVS-APM)**, $N_k = \lfloor k^{3.001} \rfloor$, $\nu^2 = 5$.

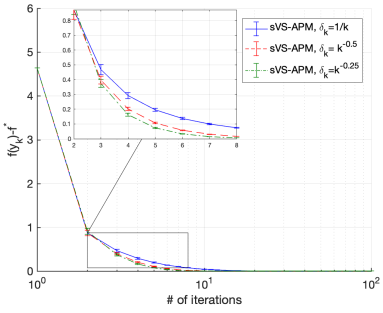

Figure 5: a.s. convergence for **(sVS-APM)**, $N_k = \lfloor k^{3.001} \rfloor$, $\nu^2 = 2$.

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 steplengths early (since $\eta_k \leq \delta_k$) on while in fixed smoothing technique, $\eta_k \leq \delta_k$ (where δ_k may be

quite small). This can be seen in the trajectories where early progress by the iterative smoothing scheme can be observed. A larger δ_k allows for larger steplengths but leads to a coarser approximation of the original problem while smaller δ_k leads to poorer progress but better approximations (See Table 7 and Figure 3).

4. a.s. convergence. Next, we implemented **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 $\delta_k \in \{1/k, 1/\sqrt{k}, 1/k^{0.25}\}$ (where $\delta_k = 1/k$ 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 Figure 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 where the variance is 5. Furthermore, our numerical studies have revealed 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 Prop. 2.

3 Risk-based economic dispatch

The optimal power flow (OPF) problem is one of the most fundamental decision-making problems[6] in power systems operations. There are a host of variants of such problems that include the modeling of reserves [3], allow for modeling security constraints [69], utilizing either DC approximations [70] or full AC formulations [71] of the power flow equations, amongst others [5]. In this chapter, we focus on the economic dispatch (ED) problem with a DC approximation of the power flow equations. With operating reserves and other regulation capacities determined in the day-ahead market [3, 4], economic dispatch decision are usually specified in a short amount of time at real time, with reserves and other regulation capacities are established. The economic dispatch of conventional generation is completed 20 minutes before the hour of delivery [2]. In power system operations, uncertainty plays a key role. Diverse formulations of stochastic optimal power flow along with different uncertainties in power system have been discussed in [1].

Paper	Problem	Model	Algorithm
[72]	DC-OPF	Stochastic	`cplex`
[73]	UC	Robust	`cplex`
[74]	ED	Robust	`cplex`
[75]	ED	Robust	Alternative Direction
[76]	UC	Robust	`cplex`
[77]	UC	Robust	Cutting plane
[78]	UC	Stochastic	Review
[79]	UC	Stochastic	Benders Decomposition / Lagrangian Relaxation
[80]	UC	Stochastic	Lagrangian relaxation
[81]	UC	Stochastic	Benders Decomposition
[82]	UC	Stochastic	Dynamic Programming
[83]	UC	Stochastic	`cplex`
[84]	UC	Stochastic	`cplex`
[85]	UC	Stochastic	Importance Sampling
[86]	DC-OPF	Stochastic	Stochastic Decomposition
[71]	AC-OPF	Stochastic	Scenario reduction by clustering
[87]	DC-OPF	Stochastic	Benders Decomposition
[88]	ED	Deterministic	Gurobi
[89]	ED	Deterministic	`cplex`
[90]	ED	Deterministic	FESTIV

Table 8: Stochastic OPF models

Most prior OPF formulations have only dealt with uncertainty in a rather rudimentary manner by choosing fixed reserve margins without using other known or estimated probabilistic information about forecast errors. Recently, one major approach to deal with such uncertainty has been through robust optimization techniques [6–10] where the uncertain parameters are assumed lie in a suitable uncertainty set and network constraints are enforced for every possible realization of uncertainty. In fact such avenues have been adopted for modeling AC power flow problems (cf. [11]).

Uncertainty in power systems also can be dealt with by adding chance constraints. In general, such avenues lead to possibly nonconvex problems [91] and more recently integer programming approaches have proven useful when contending with a sampled approximation [5, 12]. Here, the authors consider power flow problems subject to chance constraints assuming that the uncertainties are Gaussian. The chance constraint can be expressed as a second-order cone constraint, which turn out to be a convex approximation. Simulation based studies on real time dispatch are also been conducted recently. In [13], a simulation based framework is used in a power system with renewable resources, with system iterates in multiple timescales.

[6]This section has been adapted from the Ph.D. Dissertation by Wendian Wan (advised by PI), titled "Algorithms for Operation of Power Systems: Risk, Uncertainty, Discreteness, and Nonconvexity."

A variety of references consider two-stage model in economic dispatch problem, with different sources of uncertainty. Single contingencies is taken into consideration and objective also includes cost from not meeting demand in [72], where in second stage a contingency parameter is added to transmission capacity in power flow constraint. In [92], model can be extended to muti-stage with unit commitment. A general stochastic dispatch model considering real time regulation in the presence of uncertainties in the offers was proposed in [93]. Constrain on second stage generation was in a form of a random convex set with several different forms, depending on the kind of offer involved. A two-stage economic dispatch model with stochastic producers was proposed in [94], where stochastic producer capacity constraints gets its realization in second stage. In [17], the network risk limiting dispatch problem (N-RLD) was introduced under $\epsilon = 0$, which is a two stage optimization problem under stochastic demand. On second stage, we observe the realization of random demand and balance system based on real demand. In [2], a two-stage model is built for sub-hourly dispatch decisions making. Ramping limits depending on time was also considered in [2]. In general, a two-stage stochastic model can represent the different types of uncertainty with realizations of uncertain factors reveal in second stage.

3.1 Overview of economic dispatch problems

The basic economic dispatch problem requires satisfying load at minimal cost [95], as formulated next.

$$\begin{aligned} \min \quad & \sum_{i\in\mathcal{I}} c_i^g g_i \\ \text{subject to} \quad & \sum_{i\in\mathcal{I}} g_i = d \\ & g_i^{\min} \le g_i \le g_i^{\max}, \quad \forall i \in \mathcal{I} \end{aligned} \tag{EDisp}$$

where g_i, g_i^{min}, g_i^{max} denote the generation level and the minimum, and maximum capacity level associated with generator i housed at bus i, c_i^g represents the unit cost of generation at bus i, d is the total demand, and $\mathcal{I}$ denotes the set of all buses in network. Notice that the economic dispatch problem requires specifying the minimal generation decisions while meeting demand and capacity bounds. In [95], power flow constraints are also taken into consideration together with dispatch decisions. When these power flow constraints are modeled via DC load flow approximations, the resulting bus-specific phase angles need to be considered together with transmission constraints. The resulting model is specified as follows. actual power production and transmission.

$$\begin{aligned} \min \quad & \sum_{i\in\mathcal{I}} c_i^g g_i \\ \text{subject to} \quad & g_i - d_i - \sum_{j\in\mathcal{I}} B_{i,j}(\theta_i - \theta_j) = 0, && \forall i \in \mathcal{I} \\ & B_{i,j}(\theta_i - \theta_j) \le f_{i,j}^{max}, && \forall i,j \in \mathcal{I} \\ & g_i^{min} \le g_i \le g_i^{max}, && \forall i \in \mathcal{I} \\ & \theta_i^{min} \le \theta_i \le \theta_i^{max}, && \forall i \in \mathcal{I} \end{aligned} \tag{42}$$

Where θ_i denotes the phase angle at bus i, $f_{i,j}^{\max}$ represents the transmission line capacity constraint of transmission line between bus i, j, $B_{i,j}$ is the susceptance of transmission line between bus i, j,

and d_i is the demand at bus i. Single contingencies may be taken into consideration the cost of unserved demand may also be modeled [72], leading to the following model.

$$\begin{aligned} \min \quad & \mathbb{E}\left[\sum_{i\in\mathcal{I}}\left(c_i^g g_i^\omega + c_i^r r_i^\omega - c_i^d q_i^\omega\right)\right] \\ \text{subject to} \quad & g_i^\omega - q_i^\omega - \sum B_{i,j}(\theta_i^\omega - \theta_j^\omega) = 0, && \forall i \in \mathcal{I}, \forall \omega \in \Omega \\ & B_{i,j}(\theta_i^\omega - \theta_j^\omega) \le f_{i,j}^{\max}, && \forall i \in \mathcal{I}, \forall \omega \in \Omega \\ & g_i^\omega + r_i^\omega - \chi_i^\omega g_i^{\max} = 0, && \forall i \in \mathcal{I}, \forall \omega \in \Omega \\ & \lambda_i d_i \le q_i^\omega \le d_i, && \forall i \in \mathcal{I}, \forall \omega \in \Omega \end{aligned} \tag{43}$$

where r_i^ω and c^r denote reserve levels and the cost of reserves at bus i, q_i^ω and c_i^d represents unserved demand and the cost of unserved demand at bus i, Ω represents the scenario space, χ_i^ω denotes the proportion that generator i under outage would reduce in capacity by (i.e. the capacity of outage generator would be $\chi_i^\omega g_i^{\max}$ where $\chi_i^\omega \in [0,1]$), d_i represents the forecast demand, and $\lambda_i d_i$ denotes a critical level of demand that needs to be satisfied. In [92], this model is further extended to a multi-stage problems with unit commitment. A more general stochastic dispatch model was proposed in [93]

$$\begin{aligned} \min \quad & \sum_{i\in\mathcal{I}}(c_i g_i + \mathbb{E}[c_i^+(g_i^\omega - g_i)_+ - c_i^-(g_i^\omega - g_i)_-]) \\ \text{subject to} \quad & g_i - \sum_{j\in\mathcal{I}} f_{i,j} - d_i = 0, && \forall i \in \mathcal{I} \\ & g_i^\omega - \sum_{j\in\mathcal{I}} f_{i,j}^\omega - d_i = 0, && \forall i \in \mathcal{I}, \forall \omega \in \Omega \\ & (g_i, g_i^\omega) \in C_i^\omega, && \forall \omega \in \Omega, \forall i \in \mathcal{I} \\ & f_{i,j} \le f_{i,j}^{\max}, && \forall i \in \mathcal{I}, \\ & f_{i,j}^\omega \le f_{i,j}^{\max}, && \forall i \in \mathcal{I}, \forall \omega \in \Omega. \end{aligned}$$

In this model, g_i^ω denotes the second-stage ("real time") decision $f_{i,j}$ and $f_{i,j}^\omega$ represents the first and scenario-specific second-stage power flow on the transmission line, and C_i^ω represents a random convex set that may take several different forms, depending on the kind of offer involved.

- *Completely inflexible generation*: First-stage dispatched quantity, denoted by g_i cannot be varied in the second stage.

$$g_i^\omega = g_i \in [0, g_i^{\max}], \quad \forall \omega \in \Omega, \forall i \in \mathcal{I}.$$

- *Completely flexible generation*: First-stage dispatched quantity, denoted by g_i, may be varied in the second stage.

$$g_i \in [0, g_i^{\max}],\ g_i^\omega \in [0, g_i^{max}], \quad \forall \omega \in \Omega, \forall i \in \mathcal{I}.$$

- *Unpredictable or intermittent generation*: A generator with maximum capacity $g_i^{\max}$ offers a random quantity S_i.

$$g_i \in [0, g_i^{\max}],\ g_i^\omega \in [0, S_i^\omega], \quad \forall \omega \in \Omega, \forall i \in \mathcal{I}$$

- *Demand-side bid*: A quantity $-q_i \geq 0$ is bid for in the first stage while in the second-stage, this can be modified.

$$g_i \in [q_i, 0],\ g_i^\omega \in [q_i, 0], \quad \forall \omega \in \Omega, \forall i \in \mathcal{I}$$

- *Unpredictable load*: The second-stage demand-side bid g_i^ω has to be feasible with regard to a random load of size $d_i^\omega \geq 0$.

$$g_i \leq 0,\ g_i^\omega \in [-d_i^\omega, 0], \quad \forall \omega \in \Omega, \forall i \in \mathcal{I}.$$

In [14], a risk-limiting dispatch framework is introduced and this was subsequently extended in [15], where a computable closed-form formulas was derived. In [16], a related model for risk-limiting dispatch with generation limitation and network constraint was represented. In [17], a *two-stage* network risk limiting dispatch problem (N-RLD) was introduced under $\epsilon = 0$, where the first-stage represents the day-ahead scheduling while the second-stage captures real-time decisions. The overall problem requires minimizes expected cost of operation as captured by the following problem.

3.2 Two-stage stochastic economic dispatch

In the two-stage model for stochastic economic dispatch, first-stage decisions are given by slow-response generation decisions while second-stage decisions adapt to the realization of uncertainty and are tied to first-stage decisions. The nature of the uncertainty in the second-stage pertains to the randomness in real-time cost, randomness in real-time demand, line contingencies, and uncertainty in the availability of renewable The goal of two-stage model is to have determine a set of first decisions that minimize the sum of two costs: (i) the first-stage cost; and (ii) the risk-adjusted second-stage expected cost of contending with uncertainty.
We consider the stochastic economic dispatch problem faced over T hours while sub-hourly decisions associated with fast-response generators are made between hours (i.e. in a sub-hourly sense) to contend with uncertainty. We view the first-stage decisions as the hourly decisions from $t = 1, \ldots, T$ while the sub-hourly decisions are viewed as the recourse second-stage decisions Let g_t denote the first stage decisions at period t with a convex differentiable generation cost denoted by $f_0(g_t)$ while the cost of recourse decisions under realization ω in the sub-hourly period after t is denoted by $h(g_t^\omega, \omega)$.

3.3 Risk-neutral stochastic economic dispatch

In a power system network, there are several types of constraints that need consideration.

Flow balance equations

In both stages, flow balance needs to be maintained at each bus for every period. In the first stage, flow balance would be based on forecasted demand as follows:

$$g_t - B_1\theta_t - \hat{d}_t = 0,\ t = 1, \ldots, T, \tag{44}$$

where θ_t denotes the phase angle of each bus at period t, $B = (b_{i,j})_{n\times n}$ represents the susceptance matrix,

$$B_1 \triangleq B - \text{diag}\left(\sum_j b_{i,j}\right),$$

and $\hat{d}_t$ denotes a forecast demand in period t in first stage. During the sub-hourly dispatch after period t, the recourse decisions have to satisfy the following requirement.

$$g_t + g^r_{t,\omega} - B_1\theta_{t,\omega} - d_{t,\omega} - d^u_{t,\omega} - d^w_{t,\omega} \geq 0, \tag{45}$$

where $g^r_{t,\omega}$ denotes the recourse generation adjustment in period t, $\theta_{t,\omega}$ represents the second-stage phase angle for period t, and $d_{t,\omega}$ denotes the observed demand in period t under realization ω. Suppose $d^u_{t,\omega}$ represents undispatchable generation while $d^w_{t,\omega}$ denotes unavailable wind power generation for period t.

Line flow constraints

In both stages, transmission capacity constraints need to be satisfied while line contingencies are considered in the second stage. The first stage line constraints can be expressed as follows:

$$-f_{\max} \leq B_2\theta_t \leq f_{\max}, t = 1, \ldots, T \tag{46}$$

where $f_{\max}$ denotes flow bounds while B_2 is defined as follows.

$$B_2 = \begin{pmatrix} \begin{pmatrix} b_{1,\bullet} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} - \operatorname{diag}(b_{1,\bullet}) \\ \begin{pmatrix} \mathbf{0} & b_{2,\bullet} & \cdots & \mathbf{0} \end{pmatrix} - \operatorname{diag}(b_{2,\bullet}) \\ \vdots \\ \begin{pmatrix} \mathbf{0} & \mathbf{0} & \cdots & b_{n,\bullet} \end{pmatrix} - \operatorname{diag}(b_{n,\bullet}) \end{pmatrix}.$$

In the second stage, line contingencies are addressed as follows:

$$-\chi_{t,\omega} f_{\max} \leq B_2\theta_{t,\omega} \leq \chi_{t,\omega} f_{\max}, \ \forall t, \tag{47}$$

where $\chi_{t,\omega}$ is a vector represents the realization of stochastic line contingency of all transmission lines for period t with all of its elements value in $[0, 1]$.

Generation capacity constraints

Both conventional and renewable generators we have to abide by capacity constraints.

$$g^{\min} \leq g_t \leq g^{\max} \forall t \tag{48}$$

$$g^{\min}_{t,\omega} \leq g_t + g^r_{t,\omega} \leq g^{\max}_{t,\omega}, \tag{49}$$

where $g^{\min}_{t,\omega}$ and $g^{\max}_{t,\omega}$ denote second-stage capacity bounds at period t based on renewable availability.

Ramping constraints

Both conventional and renewable generators need to satisfy ramping requirements, as specified next.

$$r^{\min}_t \leq g_t - g_{t-1} \leq r^{\max}_t, \ \forall t \tag{50}$$

$$r^{\min}_{t,\omega} \leq g_t + g^r_{t,\omega} - g_{t-1} - g^r_{t-1,\omega} \leq r^{\max}_{t,\omega}, \tag{51}$$

where $r^{\min}$ and $r^{\max}$ denote the minimum and maximum ramp limit, $g_0 = 0$, and $r^{\min}_{t,\omega}$ and $r^{\max}_{t,\omega}$ denote second-stage down and up ramping limits for period t and $g^r_{0,\omega} = 0$.

Phase angle bounds

In both stages, phase angles bound are imposed in the following fashion.

$$\theta_{\min} \leq \theta_t \leq \theta_{\max}, \ \forall t \tag{52}$$

$$\theta_{\min} \leq \theta_{t,\omega} \leq \theta_{\max}. \tag{53}$$

To summarize, the two-stage stochastic economic dispatch model is defined as follow:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \sum_{t=1}^{T} [f_0(\mathbf{x}_t) + \mathbb{E}[\mathcal{Q}(\mathbf{x}_t, \omega)]] \\ \text{subject to} \quad & (44), (46), \ldots, (52). \end{aligned} \tag{s-ED}$$

where $\mathbf{x} \triangleq (g, \theta)$, $\mathcal{Q}(\mathbf{x}_t, \omega)$ is defined as follows.

$$\begin{aligned} \mathcal{Q}(\mathbf{x}_t, \omega) = \min_{g^r_{t,\omega}, \theta_{t,\omega}} \quad & f(g^r_{t,\omega}, \omega) \\ \text{subject to} \quad & (45), (47), \ldots, (53), \end{aligned} \tag{s-ED$_2^\omega$}$$

$f(g^\omega_t, \omega)$ denotes the random second-stage generation cost and $g^r_{t,\omega}$ denotes second stage decision variable at period t under random variable realization ω.

Deterministic equivalent

Suppose the ω takes on a finite number of realizations given by $\omega_1, \ldots, \omega_K$ with probabilities $p_1, \ldots, p_K$. The resulting deterministic equivalent optimization problem is given by the following.

$$\begin{aligned} \min_{\mathbf{x}_t, \mathbf{y}_{t,\omega}} \quad & \sum_{t=1}^{T} \left[f_0(\mathbf{x}_t) + \sum_{j=1}^{K} p^{\omega_j} [f(\mathbf{y}_{t,\omega}, \omega_j)] \right] \\ \text{subject to} \quad & (44), (46), \ldots, (52) \\ & (45), (47), \ldots, (53), \qquad \forall t, \omega \in \Omega. \end{aligned} \tag{DE-ED}$$

where $\mathbf{y}_\omega \triangleq (g^r_\omega, \theta_\omega)$. Generally, (DE-ED) is not solved directly since Ω can have a large cardinality, necessitating the development of decomposition and sampling schemes.

3.4 Risk-averse economic Dispatch

While the prior model employs a risk-neutral framework, this can be generalized to allow for risk preferences; a commonly employed approach utilizes the conditional value-at-risk (CVaR). Recall for a fixed level τ, the conditional value-at-risk of a random loss function $Z(\omega)$ is defined as [96]:

$$\mathbf{CVaR}_\tau(Z(\omega)) \triangleq \min_m \left\{ m + \frac{1}{1-\tau} \mathbb{E}\left[Z(\omega) - m\right]^+ \right\}. \tag{54}$$

In this setting, we choose $Z(\omega) \triangleq \mathcal{Q}(\mathbf{x}_t, \omega)$. It may be recalled that the minimization of a CVaR-based objective can be recast as follows:

$$\min_{\mathbf{x}} \{ f_0(\mathbf{x}) + \mathbf{CVaR}_\tau(\mathcal{Q}(\mathbf{x}, \omega)) \}$$

$$
\begin{aligned}
&= \min_{\mathbf{x}} \left\{ f_0(\mathbf{x}) + \min_{\mathbf{m}} \left\{ \mathbf{m} + \frac{\mathbf{1}}{\mathbf{1}-\tau} \mathbb{E}\left[\mathcal{Q}(\mathbf{x},\omega) - \mathbf{m}\right]^+ \right\} \right\} \\
&= \min_{\mathbf{x},\mathbf{m}} \left\{ \underbrace{f_0(\mathbf{x}) + \mathbf{m}}_{\triangleq c(\mathbf{z})} + \frac{1}{1-\tau} \mathbb{E}\left[\mathcal{Q}(\mathbf{x},\omega) - \mathbf{m}\right]^+ \right\} \\
&= \min_{\mathbf{z}} \left\{ c(\mathbf{z}) + \mathbb{E}\left[r(\mathbf{z},\omega)\right] \right\}, \text{ where } r(\mathbf{z},\omega) \triangleq \frac{1}{1-\tau} \left[\mathcal{Q}(\mathbf{x},\omega) - \mathbf{m}\right]^+ .
\end{aligned} \tag{55}
$$

In the next section, we discuss a smoothed accelerated gradient scheme for the two-stage stochastic convex problem:

$$
\min_{\mathbf{z}\in Z} \left[c(\mathbf{z}) + \mathbb{E}\left[r(\mathbf{z},\omega)\right] \right], \tag{r-ED}
$$

where $\mathbf{z} \triangleq (\mathbf{x},\mathbf{m})$, $c(\mathbf{z})$ is a convex differentiable function, $r(\mathbf{z},\omega)$ is a convex and nonsmooth function defined as (55), and Z is a polyhedral set defined by the constraints (44), (46), ..., (52).

3.5 Smoothing

Consider the function $r(\mathbf{z},\omega)$, defined in (55). This function has two sources of nonsmoothness: (i) The function $\mathcal{Q}(\bullet,\omega)$ is a convex nonsmooth function of $(\bullet)$; and (ii) The function $[u]^+ = \max\{u,0\}$ is a nonsmooth function of u. We intend to develop algorithms in which the gradient of a smoothed counterpart of $r(\mathbf{z},\omega)$, denoted by $r_\mu(\mathbf{z},\omega)$, is employed. Before proceeding, we recap the definition of smoothability of a convex function.

Definition 2 (Smoothable function [97])**.** *A convex function $d : \mathbb{R}^n \to \mathbb{R}$ is called (α,β)-smoothable if for any $\mu > 0$ there exists a convex differentiable function $d_\mu : \mathbb{R}^n \to \mathbb{R}$ such that the following holds for some $\alpha,\beta > 0$:*

(i) $d_\mu(x) \le d(x) \le d_\mu(x) + \beta\mu$ for all $x \in \mathbb{R}^n$.

(ii) d_μ is $\frac{\alpha}{\mu}$-smooth.

Then the function d_μ is called a $\frac{1}{\mu}$-smooth approximation of h with parameters (α,β).

3.5.1 Smoothing the recourse function $\mathcal{Q}(g,\omega)$.

Consider the recourse function $\mathcal{Q}(\mathbf{x},\omega)$ whose evaluation requires solving the following parametrized convex problem.

$$
\begin{aligned}
\min_{\mathbf{y}\in Y^\omega} \quad & d(\mathbf{y},\omega) \\
\text{subject to} \quad & W_\omega \mathbf{y} \le q_\omega - T_\omega \mathbf{x}, \qquad (\pi)
\end{aligned} \tag{P-$\mathcal{Q}(\mathbf{x},\omega)$}
$$

We make the following assumptions on (P-$\mathcal{Q}(\mathbf{x},\omega)$).

Assumption 8. *Consider the recourse problem (P-$\mathcal{Q}(\mathbf{x},\omega)$).*
(i) For every $\omega \in \Omega$, the function $d : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$ is convex and Y^ω is a closed, convex, and bounded polyhedron with a nonempty interior.
(ii) For every $\omega \in \Omega$ and every $x \in X_0$, there exists a $\bar{\mathbf{y}}(\omega)$ such that $W_\omega \mathbf{y}_\omega = q_\omega - T_\omega \mathbf{x}$ and $\bar{\mathbf{y}}_\omega \in int(Y^\omega)$.

The dual of (P-$\mathcal{Q}(\mathbf{x},\omega)$) is given by the following.

$$\max_{\pi\geq 0} \quad \underbrace{(q^\omega - T^\omega \mathbf{x})^T\pi - \bar{d}^*(W_\omega^T\pi;\omega)}_{\triangleq\varepsilon(\pi;\mathbf{x},\omega)}, \tag{D-$\mathcal{Q}(\mathbf{x},\omega)$}$$

where $\bar{d}^*(\mathbf{y};\omega)$ is the convex conjugate of $\bar{d}(y;\omega)$, defined as follows.

$$\bar{d}(\mathbf{y};\omega) \triangleq \begin{cases} d(\mathbf{y};\omega), & \mathbf{y}\in Y^\omega \\ +\infty, & \text{otherwise.} \end{cases} \tag{56}$$

By convex duality, the optimal values of (P-$\mathcal{Q}(\mathbf{x},\omega)$) and (D-$\mathcal{Q}(\mathbf{x},\omega)$) are equal for a given $\mathbf{x}$.

Lemma 14. *Suppose $d(\mathbf{y},\omega)$ is a convex function in $\mathbf{y}$ for every $\omega\in\Omega$. If for some $\mathbf{x}$, (P-$\mathcal{Q}(\mathbf{x},\omega)$) has an optimal solution. Then the dual problem (D-$\mathcal{Q}(\mathbf{x},\omega)$) has an optimal solution and the optimal values of both problems are equal.*

From the theory of stochastic programming, it is known that the recourse function $\mathcal{Q}(\mathbf{x},\omega)$ is a convex function in $\mathbf{x}$ for every ω [98, Prop. 2.21]. Consider a modified function $\mathcal{Q}_\mu(\mathbf{x},\omega)$, defined as the optimal value of this μ-regularized dual problem (D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$).

$$\max_{\pi\geq 0} \ (\varepsilon(\pi;\mathbf{x},\omega) - \mu\|\pi\|^2), \tag{D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$}$$

Lemma 15. *Suppose v_ω and $v_{\mu,\omega}$ denote the optimal values of (D-$\mathcal{Q}(\mathbf{x};\omega)$) and (D-$\mathcal{Q}_\mu(\mathbf{x};\omega)$), respectively. In addition, suppose π_ω denotes an optimal solution of (D-$\mathcal{Q}(\mathbf{x};\omega)$). Then $v_{\mu,\omega}\geq v_\omega - \mu\|\pi_\omega\|^2$.*

Proof. The result can be concluded as follows.

$$\begin{aligned} v_{\mu,\omega} &\triangleq \max_{\pi\geq 0}\ \{(\varepsilon(\pi;\mathbf{x},\omega) - \mu\|\pi\|^2)\} \\ &\geq \ \left\{(\varepsilon(\pi^*_\omega;\mathbf{x},\omega) - \mu\|\pi^*_\omega\|^2)\right\} \qquad (\text{where } \pi^*_\omega \in \arg\max_{\pi\geq 0}\{\varepsilon(\pi;\mathbf{x},\omega)\}) \\ &= \ \left\{\max_{\pi\geq 0}\varepsilon(\pi;\mathbf{x},\omega)\right\} - \mu\|\pi^*_\omega\|^2 \\ &= v_\omega - \mu\|\pi^*_\omega\|^2. \end{aligned}$$

□

Before proceeding, we show that the solution of (D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$) is bounded under the Slater regularity condition. This requires defining the Lagrangian function $\mathcal{L}(\mathbf{y},\pi,\omega)$ and the dual function $\mathcal{D}(\pi,\omega)$ associated with the primal problem, which we proceed to do next.

$$\mathcal{L}(\mathbf{y},\pi,\omega) \triangleq \left(d(\mathbf{y},\omega) + (W_\omega\mathbf{y} - q_\omega + T_\omega x)^T\pi\right) \tag{57}$$

$$\mathcal{D}(\pi,\omega) \triangleq \min_{\mathbf{y}}\ \mathcal{L}(\mathbf{y},\pi,\omega). \tag{58}$$

Lemma 16. *Consider the problem (P-$\mathcal{Q}(\mathbf{x},\omega)$) and suppose it has optimal value v_ω. Then the following hold.*
(i) Suppose for given $\mathbf{x}$ and $\omega\in\Omega$, there exists a $\bar{\mathbf{y}}\in Y^\omega$ such that $c(\bar{\mathbf{y}},\mathbf{x})<0$ where $c(\bar{\mathbf{y}},\mathbf{x})\triangleq W_\omega\bar{\mathbf{y}} - q_\omega + T_\omega\mathbf{x}$. Then the solution set of (D-$\mathcal{Q}(\mathbf{x},\omega)$) lies in a compact set.
(ii) In addition, if $\bar{\mathbf{y}}\in\cap_{\omega\in\Omega}Y^\omega$, $-c(\bar{\mathbf{y}},\omega)\geq -\bar{c}$ and $d(\bar{\mathbf{y}},\omega) - v_\omega\leq\bar{d}$ for every $\omega\in\Omega$, then the solution set of (D-$\mathcal{Q}(\mathbf{x},\omega)$) is uniformly bounded in ω, i.e. $\|\pi_\omega\|\leq\frac{\bar{d}}{\bar{c}}$ for any $\omega\in\Omega$ and for any $\pi_\omega\in SOL(D\text{-}\mathcal{Q}(\mathbf{x},\omega))$.

(iii) If $\pi_{\mu,\omega}$ denotes the optimal dual solution to (D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$), then we have that

$$\|\pi_{\mu,\omega}\| \leq \frac{d(\bar{\mathbf{y}},\omega) - v_\omega}{-\max_{1\leq j\leq m} c_j(\bar{\mathbf{y}};\omega)} + \mu\left(\frac{d(\bar{\mathbf{y}},\omega) - v_\omega}{-\max_{1\leq j\leq m} c_j(\bar{\mathbf{y}};\omega)}\right)^2.$$

(iv) In addition, if $\bar{\mathbf{y}} \in \cap_{\omega\in\Omega} Y^\omega$, $-c_j(\bar{\mathbf{y}},\omega) \geq -\bar{c}$ for $j = 1,\ldots,m$ and $d(\bar{\mathbf{y}},\omega) - v_\omega \leq \bar{d}$ for every $\omega \in \Omega$, then the solution set of (D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$ is uniformly bounded in ω and μ when $\mu \leq \bar{\mu}$, i.e. $\|\pi_{\mu,\omega}\| \leq \frac{\bar{d}}{\bar{c}} + \bar{\mu}\left(\frac{\bar{d}}{\bar{c}}\right)^2$ for any $\omega \in \Omega$, where $\pi_{\mu,\omega} = argmax_{\pi\geq 0}$(D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$).

Proof. (i) For any optimal dual solution π^*_ω, we have from strong duality,

$$\begin{aligned} v_\omega = \mathcal{D}(\pi^*_\omega,\omega) &= \inf_{\mathbf{y}\in Y^\omega} \{d(\mathbf{y},\omega) + (\pi^*_\omega)^T c_\omega(\mathbf{y};\mathbf{x})\} \\ &\leq\ d(\bar{\mathbf{y}},\omega) + (\pi^*_\omega)^T c_\omega(\bar{\mathbf{y}};\mathbf{x}) \\ &\leq d(\bar{\mathbf{y}},\omega) + \max_{1\leq j\leq m} c_{\omega,j}(\bar{\mathbf{y}};\mathbf{x}) \sum_{j=1}^m \pi^*_{\omega,i}. \end{aligned}$$

Consequently, we have the following relationship.

$$-(\max_{1\leq j\leq m} c_{\omega,j}(\bar{\mathbf{y}};\mathbf{x})) \sum_{j=1}^m \pi^*_{\omega,i} \leq d(\bar{\mathbf{y}},\omega) - v_\omega \tag{59}$$

$$\implies \|\pi^*_\omega\| \leq \sum_{j=1}^m \pi^*_{\omega,i} \leq \frac{d(\bar{\mathbf{y}},\omega) - v_\omega}{-\max_{1\leq j\leq m}(c_{\omega,j}(\bar{\mathbf{y}};\mathbf{x}))}. \tag{60}$$

(ii) By hypothesis, we have that $-\max_{1\leq j\leq m}(c_{\omega,j}(\bar{\mathbf{y}};\mathbf{x})) > \bar{c}$ and $d(\bar{\mathbf{y}},\omega) - v_\omega \leq \bar{d}$ for every $\omega \in \Omega$. Consequently, $\|\pi^*_\omega\| \leq \frac{\bar{d}}{\bar{c}}$ for all $\omega \in \Omega$.
(iii) Consider the regularized dual function $\mathcal{D}_\mu(\pi)$ and suppose its optimal value is $v_{\mu,\omega}$. Then the following sequence of inequalities hold.

$$\begin{aligned} & v_{\mu,\omega} = \mathcal{D}_\mu(\pi^*_{\omega,\mu}) \\ = & \inf_{\mathbf{y}\in Y^\omega} \{d(\mathbf{y},\omega) + (\pi^*_{\mu,\omega})^T c_\omega(\mathbf{y};\mathbf{x}) - \mu\|\pi^*_{\mu,\omega}\|^2\} \\ \leq & \ \{d(\mathbf{y},\omega) + (\pi^*_{\mu,\omega})^T c_\omega(\mathbf{y};\mathbf{x}) - \mu\|\pi^*_{\mu,\omega}\|^2\} \\ \leq & \ d(\mathbf{y},\omega) + (\pi^*_{\mu,\omega})^T c_\omega(\bar{\mathbf{y}};\mathbf{x}) \\ \leq & \ d(\mathbf{y},\omega) + \max_{1\leq j\leq m} c_{\omega,j}(\bar{\mathbf{y}};\mathbf{x}) \sum_{j=1}^m \pi^*_{\mu,\omega,i}. \end{aligned}$$

It follows that

$$\begin{aligned} \|\pi^*_{\mu,\omega}\| &\leq \frac{d(\bar{\mathbf{y}},\omega) - v_{\mu,\omega}}{-\max_{1\leq j\leq m} c_j(\bar{\mathbf{y}};\omega)} \\ &\overset{\text{Lemma 14}}{\leq} \frac{d(\bar{\mathbf{y}},\omega) - v_\omega + \mu\|\pi^*_\omega\|^2}{-\max_{1\leq j\leq m} c_j(\bar{\mathbf{y}};\omega)} \\ &\overset{(60)}{\leq} \frac{d(\bar{\mathbf{y}},\omega) - v_\omega}{-\max_{1\leq j\leq m} c_j(\bar{\mathbf{y}};\omega)} + \mu\left(\frac{d(\bar{\mathbf{y}},\omega) - v_\omega}{-\max_{1\leq j\leq m} c_j(\bar{\mathbf{y}};\omega)}\right)^2. \end{aligned}$$

(iv) By hypothesis, we have that $\min_{1\leq j\leq m}(-c_{\omega,j}(\bar{\mathbf{y}};\mathbf{x})) > \bar{c}$ and $d(\bar{\mathbf{y}},\omega) - v_\omega \leq \bar{d}$ for every $\omega \in \Omega$. Consequently, $\|\pi^*_{\mu,\omega}\| \leq \frac{\bar{d}}{\bar{c}} + \bar{\mu}\left(\frac{\bar{d}}{\bar{c}}\right)^2$ for all $\omega \in \Omega$ and for every $\mu \leq \bar{\mu}$.

□

We proceed to show that $\mathcal{Q}_\mu(\mathbf{x},\omega)$ is an (α,β)-smoothing of $\mathcal{Q}(\mathbf{x},\omega)$, where $\mathcal{Q}_\mu(\mathbf{x},\omega)$ which is defined as the optimal value of (D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$).

Lemma 17. *Suppose $\mathcal{Q}(\mathbf{x},\omega)$ is defined by the optimal value of (P-$\mathcal{Q}(\mathbf{x},\omega)$). Then the following hold:*
(i) The function $\mathcal{Q}(\mathbf{x},\omega)$ is a convex function in $\mathbf{x}$ for every $\omega\in\Omega$.
(ii) The function $\mathcal{Q}_\mu(\mathbf{x},\omega)$ is a differentiable in $\mathbf{x}$ at every ω and $\nabla_{\mathbf{x}}\mathcal{Q}_\mu(\mathbf{x},\omega) = -T_\omega^T\pi^*(\mathbf{x},\omega)$, where $\pi^*(\mathbf{x},\omega)$ denotes the optimal solution of (D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$).

Proof. [98, Prop. 2.22]. □

Proposition 4 ($\mathcal{Q}_\mu(\mathbf{x},\omega)$ satisfies $(\alpha(\omega),\beta(\omega))$-smoothability of $\mathcal{Q}(\mathbf{x},\omega)$)**.** *Consider the function $\mathcal{Q}_\mu(\mathbf{x},\omega)$ defined by (D-$\mathcal{Q}_\mu(\mathbf{x},\omega)$). Then this function satisfies the following:*
(i) The function $\mathcal{Q}_\mu(\mathbf{x},\omega)$ is $\frac{\|T_\omega\|^2}{\mu}$-smooth, i.e.

$$\|\nabla_{\mathbf{x}}\mathcal{Q}_\mu(\mathbf{x}_1,\omega)-\nabla_{\mathbf{x}}\mathcal{Q}_\mu(\mathbf{x}_2,\omega)\|\le\frac{\|T_\omega\|^2}{\mu}\|\mathbf{x}_1-\mathbf{x}_2\|,\qquad\forall\mathbf{x}_1,\mathbf{x}_2.$$

(ii) There exists a $\beta(\omega)$ such that for all $\mathbf{x}$, we have that for all $\mathbf{x}$,

$$\mathcal{Q}_\mu(\mathbf{x},\omega)\le\mathcal{Q}(\mathbf{x},\omega)\le\mathcal{Q}_\mu+\mu\beta(\omega).$$

Proof. (i) Consider an $\mathbf{x}_1,\mathbf{x}_2\in X$ and let $\pi(\mathbf{x}_1)$ and $\pi(\mathbf{x}_2)$ denote the maximizers of (D-$\mathcal{Q}_\mu(\mathbf{x}_1,\omega)$) and (D-$\mathcal{Q}_\mu(\mathbf{x}_2,\omega)$), respectively. By the strong concavity of the objective, we have that

$$\begin{aligned}&((-q_\omega+T_\omega\mathbf{x}_1+a(\pi(\mathbf{x}_1),\omega))-(-q_\omega+T_\omega\mathbf{x}_2+a(\pi(\mathbf{x}_2),\omega)))^T(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2))\\&+\mu(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2)))^T(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2))\\&\ge\mu\|\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2)\|^2,\end{aligned}\tag{61}$$

where $a(\pi(\mathbf{x}),\omega)\in\partial_\pi(\bar{d}^*_\omega(W_\omega^T\pi;\omega))$. In addition, by definition, we have that

$$(T_\omega\mathbf{x}_1-q_\omega+a(\pi(\mathbf{x}_1),\omega)+\mu\pi(\mathbf{x}_1))^T(\pi(\mathbf{x}_2)-\pi(\mathbf{x}_1))\ge0.\tag{62}$$

Adding (61) and (62), we obtain that

$$\begin{aligned}&(T_\omega\mathbf{x}_2-q_\omega+a(\pi(\mathbf{x}_2),\omega)+\mu\pi(\mathbf{x}_2))^T(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2))\\&\ge\mu\|\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2)\|^2.\end{aligned}\tag{63}$$

Consequently, by adding and subtracting $(T_\omega\mathbf{x}_1-q_\omega+a(\pi(\mathbf{x}_1,\omega)+\mu\pi(\mathbf{x}_1))^T$ $(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2))$,

$$\begin{aligned}&\underbrace{(T_\omega\mathbf{x}_1+a(\pi(\mathbf{x}_1,\omega)+\mu\pi(\mathbf{x}_1))^T(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2))}_{\le0}\\&+(T_\omega\mathbf{x}_2-T_\omega\mathbf{x}_1)^T(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2))+\underbrace{(a(\pi(\mathbf{x}_2),\omega)-a(\pi(\mathbf{x}_1),\omega))^T(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2))}_{\le\,0}\\&+\mu\underbrace{(\pi(\mathbf{x}_2)-\mu\pi(\mathbf{x}_1))^T(\pi(\mathbf{x}_1)-\pi(\mathbf{x}_2))}_{\le\,0}\end{aligned}$$

$$\geq \mu\|\pi(\mathbf{x}_1) - \pi(\mathbf{x}_2)\|^2.$$

This implies that

$$\begin{aligned}\mu\|\pi(\mathbf{x}_1) - \pi(\mathbf{x}_2)\|^2 &\leq (T_\omega \mathbf{x}_1 - T_\omega \mathbf{x}_2)^T(\pi(\mathbf{x}_1) - \pi(\mathbf{x}_2)) \\ &\leq \|T_\omega\|\|\mathbf{x}_1 - \mathbf{x}_2\|\|\pi(\mathbf{x}_1) - \pi(\mathbf{x}_2)\| \\ \implies \|\pi(\mathbf{x}_1) - \pi(\mathbf{x}_2)\| &\leq \frac{\|T_\omega\|}{\mu}\|\mathbf{x}_1 - \mathbf{x}_2\|.\end{aligned}$$

Finally, we note that $\nabla_x \mathcal{Q}_\mu(\mathbf{x}, \omega) = -T_\omega^T \pi_\omega$, where π_ω is a maximizer of (D-$\mathcal{Q}(\mathbf{x}, \omega)$), implying that

$$\begin{aligned}\|\nabla_{\mathbf{x}} \mathcal{Q}_\mu(\mathbf{x}_1, \omega) - \nabla_{\mathbf{x}} \mathcal{Q}_\mu(\mathbf{x}_2, \omega)\| &\leq \|T_\omega\|\|\pi(\mathbf{x}_1) - \pi(\mathbf{x}_2)\| \\ &\leq \frac{\|T_\omega\|^2}{\mu}\|\mathbf{x}_1 - \mathbf{x}_2\|.\end{aligned}$$

(ii) We begin by noting that $\mathcal{Q}(\mathbf{x}, \omega) = (q^\omega - T_\omega \mathbf{x})^T \pi - \bar{d}^*(W^T\pi, \omega)$ where π is a maximizer of (D-$\mathcal{Q}(\mathbf{x}, \omega)$) while $\mathcal{Q}_\mu(\mathbf{x}, \omega) = (q^\omega - T_\omega \mathbf{x})^T \pi_\mu - \bar{d}^*(W^T\pi_\mu, \omega) - \frac{1}{2}\|\pi_\mu\|^2$ where π_μ is a maximizer of (D-$\mathcal{Q}_\mu(\mathbf{x}, \omega)$). Consequently, we have that

$$\begin{aligned}\mathcal{Q}(\mathbf{x}, \omega) &= (q^\omega - T_\omega \mathbf{x})^T \pi - \bar{d}^*(W^T\pi, \omega) \\ &\geq (q^\omega - T_\omega \mathbf{x})^T \pi_\mu - \bar{d}^*(W^T\pi_\mu, \omega) \\ &\geq (q^\omega - T_\omega \mathbf{x})^T \pi_\mu - \bar{d}^*(W^T\pi_\mu, \omega) - \tfrac{1}{2}\mu\|\pi_\mu\|^2 \\ &= \mathcal{Q}_\mu(\mathbf{x}, \omega).\end{aligned}$$

In addition, it is easily seen that

$$\begin{aligned}\mathcal{Q}_\mu(\mathbf{x}, \omega) &= (q^\omega - T_\omega \mathbf{x})^T \pi_\mu - \bar{d}^*(W^T\pi_\mu, \omega) - \tfrac{1}{2}\mu\|\pi_\mu\|^2 \\ &\geq (q^\omega - T_\omega \mathbf{x})^T \pi - \bar{d}^*(W^T\pi, \omega) - \tfrac{1}{2}\mu\|\pi\|^2 = \mathcal{Q}(\mathbf{x}, \omega) - \tfrac{1}{2}\mu\|\pi\|^2.\end{aligned}$$

As a result, we have that

$$\mathcal{Q}(\mathbf{x}, \omega) \leq \mathcal{Q}_\mu(\mathbf{x}, \omega) + \tfrac{1}{2}\mu\|\pi\|^2 \leq \mathcal{Q}_\mu(\mathbf{x}, \omega) + \mu\beta(\omega),$$

where $\|\pi\|^2 \leq \beta(\omega)$ for all π, where the boundedness of π follows from the Slater regularity condition on (P-$\mathcal{Q}(\mathbf{x}, \omega)$). □

3.5.2 Smoothing the max. function

From [97], recall that the smoothing of the max. function, defined as $t(u) \triangleq [u]^+$, is given by $t_\mu(x) \triangleq \mu(\log(e^{\frac{u}{\mu}} + 1) - \log(2))$ and t_μ is $(1, \log(2))$-smoothable. We prove the relatively simple result that $t'_\mu(u) \leq \bar{t}$ for all u and for any $\mu > 0$.

Lemma 18. *Consider the function $t_\mu(x) \triangleq \mu(\log(e^{\frac{u}{\mu}} + 1) - \log(2))$. Then for any $u \in \mathbb{R}$ and any $\mu > 0$, we have that $t'_\mu(u) \leq 1$.*

Proof. It can be seen for that any u and any $\mu > 0$,

$$t'_\mu(x) = \frac{\exp\left(\frac{u}{\mu}\right)}{\exp\left(\frac{u}{\mu}\right) + 1} \leq 1.$$

□

3.5.3 Smoothing a composition of two smoothable functions

We note that $[\mathcal{Q}(\mathbf{x},\omega)-m]^+$ denotes a composition of a nonsmooth function $h(w)$ where $h(w)=\max\{w,0\}$ with another nonsmooth function $t(z)$ where $t(z)=z_1-z_2$ and $z=(z_1,z_2)$. Our intent lies in showing that under if h is (α_1,β_1) smoothable and t is (α_2,β_2) smoothable, then $p=h(t)$ is (α_3,β_3) smoothable, where the smoothability of this composite function is defined as follows.

Definition 3. *Given two convex functions $h:\mathbb{R}^m\to\mathbb{R}$ and $g:\mathbb{R}^n\to\mathbb{R}^m$. Then the function $p(z)=h(t(z))$ is said to be (α,β) smoothable if the following two conditions hold and $p_\mu(z)\triangleq h_\mu(g_\mu(z))$.*
(i) There exists a constant α such that for any $\mu>0$,

$$\|\nabla_z p_\mu(z_1)-\nabla_z p_\mu(z_2)\|\le\tfrac{\alpha}{\mu}\|z_1-z_2\|,\quad\forall z_1,z_2.$$

(ii) There exists a constant β such that for any $\mu>0$,

$$p_\mu(z)\le p(z)\le p_\mu(z)+\beta\mu.$$

Under suitable conditions, we now prove that $p=h(t)$ is a smoothable function when h and t are smoothable.

Lemma 19. *Suppose $h:\mathbb{R}\to\mathbb{R}$ is a non-decreasing nonnegative convex function and is (α_1,β_1)-smoothable. In addition, if h_μ denotes an (α_1,β_1) smoothing of h, then h_μ is assumed to be a non-increasing and nonnegative function. Suppose $t:\mathbb{R}^n\to\mathbb{R}^m$ is (α_2,β_2)-smoothable convex function. In addition, suppose $\|\nabla_w h(w)\|\le C_1$ for all w and $\|\nabla_x t(x)\|\le C_2$ for all x. Then $p(z)=h(t(z))$ is (α,β)-smoothable.*

Proof. Since $\nabla_z(h_\mu(g_\mu(z))=h'_\mu(t_\mu(z))\nabla_z t_\mu(z)$, we have the following by adding and subtracting terms and invoking the triangle inequality.

$$\begin{aligned}\|\nabla h_\mu(t_\mu(z_1))-\nabla h_\mu(t_\mu(z_2)\| &= \|h'_\mu(t_\mu(x_1))\nabla_z t_\mu(z_1)-h'_\mu(t_\mu(z_2))\nabla_z t_\mu(z_2)\|\\ &\le \|h'_\mu(t_\mu(z_1))\nabla_z t_\mu(z_1)-h'_\mu(t_\mu(z_1))\nabla_z t_\mu(z_2)\|\\ &+\|h'_\mu(t_\mu(z_1))\nabla_z t_\mu(z_2)-h'_\mu(t_\mu(z_2))\nabla_z t_\mu(z_2)\|\\ &\le \|h'_\mu(t_\mu(x_1))\|\|\nabla_z t_\mu(z_1)-\nabla_z t_\mu(z_2)\|\\ &+\|\nabla_z t_\mu(x_2)\|\underbrace{\|h'_\mu(t_\mu(z_1))-h'_\mu(t_\mu(z_2))\|}_{\textbf{Term b}}.\end{aligned}\tag{64}$$

Since $h:\mathbb{R}\to\mathbb{R}$ is (α_1,β_1)-smoothable and $t:\mathbb{R}^n\to\mathbb{R}$ is (α_2,β_2)-smoothable, it follows that for all z_1,z_2,

$$\begin{aligned}\|h'_\mu(t_\mu(z_1))-h'_\mu(t_\mu(z_2))\|&\le\frac{\alpha_1}{\mu}\|t_\mu(z_1)-t_\mu(z_2)\|\\ \|\nabla t_\mu(z_1)-\nabla t_\mu(z_2)\|&\le\frac{\alpha_2}{\mu}\|z_1-z_2\|.\end{aligned}$$

Recall that by the differentiability of $t_\mu(z)$ and the mean-value theorem, for some $\gamma\in[0,1]$,

$$\begin{aligned}t_\mu(z_2)&=t_\mu(z_1)+\nabla_z t_\mu(z_1+\gamma(z_2-z_1))^T(z_2-z_1)\\ \Rightarrow\|t_\mu(z_2)-g_\mu(z_1)\|&\le\|\nabla_z t_\mu(z_1+\gamma(z_2-z_1))\|\|z_1-z_2\|\\ &\le C_2\|z_1-z_2\|.\qquad\text{(by assumption).}\end{aligned}\tag{65}$$

By (α_1, β_1)-smoothability of h and (65), Term b can be bounded as follows.

$$\textbf{Term b} \leq \tfrac{\alpha_2}{\mu}\|t_\mu(z_1) - t_\mu(z_2)\| \leq \frac{\alpha_2 C_2}{\mu}\|z_1 - z_2\|. \tag{66}$$

From (64), we have that for any z_1, z_2,

$$\begin{aligned}\|\nabla h_\mu t_\mu(z_1) - \nabla h_\mu(t_\mu(z_2)\| &\leq \frac{C_1\alpha_2}{\mu}\|z_1 - z_2\| + \frac{C_2\alpha_1}{\mu}\|z_1 - z_2\| \\ &\leq \tfrac{\alpha}{\mu}\|z_1 - z_2\|, \text{ where } \alpha \triangleq (C_1\alpha_2 + C_2\alpha_1).\end{aligned} \tag{67}$$

Since $h : \mathbb{R} \to \mathbb{R}$ is (α_1, β_1)-smoothable and $t : \mathbb{R}^n \to \mathbb{R}$ is (α_2, β_2)-smoothable, we have for any $z, \mu > 0$,

$$h_\mu(t(z)) \leq h(t(z)) \leq h_\mu(t(z)) + \beta_1\mu \tag{68}$$

$$t_\mu(z) \leq t(z) \leq t_\mu(z) + \beta_2\mu. \tag{69}$$

Since $h(\bullet)$ is a nondecreasing function,

$$\begin{aligned}
h_\mu(t_\mu(z)) \leq h_\mu(t(z)) &\leq h_\mu(t_\mu(z) + \beta_2\mu) && (h_\mu \text{ non-dec., (69)}) \\
h(t(z)) &\leq h(t_\mu(z) + \beta_2\mu) && ((69), h \text{ nondec.}) \\
h_\mu(t_\mu(z)) &\leq h_\mu(t(z)) \\
&\leq h(t(z)) && (\text{From (68)}) \\
&\leq h_\mu(t(z)) + \beta_1\mu && (\text{From (68)}) \\
&\leq h_\mu(t_\mu(z) + \beta_2\mu) + \beta_1\mu. && (\text{From (69)})
\end{aligned}$$

Since h_μ is a convex and positive function, for $\kappa \in [0, 1]$, we have the following by the mean-value theorem.

$$h_\mu(t_\mu(z) + \beta_2\mu) = h_\mu(t_\mu(z)) + h'_\mu(t_\mu(z) + \kappa\beta_2\mu)\beta_2\mu.$$

Since $h'_\mu(t_\mu(z) + \kappa\beta\mu) \leq C_1$, it follows that

$$h_\mu(t_\mu(z) + \beta_2\mu) \leq h_\mu(t_\mu(z)) + C_1\beta\mu,$$

implying that

$$h_\mu(t_\mu(z)) \leq h(t(z)) \leq h_\mu(t_\mu(x)) + \underbrace{(\beta_1 + C_1\beta_2)}_{\triangleq\beta}\mu. \tag{70}$$

From (67) and (70), $h(t(z))$ is (α, β)-smoothable. □

3.5.4 Smoothing $r(\mathbf{z}, \omega)$

Since $r(\mathbf{z}, \omega)$ is a consequence of a composition of a nonsmooth function (specifically the max function) on an affine translation of another nonsmooth function (specifically the recourse function $\mathcal{Q}(\bullet, \omega)$), we may utilize the results from the prior subsection. Specifically, let $t(\mathbf{z}, \omega) \triangleq \mathcal{Q}(\mathbf{x}, \omega) - m$, where $\mathbf{z} = (g, \theta, m)$. Furthermore, suppose $h(u) = [u]^+$.

Lemma 20. *Consider the functions $t(\mathbf{z},\omega) = \mathcal{Q}(\mathbf{x},\omega)$ and $h(u) = [u]^+$. Then the following hold.*
(i) The function $h_\mu(u) = \mu\log(e^{\frac{u}{\mu}}+1)$ represents a $(1,\log(2))$ smoothing of h and $0 \le h'(u) \le 1$ for all u.
(ii) The function $t(\mathbf{z},\omega) = \mathcal{Q}(\mathbf{x},\omega) - m$ is a convex (α_2,β_2)-smoothable function and $\|\nabla_{\mathbf{z}} t(\mathbf{z},\omega)\| \le C_2$ for all ω.

Proof. (i) Follows immediately from [97] and Lemma 18.

(ii) Since $\mathcal{Q}_\mu(\mathbf{x},\omega)$ is $(\alpha_2(\omega),\beta_2(\omega))$-smoothable, we have that $\mathcal{Q}_\mu(\mathbf{x},\omega) - m$ satisfies the following for any $\mathbf{x}$.

$$\mathcal{Q}_\mu(\mathbf{x},\omega) - m \le \mathcal{Q}(\mathbf{x},\omega) - m \le \mathcal{Q}_\mu(\mathbf{x},\omega) - m + \beta_2(\omega)\mu. \tag{71}$$

In addition, we have that

$$\begin{aligned}
\|\nabla_{\mathbf{z}} r(\mathbf{z}_1,\omega) - \nabla_{\mathbf{z}} r(\mathbf{z}_2,\omega)\| &= \left\| \begin{pmatrix} \nabla_g(\mathcal{Q}_\mu(\mathbf{x}_1,\omega)-m) - \nabla_g(\mathcal{Q}_\mu(\mathbf{x}_2,\omega)-m) \\ \nabla_\theta(\mathcal{Q}_\mu(\mathbf{x}_1,\omega)-m) - \nabla_\theta(\mathcal{Q}_\mu(\mathbf{x}_2,\omega)-m) \\ \nabla_m(\mathcal{Q}_\mu(\mathbf{x}_1,\omega)-m) - \nabla_m(\mathcal{Q}_\mu(\mathbf{x}_2,\omega)-m) \end{pmatrix} \right\| \\
&= \left\| \begin{pmatrix} \nabla_g \mathcal{Q}_\mu(\mathbf{x}_1,\omega) - \nabla_g \mathcal{Q}_\mu(\mathbf{x}_2,\omega) \\ 0 \\ 0 \end{pmatrix} \right\| \\
&\le \frac{\alpha_2(\omega)}{\mu}\|g_1 - g_2\| \le \frac{\alpha_2(\omega)}{\mu}\|\mathbf{z}_1 - \mathbf{z}_2\|.
\end{aligned}$$

Finally, it is relatively easy to see that

$$\|\nabla_{\mathbf{z}} t(\mathbf{z},\omega)\| = \left\| \begin{pmatrix} \nabla_{\mathbf{x}} \mathcal{Q}_\mu(\mathbf{x},\omega) \\ -1 \end{pmatrix} \right\| \le \|\nabla_{\mathbf{x}} \mathcal{Q}_\mu(\mathbf{x},\omega)\| + 1 \le C(\omega) + 1,$$

for all $\mathbf{x}$ where the last inequality follows from observing that

$$\|\nabla_{\mathbf{x}} \mathcal{Q}_\mu(\mathbf{x},\omega)\| = \| - T_\omega^T \pi_\omega\| \le \|T_\omega\| \left(\frac{\bar{d}}{\bar{c}} + \bar{\mu}\left(\frac{\bar{d}}{\bar{c}}\right)^2 \right).$$

□

We may now claim the smoothability of $r(\mathbf{z},\omega)$.

Proposition 5. *Suppose Assumption 8 holds. Consider function $r(\mathbf{z},\omega)$ defined in (55). Then $r(\mathbf{z},\omega)$ is a convex and (α,β)-smoothable function.*

Proof. From Lemma 20, we have that $h(u)$,$t(\mathbf{z},\omega)$ satisfy the requirement of Lemma 19. Following Lemma 19 we have that $r(\mathbf{z},\omega) = h(t(\mathbf{z},\omega))$ is (α,β)-smoothable. □

We may then define the smoothed approximation of $r(\mathbf{z},\omega)$ as follows.

$$r_\mu(\mathbf{z},\omega) \triangleq m + \frac{\mu\log\left(\exp\left(\frac{\mathcal{Q}_\mu(\mathbf{x},\omega)-m}{\mu}\right)+1\right) - \mu\log(2)}{1-\tau}. \tag{72}$$

As a consequence, we have that

$$\nabla_{\mathbf{z}} r_\mu(\mathbf{z},\omega) = \begin{pmatrix} \frac{1}{1-\tau}\left(\frac{e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega)-m}{\mu}}}{e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega)-m}{\mu}}+1}\right)\nabla\mathcal{Q}_\mu(\mathbf{x},\omega) \\ 0 \\ 1-\frac{1}{1-\tau}\left(\frac{e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega)-m}{\mu}}}{e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega)-m}{\mu}}+1}\right) \end{pmatrix}. \tag{73}$$

We know $\mathcal{Q}_\mu(\mathbf{x},\omega)$ is (α,β)-smooth approximation of $\mathcal{Q}(\mathbf{x},\omega)$. Thus, $r_\mu(\mathbf{z},m)$ is a smooth approximation of $r(\mathbf{z},m)$ with its gradients Lipschitz constant μ.

3.6 A variance-reduced smoothed accelerated scheme for two-stage risk-averse problems

While the prior section has analyzed the smoothing of the risk-adjusted recourse function, in this section, we utilize a variance-reduced smoothed accelerated scheme for such a class of problems. In Section 3.6.1, we provide a brief review of decomposition and Monte-Carlo sampling techniques for resolving two-stage stochastic convex problems, possibly complicated by risk-aversion. In Section 3.6.5, we introduce a recently developed variance-reduced smoothed accelerated scheme and show how it may be extended to contend with risk-averse regimes. Finally, in Section 3.6.6, we review the convergence statements inherited from this scheme.

3.6.1 A review of Monte-Carlo sampling schemes for 2-stage programs

Traditionally, schemes for resolving two-stage stochastic programs have differed based on whether the sample-space of the second-stage problem is finite or infinite. In the case of the former, decomposition techniques have proven useful in developing techniques that scale with the cardinality of Ω. Amongst the earliest of these was the L-shaped method [99] while augmented Lagrangian [100] and splitting methods [101] have also been utilized. A more comprehensive review of decomposition schemes can be found in [102]. When the sample-space is infinite, these avenues cannot be adopted and one has to resort to Monte-Carlo sampling schemes. We review three avenues for resolving such problems.

3.6.2 Stochastic cutting plane methods

Stochastic decomposition (SD) techniques decompose the stochastic elements of a problem from deterministic data, combining successive approximation methods from mathematical programming with sampling approaches. Unlike other sampling methods, SD leverages the special structure of linear programming problems. When the second-stage problems are linear, this implies that the second-stage recourse function is a piece-wise linear function. Cutting-plane techniques originate from the work by Kelley [103] in which the following algorithm was proposed to solve the following convex problem.

$$\min_{x\in X} \quad c^T x + \mathcal{Q}(x), \tag{74}$$

where $\mathcal{Q}(\cdot)$ is a convex function and X is a compact, convex, and nonempty set. The basic idea of cutting plane algorithm is as below: Such avenues have been extended to accommodate two-stage stochastic linear programs by Van Slyke and Wets [99] but only allow for finite sample-spaces . To accommodate general sample-spaces, the stochastic decomposition (SD) scheme was proposed by

Algorithm 3 Cutting-plane scheme

initialization: $x^1, k = 0, \ell_0(x) = -\infty, u_0 = c^T x^1 + f(x^1)$ and $l_0 = -\infty$;
while $u_k - l_k > 0$ **do**
    $k = k + 1$. Find (α^k, β^k) such that

$$\begin{aligned} \mathcal{Q}(x^k) &= \alpha_k + \beta_k x^k \\ \mathcal{Q}(x) &\geq \alpha_k + \beta_k x \quad \forall x \in X \end{aligned}$$

    Update $u_k = \min\{u_{k-1}, c^T x^k + \mathcal{Q}(x^k)\}, \ell_k(x) = \max\{\ell_{k-1}(x), \alpha_k + \beta_k x\}$.
    Update $l_k = \min_{x \in X}\{c^T x + \ell_k(x)\}$ where x_{k+1} is a solution to (74).
end while
$x^* = x_{k+1}$
return x^*

Higle and Sen [104] in 1991. Consider the problem (74) where $\mathcal{Q}(x) \triangleq \mathbb{E}[\mathcal{Q}(x, \omega)]$, where $\mathcal{Q}(x, \omega)$ is an optimal value of

$$\max_{\pi} \ \left\{ (h_\omega - T_\omega x)^T \pi \mid W^T \pi \leq q_\omega \right\}. \tag{LP-S2D}$$

This scheme approximates the recourse function $\mathbb{E}\left[Q(x, \omega)\right]$ through a sequence of piecewise linear approximations. Within any given major iteration, each piece of the piecewise linear approximation is derived from a conditionally independent set of observations. As part of the scheme [104], a set V_k is constructed by solving one subproblem per iteration and dual vector obtained is added to this set. Formally, the update of V_k is defined as follows.

$$V_k := V_{k-1} \cup \pi^k_{\omega_k},$$

where $\pi^k_{\omega^k}$ is a solution to the following problem:

$$\begin{aligned} \max \quad & [h_{\omega^k} - T_{\omega^k} x^k]^T \pi \\ \text{subject to} \quad & W^T \pi \leq q_{\omega^k}. \end{aligned}$$

Akin to the cutting plane scheme, we may obtain the piecewise linear outer-approximation $\eta_k(x)$ for the recourse function with the following form:

$$\eta_k(x) := \max\{\alpha^k_t + \beta^k_t x \ \mid \ t = 1, \ldots, k\}.$$

The stochastic decomposition (SD) algorithm is formally defined in Algorithm 5 and further details can be found from [104].

3.6.3 Sample-average approximation

In sample-average approximation theory (also referred to as exterior sampling), samples are generated outside of an optimization procedure. Consequently, the resulting sample average approximation (SAA) problems are solved by deterministic optimization algorithms. One of the advantages of SAA is that this method separates sampling procedures and optimization techniques. Consider the following stochastic programming problem:

$$\min_{x \in X} \quad f(x), \text{ where } f(x) \triangleq \mathbb{E}\left[F(x, \xi(\omega))\right], \tag{75}$$

Algorithm 4 Stochastic Decomposition Algorithm

initialization: $k = 0, V_0 = \emptyset, \eta_0(x) = -\infty, x^1 \in X$ L is given;
while iteration $k < K_{max}$ **do**
    $k = k + 1$. Randomly generate an observation of ω, ω^k, independent of any previously generated observations;
    Solve subproblem to get solution $\pi^k_{\omega^k}$

$$\begin{aligned} \max \quad & [h_{\omega^k} - T_{\omega^k} x^k]^T \pi \\ \text{subject to} \quad & W^T \pi \leq q_{\omega^k}. \end{aligned}$$

    Update $V_k = V_{k-1} \cup \pi^k_{\omega_k}$
    Determine the coefficients of the k^{th} piecewise linear approximation of recourse function (α^k, β^k) such that

$$\alpha^k_k + \beta^k_k x = \frac{1}{k} \sum_{t=1}^{k} \pi^k_t (h_{\omega^t} - T_{\omega^t} x)$$

    where $\pi^k_t \in \arg\max\{\pi^T (h_{\omega_t} - T_{\omega_t} x^k) | \pi \in V_k\}$
    Update the coefficients of all previously generated cuts.

$$\alpha^k_t = \frac{k-1}{k} \alpha^{k-1}_t + \frac{1}{k} L \quad , \quad \beta^k_t = \frac{k-1}{k} \beta^{k-1}_t.$$

    Update $\eta_k(x) = \max\{\alpha^k_t + \beta^k_t x \mid t = 1, \ldots, k\}$.
    Solve $\min_{x \in X}\{c^T x + \eta_k(x)\}$ where x_{k+1} is the solution to this problem.
end while
$x^* = x_{k+1}$
return x^*

$X \subseteq \mathbb{R}^n$ is a closed and convex set, $\xi : \Omega \to \mathbb{R}^d$ is a random vector, and the associated probability space is denoted by $(\Omega, \mathcal{F}, \mathbb{P})$. Unless stated otherwise, the expectation is assumed to be well-defined and finite valued for all $x \in X$, which implies for every $x \in X$ the value of $F(x, \omega)$ for every $\omega \in \Omega$ is finite. Suppose we have a sample $\omega^1, \ldots, \omega^N$ of N realizations of the random vector ω. This sample is generated by Monte Carlo sampling and for any $x \in X$, we estimate the expected value $f(x)$ by the sample-average $f_N(x) \triangleq \frac{1}{N} \sum_{j=1}^{N} F(x, \omega^j)$ by averaging values $F(x, \omega^j), j = 1, \ldots, N$. The resulting sample average approximation (SAA) of the true problem is defined as follows.

$$\min_{x \in X} \hat{f}_N(x), \text{ where } \hat{f}_N(x) \triangleq \frac{1}{N} \sum_{j=1}^{N} F(x, \omega^j).$$

Note that $\hat{f}_N(x)$ can be viewed as the expectation taken with respect to the empirical measure associated with a probability mass function $\{\frac{1}{N}, \ldots, \frac{1}{N}\}$. By the law of large numbers (LLN), under suitable regularity conditions $\hat{f}_N(x)$ converges to $f(x)$ pointwise with probability one as $N \to \infty$. Moreover, by the classical LLN, this convergence holds if the sample is independent and identically distributed. Much of the research on SAA theory focuses on proving that the estimator for the optimal value converges to the true value as $N \to \infty$ with probability one. Related statements can be developed for the solution set. In addition, rates of convergence can also be derived for such schemes.

Consistency of SAA estimators was investigated by tools of epi-convergence analysis by King and Wets [105] and Robinson [106] while asymptotic of SAA estimators of optimal solutions of stochastic programs were discussed by King and Rockefeller [107] and Shapiro [108]. A detailed exposition of recent theoretical findings can be found in [98]. It is worth emphasizing that this avenue is not an algorithm in the conventional sense but represents an avenue for approximation.

3.6.4 Stochastic approximation methods

Stochastic approximation schemes originate from the seminal paper by Robbins and Monro [109] while asymptotics can be found in the research by Kushner and Clark [110] and Nevelson and Hasminskii [111]. Longer step averaging schemes was developed in Polyak [112] and these ideas were presented in a different form by Nemirovski and Yudin [113]. Stochastic quasi-gradient techniques are closely related to stochastic approximation and early work focused on the solution of two-stage stochastic linear programs [114]. This avenue saw significant subsequent study by Gaivoronski [115], Wets [116], amongst others. For a given stochastic convex optimization (75) where f is a differentiable function, given an $x_0 \in X$, a standard SA scheme would be based on the following update rule:

$$x_{k+1} := \Pi_X \left(x_k - \eta_k (\nabla_x f(x_k) + w_k) \right), \ k \geq 0$$

where $w_k := \nabla_x f(x_k; \omega_k) - \nabla_x f(x_k)$ and $\nabla_x f(x, \xi_\omega)$ is referred to as $\nabla_x f(x, \omega)$. An variable sample-size stochastic approximation scheme (VSSA) was proposed in [117, 118] in which the sequence $\{x_{k+1}\}$ would have the following update rule:

$$x_{k+1} := \Pi_X \left(x_k - \eta_k \frac{\sum_{j=1}^{N_k} \nabla_x f(x_k, \xi_{j,k})}{N_k} \right), \ k \geq 0.$$

In such a scheme, an increasingly unbiased estimate of the gradient is employed, leading to improved iteration complexity of the scheme. Following the idea in [117–119], we introduce a variant of this scheme for solving two-stage stochastic programs in the next section.

Specifically, in this chapter, we revisit stochastic quasigradient methods which has traditionally been plagued by the same challenges as stochastic approximation. In particular, the convergence rate was $\mathcal{O}(\frac{1}{\sqrt{k}})$ and the empirical behavior varies significantly with the choice of step length sequence. In [120], we introduce three key modifications to the standard stochastic approximation framework by (i) utilizing a two-step accelerated scheme, (ii) incorporating a smoothing of the recourse function by regularizing the second-stage dual; (iii) and leveraging variance reduction. We develop a foundation to allow for applying this framework to risk-averse two-stage problems which allows for recovering the optimal rate of $\mathcal{O}(1/k)$. Next, we describe this scheme.

3.6.5 Variance-reduced smoothed accelerated scheme

We now introduce the variable sample-size accelerated proximal method (VS-APM) first presented in [120] and apply it to (r-ED).

$$\min_{\mathbf{z}\in Z} \ \mathbb{E}[h(\mathbf{z},\omega)], \ \text{where } h(\mathbf{z},\omega) \triangleq (c(\mathbf{z}) + r(\mathbf{z},\omega)). \tag{r-ED}$$

This framework incorporates three aspects in extending standard stochastic approximation schemes.

(i) *Smoothing.* The first change from standard stochastic approximation schemes lies in utilizing the gradient of a smoothed objective, where the smoothing parameter sequence is driven to zero. The resulting scheme can be formalized as follows, given a $z_0 \in Z$.

$$\mathbf{z}_{k+1} := \Pi_Z \left[z_k - \gamma_k (\nabla_{\mathbf{z}} h_{\mu_k}(\mathbf{z}_k) + w_k) \right], \qquad k \geq 0. \tag{76}$$

In contrast with standard stochastic approximation, we employ the sampled gradient $\nabla_{\mathbf{z}} h_{\mu_k}(\mathbf{z}_k) + w_k$ where $\nabla_{\mathbf{z}} h_{\mu_k}(\mathbf{z}_k) + w_k = \nabla_{\mathbf{z}} c(\mathbf{z}_k) + \nabla_{\mathbf{z}} r_{\mu_k}(\mathbf{z}_k, \omega_k)$.

(ii) *Variance-reduction.* In traditional stochastic approximation schemes, a single sample $\nabla_{\mathbf{z}} r_{\mu_k}(\mathbf{z}_k, \omega_k)$ or a fixed batch-size of samples is utilized. However, such avenues lead to biased gradients (where the conditional bias does not diminish to zero). Instead, we propose a variance-reduced scheme given by the following.

$$\mathbf{z}_{k+1} := \Pi_Z \left[z_k - \gamma_k (\nabla_{\mathbf{z}} h_{\mu_k}(\mathbf{z}_k) + \bar{w}_k) \right], \qquad k \geq 0 \tag{77}$$

where $\nabla_{\mathbf{z}} h_{\mu_k}(\mathbf{z}_k) + \bar{w}_k = \frac{\sum_{j=1}^{N_k} \nabla_{\mathbf{z}} c(\mathbf{z}_k) + \nabla_{\mathbf{z}} r_{\mu_k}(\mathbf{z}_k, \omega_{j,k}):}{N_k}$. In fact, the conditional bias of the gradients diminishes to zero and this scheme starts mimicking an inexact gradient scheme.

(iii) *Acceleration.* Finally, we introduce an accelerated scheme by utilizing the following two-step rule.

$$\zeta_{k+1} := \Pi_Z \left[z_k - \gamma_k (\nabla_{\mathbf{z}} h_{\mu_k}(\mathbf{z}_k) + \bar{w}_k) \right], \qquad k \geq 0 \tag{78}$$

$$\mathbf{z}_{k+1} := (1 + \alpha_k)\zeta_{k+1} - \alpha_k \zeta_k, \qquad k \geq 0. \tag{79}$$

Note that α_k are prescribed sequences and this avenue was first suggested for solving convex programs with differentiable objectives by Nesterov [**?**]. The resulting accelerated scheme improved the convergence rate from $\mathcal{O}(1/k)$ to $\mathcal{O}(1/k^2)$. Similar benefits are expected to accrue here when step length sequences, smoothing sequences, and sample-size sequences are chosen appropriately. Collectively, this scheme is referred to as a variable sample-size accelerated proximal scheme (VSAPM) [120].

Algorithm 5 VS-APM for two-stage risk-based ED

initialization: $\lambda_1 = 1, \gamma_0, \mathbf{y}_0 = \mathbf{z}_1, M_0 = 0, N_k = 1, k = 1$;
while $k < K$ **do**
  Generate N_k samples and compute $\nabla_{\mu_k} h(\mathbf{z}_k, \omega_{1,k}), \ldots, \nabla_{\mu_k} h(\mathbf{z}_k, \omega_{N_k,k})$.
  Update

$$\mathbf{y}_{k+1} := \Pi_Z \left[\mathbf{x}_k - \eta_k \frac{\sum_{j=1}^{N_k} \nabla_{\mu_k} h(\mathbf{z}_k, \omega_{j,k})}{N_k} \right]. \tag{80}$$

  $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$; $\gamma_k := \frac{\lambda_k - 1}{\lambda_{k+1}}$,
  Update

$$\mathbf{z}_{k+1} := (1 + \gamma_k)\mathbf{y}_{k+1} - \gamma_k \mathbf{y}_k). \tag{81}$$

  Update $k := k + 1$, N_k and μ_k.
end while
return $\mathbf{z}_K$.

The scheme is formally stated in Algorithm 5. It may be recalled that

$$r_\mu(\mathbf{z}, \omega) = m + \frac{1}{1-\tau} \mu \left[\log \left(e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega) - m}{\mu}} + 1 \right) \right], \tag{82}$$

where $\mathbf{z} = (\mathbf{z}, m)$. Furthermore, the gradient $\nabla_{\mathbf{z}} r_\mu(\mathbf{z}, \omega)$ is defined as follows.

$$\begin{aligned} \nabla_{\mathbf{x}} r_\mu(\mathbf{z}, \omega) &= \left(\frac{1}{1-\tau} \right) \left(\frac{e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega) - m}{\mu}}}{e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega) - m}{\mu}} + 1} \right) \nabla \mathcal{Q}_\mu(\mathbf{x}, \omega), \\ \nabla_m r_\mu(\mathbf{z}, \omega) &= 1 - \left(\frac{1}{1-\tau} \right) \left(\frac{e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega) - m}{\mu}}}{e^{\frac{\mathcal{Q}_\mu(\mathbf{x},\omega) - m}{\mu}} + 1} \right), \end{aligned} \tag{83}$$

where $\nabla_{\mathbf{x}} \mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_k) = -T(\omega_k)^T \pi(\mathbf{x}_k, \omega_k)$ and $\pi(\mathbf{x}_k, \omega_k)$ is a solution of the smoothed second-stage dual problem (D-$\mathcal{Q}_\mu(\mathbf{x}, \omega_{j,k})$).

3.6.6 Convergence theory

We now recall the two main assumptions for claiming convergence from (VS-APM) from [120]. Of these, the first requires that the objective of the original problem is indeed smoothable and the distance of the initial iterate to an optimal solution can be bounded.

Assumption 9. *(i) The function $\mathbb{E}[r(\mathbf{z}, \omega)]$ is (α, β) smoothable; (ii) There exists a C such that $\|\mathbf{z}_1 - \mathbf{z}^*\|^2$ where $\mathbf{z}^*$ is a solution to the original problem.*

Next, we require that the noise sequence $\bar{w}_k$, defined as follows.

$$\bar{w}_k \triangleq \nabla_{\mathbf{z}} \widehat{r}_{\mu_k}(\mathbf{z}_k) - \nabla_{\mathbf{z}} \bar{r}_{\mu_k}(\mathbf{z}_k), \text{ where } \nabla_{\mathbf{z}} \widehat{r}_{\mu_k}(\mathbf{z}_k) \triangleq \frac{\sum_{j=1}^{N_k} \nabla_{\mathbf{z}} r_{\mu_k}(\mathbf{z}_k, \omega_{j,k})}{N_k}. \tag{84}$$

Assumption 10. *Consider the sequence* $\{\bar{w}_k\}$ *where* $\bar{w}_k$ *is defined as* (84). *Then there exists a scalar* $\nu > 0$, *such that* $\mathbb{E}[\|\bar{w}_k\|^2 \mid \mathcal{F}_k] \leq \frac{\nu^2}{N_k}$ *and* $\mathbb{E}[\bar{w}_k \mid \mathcal{F}_k] = 0$ *holds almost surely for all* k, *where* $\mathcal{F}_k \triangleq \sigma\{\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_{k-1}\}$.

We may now formally state the convergence statement from [120].

Proposition 6. *Consider the sequence* $\{\mathbf{z}_k\}$ *generated from (sVS-APM) where* $\mu_k = 1/k$, $\eta_k = 1/2k$, *and* $N_k = \lfloor k^a \rfloor$, *where* $a > 1$. *Suppose Assumption 9 and 10 hold. Then the following hold.*

(i) If $\bar{C} \triangleq \frac{2\nu^2 a}{a-1} + 4C^2 + B^2$, *then the following holds for* $K \geq 1$.

$$\mathbb{E}[h(\mathbf{z}_k, \omega_k)] - \mathbb{E}[h(\mathbf{z}^*, \omega)] \leq \frac{\bar{C}}{K}. \tag{85}$$

(ii) Let $\epsilon \leq \bar{C}/2$ *and* K *is such that* $\mathbb{E}[h(\mathbf{z}_k, \omega_k)] - \mathbb{E}[h(\mathbf{z}^*, \omega)] \leq \epsilon$. *Then* $\sum_{k=1}^{K} N_k \leq \mathcal{O}(\frac{1}{\epsilon^{1+a}})$.

We now provide some results that allows us to claim that such Prop. 6 can be invoked.

Lemma 21. *Consider the noise sequence* $\bar{w}_k$ *defined in* (84). *Then this sequence satisfies the following: (i)* $\mathbb{E}[\bar{w}_k \mid \mathcal{F}_k] = 0$ *a.s. for every* $k \geq 1$; *(ii)* $\mathbb{E}[\bar{w}_k \mid \mathcal{F}_k] \leq \frac{\nu^2}{N_k}$ *a.s. for every* $k \geq 1$.

Proof. (i) Recall that $r_\mu(\mathbf{z}) = \mathbb{E}[r_\mu(\mathbf{z}, \omega)]$. Since $r_\mu(\mathbf{z}, \omega)$ is a continuously differentiable convex function in $\mathbf{z}$ for every ω, it follows that we may interchange derivatives and expectations in claiming that $\nabla_{\mathbf{z}} r_\mu(\mathbf{z}) = \mathbb{E}[\nabla_{\mathbf{z}} r_\mu(\mathbf{z}, \omega)]$ (cf. [98, Theorem 7.44]). Consequently, if $\mathbf{z}_k$ is adapted to $\mathcal{F}_k$ and $\bar{w}_k$ is defined as (84), it follows that $\mathbb{E}[\bar{w}_k \mid \mathcal{F}_k] = 0$ in an a.s. sense.
(ii) Next, we note that $\bar{w}_k$ is a sample-average of a set of i.i.d random variables with mean zero. Consequently, if $\mathbb{E}[\|w_i\|^2 \mid \mathcal{F}_k] \leq \nu^2$ for $i = 1, \ldots, N_k$ in an a.s. fashion, it follows that $\mathbb{E}[\|\bar{w}_k\|^2 \mid \mathcal{F}_k] \leq \frac{\nu^2}{N_k}$ in a.s. sense. It remains to show that $\mathbb{E}[\|w_i\|^2 \mid \mathcal{F}_k] \leq \nu^2$ a.s. .

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla_{\mathbf{z}} r_{\mu_k}(\mathbf{z}_k, \omega_k) - \mathbb{E}[\nabla_{\mathbf{z}} r_{\mu_k}(\mathbf{z}_k, \omega)]\|^2\right] \leq \mathbb{E}\left[\|\nabla_{\mathbf{z}} r_{\mu_k}(\mathbf{z}_k, \omega_k)\|^2\right] \\
&= \mathbb{E}\left[\|\nabla_{\mathbf{x}} r_{\mu_k}(\mathbf{z}_k, \omega_k)\|^2\right] + \mathbb{E}\left[\|\nabla_m r_{\mu_k}(\mathbf{z}_k, \omega_k)\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{1-\tau}\left(\frac{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k}) - m_k}{\mu_k}} \nabla \mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k})}{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k}) - m_k}{\mu_k}} + 1}\right)\right\|^2\right] \\
&+ \mathbb{E}\left[\left\|\left(1 - \frac{1}{1-\tau}\left(\frac{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k}) - m_k}{\mu_k}}}{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k}) - m_k}{\mu_k}} + 1}\right)\right)\right\|^2\right].
\end{aligned}$$

We observe that the first term can be bounded as follows.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{1-\tau}\left(\frac{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k}) - m_k}{\mu_k}} \nabla \mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k})}{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k}) - m_k}{\mu_k}} + 1}\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{(1-\tau)}\right)^2 \left\|\left(\frac{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k}) - m_k}{\mu_k}} \nabla \mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k})}{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k, \omega_{j,k}) - m_k}{\mu_k}} + 1}\right)\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left(\tfrac{1}{(1-\tau)}\right)^2 \left\| \frac{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k,\omega_{j,k})-m_k}{\mu_k}}}{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k,\omega_{j,k})-m_k}{\mu_k}}+1} \nabla \mathcal{Q}_{\mu_k}(\mathbf{x}_k,\omega_{j,k}) \right\|^2\right] \\
&\leq \mathbb{E}\left[\left(\tfrac{1}{(1-\tau)}\right)^2 \|\nabla \mathcal{Q}_{\mu_k}(\mathbf{x}_k,\omega_{j,k})\|^2\right] \\
&= \mathbb{E}\left[\left(\tfrac{1}{(1-\tau)}\right)^2 \| - T^T_{\omega_{j,k}} \pi_{\omega_{j,k}}\|^2\right] \\
&\leq \mathbb{E}\left[\left(\tfrac{\bar{\pi}}{(1-\tau)}\right)^2 \|T_{\omega_{j,k}}\|^2\right] \leq \nu_1^2.
\end{aligned}
$$

The second term can be similarly bounded.

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \left(1 - \frac{1}{1-\tau}\left(\frac{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k,\omega_{j,k})-m_k}{\mu_k}}}{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k,\omega_{j,k})-m_k}{\mu_k}}+1}\right)\right)\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\| \left(1 - \frac{1}{1-\tau}\left(\frac{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k,\omega_{j,k})-m_k}{\mu_k}}}{e^{\frac{\mathcal{Q}_{\mu_k}(\mathbf{x}_k,\omega_{j,k})-m_k}{\mu_k}}+1}\right)\right)\right\|^2\right] \\
&\leq (2 + \frac{2}{(1-\tau)^2}) \triangleq \nu_2^2.
\end{aligned}
$$

It follows that $\mathbb{E}[\|w_i\|^2 \mid \mathcal{F}_k] \leq \nu^2 = \nu_1^2 + \nu_2^2$. □

3.7 Numerical Studies

In this section, we apply our scheme to the resolution of two-stage stochastic economic dispatch problems to two sets of problems. In Section 3.7.1, we review the model for generation of wind realizations and compare our scheme with the stochastic decomposition and stochastic quasi-gradient counterparts in Section 3.7.2 based on an IEEE 118-bus system which contains 19 generators, 35 synchronous condensers, 177 lines, 9 transformers, and 91 loads. The impact of risk is examined in Section 3.7.3. Finally, we conclude the section by examining the performance of this scheme on test problems sourced from ARPA-E's Grid Optimization competition.

3.7.1 Autoregressive Moving Average Model

In this subsection, we review the statistical model employed for developing wind forecasts and generating demand scenarios. A review of multi area wind speed and wind power scenario generation methods has been provided in [121]. It is worth noting that ARMA techniques have been used in developing stochastic optimization schemes for power system dispatch [122]. Our focus is on autoregressive moving average (ARMA) models and use data from ERCOT's hourly wind generation during 2009, 2010, and 2011 to test the resulting models. In an ARMA model [123], wind speed y_t in period t consists of the weighted sum of past observations and a weighted sum of independent shocks defined as follows:

$$
\text{ARMA}(p,q): \; y_t = \mu_0 + \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{k=1}^{q} \theta_k \epsilon_{t-k} + \epsilon_t.
$$

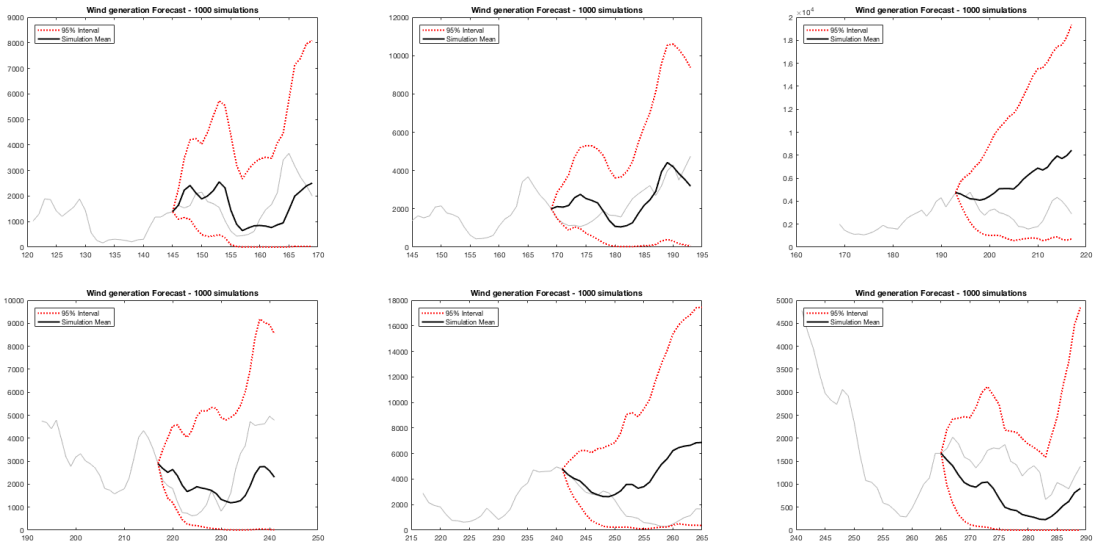

Figure 6: ARMA: Test data v.s. prediction

where $y_{t-1}, \ldots, y_{t-p}$ represent past observations (AR) while $\epsilon_{t-1}, \ldots, \epsilon_{t-q}$ are past innovations (MA). All of ϵ_ts are identical and independent centered Gaussian variables (white noise processes). By solving the Yule-Walker equations, the coefficients ϕ_js and θ_ks can be estimated. As wind speed over large geographical area is generally believed to follow a Weibull distribution [124], there is a need of a normalization transformation given by $y = N^{-1}\left[F(w)\right]$, where w denotes the time series representing the wind generation data, F denotes the cumulative distribution function (CDF) of the Weibull distribution associated with the stochastic process, and N denotes the standard normal CDF. Our preliminary tests are captured in Figure 6 where six sets of predictions are provided with 95% confidence intervals.

3.7.2 Performance comparison for stochastic economic dispatch

We now compare the performance of the proposed VS-APM scheme with stochastic decomposition and standard stochastic quasi-gradient on an IEEE 118-bus system with 19 generators, 35 synchronous condensers, 177 lines, 9 transformers, and 91 loads. All experiments were implemented in Matlab R2017a with `cplex` employed for solving LPs and QPs.
We begin by comparing VS-APM with stochastic decomposition in a setting where the simulation budget is 1000. In Table 9, we compare the behavior of (VS-APM) with SD on 10 problem instances. We observe that (VS-APM) takes less than 1% of the time taken by SD while producing similar objective values. This difference is because (SD) is contending with increasingly larger first-stage problems with (VS-APM) does not have this challenge. Note that the objective value is generated by re-sampling with 20 scenarios. A comparison with standard stochastic gradient provided in Table 10 reveals similar benefits in terms of computational time. In this instance, the key benefit lies in taking far less first-stage projection steps, a consequence of utilizing variance reduced schemes.

3.7.3 Risk-based Economic Dispatch

We now consider the risk-based model in settings where the variance of demand is raised from 10 to 50 in steps of 10 while wind penetration is raised from 10% to 30%. We compare the risk-neutral

$\|\Omega\|$	Iter_SD	SD_mean	Time_SD	Iter_VSAPM	VSAPM_mean	ObjDiff	Time_VSAPM	TimeDiff
1000	1000	1.24E+04	4.14E+03	136	1.22E+04	-1.46%	3.06E+01	0.74%
1000	1000	3.49E+04	3.84E+03	136	3.55E+04	1.83%	2.75E+01	0.71%
1000	1000	7.12E+04	3.90E+03	136	7.16E+04	0.61%	2.89E+01	0.74%
1000	1000	1.11E+04	4.10E+03	136	1.10E+04	-0.50%	2.95E+01	0.72%
1000	1000	3.39E+04	4.35E+03	136	3.42E+04	0.88%	2.78E+01	0.64%
1000	1000	7.15E+04	4.25E+03	136	6.99E+04	-2.22%	2.79E+01	0.66%
1000	1000	1.05E+04	3.83E+03	136	1.02E+04	-1.94%	2.88E+01	0.75%
1000	1000	3.32E+04	3.82E+03	136	3.32E+04	0.07%	2.78E+01	0.73%
1000	1000	6.83E+04	3.82E+03	136	6.83E+04	-0.06%	2.86E+01	0.75%
1000	1000	1.04E+04	3.83E+03	136	1.02E+04	-2.38%	2.82E+01	0.74%

Table 9: VS-APM and SD

Iter	SA_mean	Iter	V_mean	Diff	Iter	V_mean	Diff
1000	1.24E+04	65	1.22E+04	-1.54%	136	1.22E+04	-1.58%
1000	3.51E+04	65	3.64E+04	3.67%	136	3.55E+04	1.19%
1000	7.16E+04	65	7.36E+04	2.77%	136	7.16E+04	-0.05%
1000	1.12E+04	65	1.10E+04	-1.73%	136	1.10E+04	-1.73%
1000	3.39E+04	65	3.50E+04	3.19%	136	3.42E+04	0.98%
1000	6.99E+04	65	7.16E+04	2.48%	136	6.99E+04	-0.03%
1000	1.04E+04	65	1.02E+04	-1.61%	136	1.02E+04	-1.56%
1000	3.31E+04	65	3.39E+04	2.40%	136	3.32E+04	0.44%
1000	6.83E+04	65	7.00E+04	2.48%	136	6.83E+04	-0.03%
1000	1.03E+04	65	1.02E+04	-1.52%	136	1.02E+04	-1.49%
1000	3.27E+04	65	3.33E+04	1.76%	136	3.28E+04	0.30%
1000	6.69E+04	65	6.86E+04	2.45%	136	6.70E+04	0.09%
1000	1.21E+04	65	1.21E+04	-0.09%	136	1.21E+04	-0.54%
1000	3.26E+04	65	3.33E+04	2.04%	136	3.28E+04	0.63%

Table 10: VS-APM and SA

solution with the risk-averse solution in Table 11 where $\tau = 0.2$. It is observed that the conditional value of risk increases as variance in demand grows. In addition, we note that the CVaR associated with the risk-neutral solution (CVaR_Mean) is significantly higher than that with the risk-averse solution (CVAR_rED). In effect, solving a risk-neutral model leads to higher risk exposure. We also observe that the value of the stochastic solution (VSS_CVaR) increases as σ and wind penetration levels grow. We conduct further tests on IEEE test networks from MATPOWER and find that

omega_av	wind_per	sigma	fv_Mean	fv_sED	VSS	CVaR_Mean	CVaR_rED	VSS_CVaR
1000	0.2	1	5.68E+04	3.51E+04	2.18E+04	8.09E+04	5.31E+04	2.78E+04
1000	0.3	1	1.03E+05	7.08E+04	3.19E+04	1.45E+05	1.06E+05	3.86E+04
1000	0.4	1	1.58E+05	1.09E+05	4.84E+04	2.17E+05	1.63E+05	5.41E+04
1000	0.2	10	5.70E+04	3.54E+04	2.16E+04	8.45E+04	5.74E+04	2.71E+04
1000	0.3	10	1.03E+05	7.08E+04	3.20E+04	1.49E+05	1.08E+05	4.10E+04
1000	0.4	10	1.58E+05	1.09E+05	4.83E+04	2.20E+05	1.65E+05	5.46E+04
1000	0.2	20	5.77E+04	3.63E+04	2.14E+04	9.62E+04	7.05E+04	2.58E+04
1000	0.3	20	1.03E+05	7.09E+04	3.25E+04	1.61E+05	1.20E+05	4.03E+04
1000	0.4	20	1.58E+05	1.09E+05	4.84E+04	2.35E+05	1.77E+05	5.81E+04
1000	0.2	30	5.90E+04	3.75E+04	2.15E+04	1.14E+05	8.54E+04	2.86E+04
1000	0.3	30	1.04E+05	7.12E+04	3.31E+04	1.79E+05	1.35E+05	4.36E+04
1000	0.4	30	1.58E+05	1.09E+05	4.86E+04	2.52E+05	1.92E+05	6.04E+04
1000	0.2	50	6.33E+04	4.08E+04	2.24E+04	1.51E+05	1.18E+05	3.26E+04
1000	0.3	50	1.07E+05	7.28E+04	3.45E+04	2.20E+05	1.73E+05	4.68E+04
1000	0.4	50	1.60E+05	1.10E+05	4.94E+04	2.94E+05	2.28E+05	6.63E+04

Table 11: Value of Stochastic Solution

risk-neutral solutions lead to higher risk exposure than risk-averse solutions (see Table 12).

	Mean Cost	Two-stage Cost	Risk Cost	Mean worst 20%	Two-stage worst 20%	Risk worst 20%
IEEE118B	59533.26729	44793.66107	59530.76254	268179.7149	268179.1903	198998.0665
IEEE145	46916705.35	46851098.4	46853310.19	212284415.2	212077890.7	212070666.7
IEEE300A	47577079.24	39356294.75	39689745.57	64922699.66	56087603.98	55719991.89
IEEE300B	87018479.75	70139581.54	70877282.34	97323306.35	79622883.48	78862567.52
IEEE300C	86642730.74	69798253.99	70560254.99	89515610.86	73364330.13	72596519.46

Table 12: Comparison across different networks

We further examine the impact of variance for the IEEE 300 bus system where demand is assumed to follow a non-normal (beta) distribution. We note that the risk exposure grows as the variance increases and risk-averse models are able to better manage this exposure.

Mean Cost	Two Stage Cost	Risk Cost	Mean worst 20%	Two Stage worst 20%	Risk worst 20%	Variance
86644497.45	69773646.04	70537357.47	89044111.67	72892328.35	72123658.01	0.00507185
86886563.58	70095760.05	70834083.12	92983380.85	76200148.71	75475844.68	0.01984127
87601203.2	70675269.16	71397447.66	98226025.94	80316099.73	79545661.2	0.0375
90436594.28	73754429.32	74431960.72	111269137.4	93425043.78	92685828.32	0.06857143

Table 13: Changing variance on IEEE 300 system

3.7.4 Case study: ARPA-E Network

ARPA-E networks are networks that been put in use of Grid Optimization (GO) Competition. The goal is have a real-time matching of instantaneous electricity generation and demand, which requires utilities, grid operators, and other stakeholders to use a variety of sophisticated software operating across a wide range of timescales.

One test network among those was chosen and modified to be tested on. The numeric test is conducted on ARPA-E "Original Dataset Real-Time Network_01-10R". This network contains 500 buses, with 90 generators, 468 branches, 262 transformers and 371 contingencies with each contingency represents one generator failure or an branch or transformer failure. The original network is for ACOPF Some necessary modifications are made to conduct DCOPF experiment. DCOPF is an relaxation to original thus in order to tighten constraints to introduce recourse to second stage, loads are modified to double loads.

Mean Cost	Risk Cost	Two-stage Cost	Mean Risk	Risk Risk	Two-stage Risk	Samples
187878.701	187878.701	187878.701	245966.5256	245966.5256	245966.5256	0
187878.701	176647.6331	173831.485	245966.5256	200140.8446	205645.2917	7967
187878.701	176645.5224	173831.4252	245966.5256	200140.6522	205635.3819	7968
187878.701	176486.4437	173829.7477	245966.5256	200132.5389	205285.9083	15934
187878.701	176487.8997	173829.7249	245966.5256	200132.4342	205294.8453	15935
187878.701	176747.7048	173828.9865	245966.5256	200128.5912	207123.1018	23901
187878.701	176778.0333	173828.9672	245966.5256	200128.5051	207307.6371	23902
187878.701	176723.3841	173828.4069	245966.5256	200125.7028	206871.7545	31868
187878.701	176723.4163	173828.3928	245966.5256	200125.6382	206871.9366	31869
187878.701	176735.4956	173828.1031	245966.5256	200124.0864	206818.1594	39835
187878.701	176735.5698	173828.0869	245966.5256	200124.0095	206818.7758	39836

Table 14: ARPA-E network

Comparing against mean value solution, we can find stochastic solution provides better average performance in terms of overall cost of both pre-contingency and post contingency. With mean value solution gets final average cost of 187878.70, risk-based model get 176735.57 and standard two-stage model gets 173828.09, it shows that both risk-based model and standard two-stage model can reduce expected cost but in terms of average cost standard two-stage model perform best. However, when considering worst 20% scenarios, with mean value solution gets final risk of 245966.53, risk-based

model get 200124.01 and standard two-stage model gets 206818.78, it shows that both risk-based model and standard two-stage model can reduce risk but risk-based model perform best.

4 Mixed-integer nonlinear stochastic optimization

We consider the following two-stage stochastic program with integers in first and second stages[7], defined as follows.

$$\begin{aligned} &\min_{x\in S} \; f(x) + \mathbb{E}[Q(x,\omega)] \\ &\text{subject to } \underline{x} \le x \le \bar{x} \\ &\qquad\qquad x \in \mathbb{R}_+^{n_1-p_1} \times \mathbb{Z}_+^{p_1}, \end{aligned} \tag{SNIP}$$

where $Q(x,\omega)$ is the optimal value of the second-stage problem, defined as

$$\begin{aligned} &\min_{y} \; q(y,\omega) \\ &\text{subject to } W(y,\omega) + T(x,\omega) \preceq_C 0 \\ &\qquad\qquad \underline{y} \le y \le \bar{y} \\ &\qquad\qquad y \in \mathbb{R}_+^{n_2-p_2} \times \mathbb{Z}_+^{p_2}. \end{aligned} \tag{(Sub(x,ω))}$$

Above, n_1, n_2, p_1, p_2 are nonnegative integers with $p_1 \le n_1$ and $p_2 \le n_2$, x represents the first-stage decisions and y represents the second-stage decisions, and ω represents the uncertain data for the second-stage with known distribution. $S \subset \mathbb{R}^{n_1}$ is a convex set and $f : \mathbb{R}^{n_1} \to \mathbb{R}$ is a convex function. $q : \mathbb{R}^{n_2} \times \Omega \to \mathbb{R}$ is a convex function, $W = (w_1, \ldots, w_l) : \mathbb{R}^{n_2} \times \Omega \to \mathbb{R}^l$, $T = (t_1, \ldots, t_l) : \mathbb{R}^{n_1} \times \Omega \to \mathbb{R}^l$ and $C \subset \mathbb{R}^l$ is a closed convex cone. Furthermore, we assume that both f is convex and $q(\bullet, \omega)$ is convex for every $\omega \in \Omega$.

4.1 Literature review

A variety of applications in energy planning [125], manufacturing [126] logistics [127], etc. can be formulated as two-stage stochastic integer programs. In general, stochastic integer programming problems combine the difficulty of stochastic programming with integer programming. We first briefly review some important progress in theory and algorithms for solving the two-stage SIP.

(a) *Evaluating the expected second-stage cost for a fixed first-stage decision.* If the distribution of the uncertain parameters is continuous or the number of possible realizations is extremely large, then it is generally impossible to evaluate $\mathbb{E}[Q(x,\omega)]$ exactly. In this case, one possible solution to resort to approximating the underlying probability distribution by a manageable distribution. If the underlying distribution is continuous one may approximate it via discretization (see [128,129]). Another way is using statistical estimates of the expected value function via Monte Carlo sampling which has two kinds of methods. In interior sampling approaches, the estimation of $\mathbb{E}[Q(x,\omega)]$ is carried within the algorithm used to optimize this function [130]. In exterior sampling approaches, the sampling and optimization are decoupled. A Monte Carlo sample of the uncertain parameters is generated, and the expectation objective in the problem is replaced by a sample average [131,132]. An evaluation of the expected second-stage objective value requires solving many similar integer programs. When the second-stage variables are pure integer, several proposals which leverage Groebner basis and other test set based methods from computational algebra for exploiting IP problem similarity have been considered in [133–135]. For the case of mixed-integer subproblems, if a cutting plane method is used, then under some conditions it is possible to transform a cut (or a

[7] This section has been adapted from a working paper with Shisheng Cui and the PI as well as the Ph.D. Dissertation by Wendian Wan (advised by PI), titled "Algorithms for Operation of Power Systems: Risk, Uncertainty, Discreteness, and Nonconvexity."

valid inequality) derived for one of the second-stage subproblems into a cut for another subproblem by exploiting similarity [136, 137].

(b) *Optimizing the expected second-stage cost.* Much of the development in SIP has been towards the difficulty of optimizing $f(x) := c^T x + \mathbb{E}[Q(x, \omega)]$. We consider the following classifications.

(i) *Convex approximations of the value function.* There are results for constructing convex approximations of general integer recourse functions (SIPs involving pure integer second-stage variables) by perturbing the underlying distribution have been obtained [138]. These convex approximating functions are amenable for optimization and can be used to provide strong lower bounds within some of the algorithms for optimizing $f(x)$.

(ii) *Stage-wise decomposition algorithms.* This class of algorithms adopt the natural viewpoint of optimizing the objective function $f(x) := c^T x + \mathbb{E}[Q(x, \omega)]$ over the set of feasible first-stage decisions (say denoted by X). For SIPs with binary first-stage variables and mixed-integer second-stage variables, the integer L-shaped method [139] is widely used. Using disjunctive programming techniques, it is possible to derive cuts from the solutions [137, 140]. For SIPs where the first-stage variables are not necessarily all binary, dual functions from the second-stage integer program can, in principle, be used to construct cuts to build the approximation [141]. If the second-stage variables are pure integer (and the first-stage variables are mixed-integer), then it can be shown that $\mathbb{E}[Q(x, \omega)]$ is piece-wise constant over subsets that form a partitioning of the feasible region of x [134]. By exploiting certain monotonicity properties, the subsets can be enumerated efficiently within a branch-and-bound strategy [142]. Additional properties of the MIP value function $Q(x, \omega)$, such as sub-additivity, can be used to improve the method [143].

(iii) *Scenario-wise decomposition.* Copies of the first-stage variable are introduced for each scenario as the non-anticipativity constraint. Consider the Lagrangian dual problem obtained by relaxing the non-anticipativity constraints through the introduction of Lagrange multipliers. For a given set of multipliers, the problem is separable by scenarios, thus the dual function can be evaluated in a decomposed manner. Optimization of the dual function can be performed using standard non-smooth optimization techniques. However, owing to the non-convexities, there exists a duality gap, and one needs to resort to a branch-and-bound strategy to prove optimality [144].

(iv) *Cuts for deterministic equivalent MIP.* If the number of scenarios (assuming a finite distribution setting) is not astronomical then a possible approach is to directly solve the deterministic equivalent MIP, using a solver such as CPLEX or EXPRESS. One of the most important features of these solvers is the generation of cutting planes. In a stochastic integer program the constraint system is repeated with small changes for each scenario. Thus it is possible to effectively combine cuts from multiple rows corresponding to different scenarios [145]. Such cuts have been shown to significantly improve performance over single row cuts. An open issue is that such multi-row cuts links second stage variables across multiple scenarios, and hence destroys decomposability.

Since a key factor for solving two-stage SIP is solving the integer programming problem, we briefly review the two basic ideas.

(i) *Cut based methods (branch-and-cut/benders).* The basic idea is first solving a relaxation of the problem where some constraints are ignored or replaced with less stringent constraints.

This gives a lower bound on the true optimal value. If the relaxation solution is feasible, it is optimal. Otherwise, divide the feasible region (branch) and repeat. The total running time is related to two factors, time to process a node and number of nodes. Both can be very important. For very large instances (as in stochastic programming), solving a single relaxation can be too time-consuming. Number of nodes can grow exponentially in number of decision variables if do not prune often enough. Thus, it needs solving relaxations fast and obtaining strong relaxations so that can prune high in tree. We consider branch-and-cut to reduce number of nodes to explore with improved relaxation bounds and add inequalities required to define feasible region. This approach is the heart of all modern MIP solvers. Its basic idea includes solving current LP relaxation, attempting to generate valid inequalities that cut off the relaxed solution and if cuts found, adding to LP relaxation and repeat. There are two general approaches for SMIP with continuous recourse: benders with MIP master problem; branch-and-cut adding Benders cuts (and others) in tree. When we have mixed binary variables in the second-stage only, we can refer to [146] (lift-and-project cuts), [147] (reformulation linearization technique) and [148] (disjunctive cuts from branch-and-cut tree).

(ii) *Lagrangian relaxation based methods (dual decomposition).* The idea for these methods is creating copies of the first-stage decision variables for each scenario by adding nonanticipativity constraints. Then we relax these constraints using Lagrangian relaxation with dual vectors and each subproblem is a deterministic mixed-integer program. We could leverage subgradient algorithm to solve but it's slow in practice and very sensitive to step-size choices. Cutting plane algorithm with bundle-regularization techniques has good performance (see [149–153]). The general idea is adding objective term or constraint to encourage/require RMP solutions to not move "too far" in consecutive iterations. An elegant algorithm named progressive hedging [154] is used for solving primal and dual for convex stochastic programs, which is equivalent to alternating direction method of multipliers. It can be applied to solving quadratic MIP subproblems [155, 156]. If the primal subproblem solutions are not equal when we have (approximately) solved Lagrangian dual, we have two options: find a heuristic solution or use dual decomposition [144, 157]. To reduce the number of nonanticiptivity constraints, a scenario bundling approach was proposed [156, 158–160]. The ideas is partitioning the scenario set and treat scenarios within each "bundle" as a single scenario when doing decomposition.

(iii) *Stochastic mixed-integer programs with convex relaxations.* For stochastic convex program with mixed-integer variables, to the best of our knowledge, there are only two related work. In [161], Li and Grossmann proposed an improved L-shaped method for two-stage convex mixed-binary nonlinear stochastic programs. They included both Benders cuts and Lagrangean cuts in the Benders master problem. Although advantages over commercial global solvers were shown, the convergence guarantee was not presented due to the duality gap of the Lagrangean cuts and the integrality gap of the strengthened Benders cuts. In their other paper [162], they proved finite ϵ-convergence for the same problem by combining Benders decomposition with branch and bound scheme by constructing the convex hull of the MINLP subproblem for each scenario in closed-form.

4.2 Proposed scheme

Our scheme first develops a branching scheme applied to the master problem. Under the assumption that this relaxation is convex, we can apply variance-reduced schemes to a regularized problem. This leads to geometrically convergent sequences allowing us to resolve this subproblem efficiently. However, this does not suffice since the tree grows in accordance with second-stage discreteness.

Then we consider developing pruning rules that leverages problem structure across scenarios, thus we intend to develop more efficient branching schemes. Collectively, this is expected to lead to branching techniques that can contend with a high degree of nonlinearity. Extensions to this scheme will allow for nonconvex relaxations.

Throughout, we assume Ω has finite cardinality and define $X \triangleq \mathbb{Z}_+^{p_1+|\Omega|p_2} \cap \{(x_1, y_{11}, \ldots, y_{|\Omega|1}) : \underline{x}_1 \leq x_1 \leq \bar{x}_1, \underline{y}_{k1} \leq y_{k1} \leq \bar{y}_{k1},\ \forall k = 1, \ldots, |\Omega|\}$, $R \triangleq \mathbb{R}_+^{(n_1-p_1)+|\Omega|(n_2-p_2)} \cap \{(x_2, y_{12}, \ldots, y_{|\Omega|2}) : \underline{x}_2 \leq x_2 \leq \bar{x}_2, \underline{y}_{k2} \leq y_{k2} \leq \bar{y}_{k2},\ \forall k = 1, \ldots, |\Omega|\}$, $D \triangleq \{(x, y_1, \ldots, y_{|\Omega|}) \mid x \in S,\ W_k(y_k) + T_k(x) \preceq_C 0,\ \forall k = 1, \ldots, |\Omega|\}$. Let $z \triangleq (x, y_1, \ldots, y_{|\Omega|})$. The reason for distinguishing the sets X, R and D is that we are going to treat X directly, as a simple set. It is noted that in our problem, X is a finite set. Hence, in short, we consider the following convex program with integer variables:

$$\min_{z \in X \times R \cap D} f(x) + \sum_k p_k q_k(y_k) := F(z).$$

We present a computational framework for addressing such a problem by combining the variance-reduced stochastic approximation (VRSA) with a branching scheme. Notably, this scheme is a stochastic approximation scheme that combines smoothing, acceleration, and variance reduction. Such a framework is fairly adaptable and can allow for a broad range of risk-based convex models. The algorithm can be summarized as follows:

Algorithm 6 VRSA-BB

initialization: $f_u \leftarrow \infty$, $f_l \leftarrow -\infty$; $X_0 \leftarrow \emptyset$, $Y_{0k} \leftarrow \emptyset$, $\forall k \in \{1, \cdots, |\Omega|\}$; $(x^*, y^*) \leftarrow 0$;
`node_list`$\leftarrow \emptyset$, `candidate_list`$\leftarrow \emptyset$;
Node 0 := $(f_u, f_l, X_0, Y_0 := (Y_{0k})_{k=1}^{|\Omega|})$, add node 0 to `node_list`; $i \leftarrow 1$; Set number of iterations T;
while `node_list` is not empty **do**
    Choose a node n with smallest f_{n_l} from `node_list`; $f_s \leftarrow \infty$, $id \leftarrow 0$;
    Solve the subproblem by `VRSA`(X_n, Y_n, T), get $(f_m, (x_m, y_m)), X_m, Y_m$;
    Remove node n from `node_list`;
    if $n \in$ `candidate_list` **then**
        Remove node n from `candidate_list`;
    end if
    if no available partitions **then**
        Set $f_{n_l} \leftarrow f_m$, $f_{n_u} \leftarrow f_m$, $X_n \leftarrow X_m$, $Y_n \leftarrow Y_m$;
        Add node n to `node_list`;
        Add node n to `candidate_list`;
        if $f_m < f_u$ **then**
            $(x^*, y^*) \leftarrow (x_m, y_m)$; $f_u \leftarrow f_m$;
        end if
    else
        if integer constraints are all satisfied **then**
            $id \leftarrow 1$;
        end if
        Choose a variable $v_m \in (x_m, y_m)$ to branch; Let a_m be the value of v_m.
        if $v_m \in x_m$ **then**
            Add cut to X_m and generate X_{m1} and X_{m2}, respectively;
            for $w \in \{1, 2\}$ **do**
                Check X_{mw} feasibility;
                if feasible **then**
                    if $id = 1$ and $\{x_j = a_m\} \subset X_{mw}$ **then**
                        Set $i \leftarrow i + 1$, $f_{i_l} \leftarrow f_m$, $f_{i_u} \leftarrow f_m$, $X_i \leftarrow X_{mw}$, $Y_i \leftarrow Y_m$;

```
                    Add node i to candidate_list;
                    if f_m < f_u then
                        (x*, y*) ← (x_m, y_m); f_u ← f_m;
                    end if
                else
                    Set i ← i + 1, f_{i_l} ← f_m, f_{i_u} ← f_s, X_i ← X_mw, Y_i ← Y_m;
                end if
                Add node i to node_list;
            end if
        end for
    else
        v_m ∈ y_m, find the corresponding scenario k;
        Add cut to Y_mk and generate Y_mk1 and Y_mk2, respectively;
        for w ∈ {1, 2} do
            Check Y_mkw feasibility
            if feasible then
                Determine the feasible set of x; Add cut to X_mw;
                if id = 1 and {y_j = a_m} ⊂ Y_mkw then
                    Set i ← i + 1, f_{i_l} ← f_m, f_{i_u} ← f_m, X_i ← X_mw, Y_i ← Y_mw;
                    Add node i to candidate_list;
                    if f_m < f_u then
                        (x*, y*) ← (x_m, y_m); f_u ← f_m;
                    end if
                else
                    Set i ← i + 1, f_{i_l} ← f_m, f_{i_u} ← f_s, X_i ← X_mw, Y_i ← Y_mw;
                end if
                Add node i to node_list;
            end if
        end for
    end if
end if
T_i ← T_i + 1;
end while
```

Algorithm 7 VRSA(X, Y, T)

```
initialization: γ > 0;
Read X_n, Y_n, T from node n; Initialize x_0;
for t ∈ {0, 1, ..., T − 1} do
    N_t ← ⌊a^{t+1}⌋, a > 1; K_{N_t} ← N_t samples from Ω;
    for k ∈ {1, 2, ..., K_{N_t}} do
        Solve second stage problem on Y_nk with x_t; Return dual solution z_tk;
    end for
    x_{t+1} ← Π_{X_n} ( x_t − γ(∇f(x_t) + (1/N_t) Σ_{k=1}^{K_{N_t}} (∇T_k(x_{t−1})^T z_tk)) );
end for
for k ∈ {1, 2, ..., |Ω|} do
    Solve second stage problem on Y_nk with x_T; Return primal solution y_Tk;
end for
return (f(x_T, y_T), (x_T, y_T)), X_n, Y_n;
```

The key difference between VRSA-BB and the classical branch and bound scheme is that VRSA-BB employs a stochastic method to solve each continuous relaxation problem. The essential is that the second stage problem can be divided into many small problems and they can be solved by VRSA efficiently. To prove the almost sure convergence of VRSA-BB, we need the following preliminary

knowledge on stochastic branch and bound schemes.
In the branch and bound method the original set X is sequentially subdivided into subsets X^p generating a partition $\mathcal{P}$ of X (or of its part) such that $\bigcup_{X^p\in\mathcal{P}} X^p = X$. Consequently, the original problem is subdivided into subproblems

$$\min_{z\in X^p\times R\cap D} \ F(z),\ X^p \in \mathcal{P}.$$

Let $F^*(X^p)$ denote the optimal value of this subproblem. Clearly, we have the optimal value of the original problem equals

$$F^*(X) = \min_{X^p\in\mathcal{P}} \ F^*(X^p).$$

Before proceeding, we describe the concept of stochastic bounds. First, we need the following two assumptions.

Assumption 11. There exist functions $L : 2^X \times R \to \mathbb{R}\cup\{-\infty\}$ and $U : 2^X \times R \to \mathbb{R}\cup\{+\infty\}$ such that for each $X^p \subset X$ (for brevity, we use $L(X^p)$ and $U(X^p)$ denote the function value, respectively),

$$L(X^p) \leq F^*(X^p) \leq U(X^p),$$
$$U(X^p) = F(z') \text{ for some } z' \in X^p \times R, \text{ if } U(X^p) < +\infty$$

and if X^p is degenerated into a singleton then

$$L(X^p) = F^*(X^p) = U(X^p).$$

In general, in stochastic problems the bounds $L(X^p)$ and $U(X^p)$ can hardly be computed exactly. Therefore we can only assume that some statistical estimates of $L(X^p)$ and $U(X^p)$ can be obtained.

Assumption 12. In some probability space $(\Omega, \Sigma, \mathbb{P})$, for each subset $X^p \subset X$, there exist sequences of random estimates $\xi^l(X^p,\omega),\ l = 1, 2, \dots,$ and $\eta^m(X^p,\omega),\ m = 1, 2, \dots,\ \omega \in \Omega$, such that

$$\lim_{l\to\infty} \xi^l(X^p,\omega) = L(X^p) \text{ a.s.},$$
$$\lim_{m\to\infty} \eta^m(X^p,\omega) = U(X^p) \text{ a.s.}.$$

Moreover, there exists z^P_m such that $F(z^P_m) = \eta^m(X^p,\omega)$ if $\eta^m(X^p,\omega) < +\infty$, for $m = 1, 2, \dots,$ $\omega \in \Omega$.

From now on, for brevity, we skip the argument ω from the random indices l and m, random partitions $\mathcal{P}$ and random sets. Next, we briefly introduce a general stochastic branch and bound framework.

Algorithm 8 SBB

1: **initialization:** Initial partition $\mathcal{P}_0 = \mathcal{P}'_0 = X$; $\xi_0 = \xi^{l_0}(X)$; $\eta_0 = \eta^{m_0}(X)$; $k = 1$;
2: **partitioning:** Select the record subset $Y^k \in \operatorname{argmin}\{\xi_k(X^p) : X^p \in \mathcal{P}_k\}$; If $\min_{X^p\in\mathcal{P}_k} \eta_k(X^p) < +\infty$, select a set $X^k \in \operatorname{argmin}\{\eta_k(X^p) : X^p \in \mathcal{P}_k\}$ and a solution $z^k \in \{(x, r) : x \in X_k, F(x, r) = \eta_k(X^k)\}$;
3: **if** Y^k is a singleton **then**
4: $\quad \mathcal{P}_k = \mathcal{P}_{k-1}$;
5: **else**
6: $\quad \mathcal{P}''_k(Y^k) = \{Y^k_i, i = 1, 2, \dots\}$, such that $Y^k = \bigcup_i Y^k_i$ and $Y^k_i \cap Y^k_j = \emptyset$ for $Y^k_i, Y^k_j \in \mathcal{P}''_k, i \neq j$;

7: $\mathcal{P}_k = (\mathcal{P}'_k \setminus Y^k) \cup \mathcal{P}''_k(Y^k)$;
8: **end if**
9: **bound estimation:** For all subsets $X^p \in \mathcal{P}_k$, select estimates $\xi_k(X^p) = \xi^{l_k(X^p)}(X^p)$ and $\eta_k(X^p) = \eta^{m_k(X^p)}(X^p)$ for $L(X^p)$ and $U(X^p)$;
10: **deletion:** $\mathcal{P}'_k = \mathcal{P}_k \setminus \{X^p : X^p \cap D = \emptyset\}$; k:=k+1; Go to partitioning;

Now we are ready to state the following lemma which is important to prove a.s. convergence of our scheme. We denote by $Z^* = X^* \times R^*$ the solution set.

Lemma 22. *Suppose Assumptions 11 and 12 hold. Consider partitions following Algorithim 3. Assume that if a subset $X' \subset \mathcal{P}_k$ for infinitely many k, then a.s. $\lim_{k\to\infty} l_k(X') = \lim_{k\to\infty} m_k(X') = \infty$ a.s.. Define recurrent record sets as those, which are record sets for infinitely many k. Then the following holds*

(i) Almost surely, there exits an iteration number k_0 such that for all $k \geq k_0$, Y_k are singletons and recurrent, and $Y_k \subset X^*$; All approximate sets X_k are recurrent;

(ii) z^k converges to Z^* a.s..

Proof. (i) Please refer to [163, Theorem 3.].
(ii) Consider an approximate set X_k, where $k \geq k_0$. By definition

$$\eta^{m_k(X^k)}(X^k) \leq \eta^{m_k(Y^k)}(Y^k),$$

where $Y^k \subset X^*$. Because for all $k \geq k_0$ all record sets are recurrent, then by assumption, $m_k(Y^k) \to \infty$, so $\eta^{m_k(Y^k)}(Y^k) \to U(Y^k) = F^*(Y^k) = F^*$. Thus,

$$\limsup_{k\to\infty} \eta^{m_k(X^k)}(X^k) \leq F^*.$$

Due to finiteness of the partition and the finiteness of the sets, there are finite unique elements in $\{X^k, k \geq k_0\}$. According to (i), for all $k \geq k_0$, X_k is recurrent, which means for each element $X' \in \{X^k, k \geq k_0\}$, we have $X' = X^k$ for infinitely many k. We arbitrarily choose a $X' \in \{X^k, k \geq k_0\}$. X' is recurrent, which means there is a subsequence $\mathcal{K}$ such that $X^k = X'$, $\forall k \in \mathcal{K}$. Then we have

$$F(z') = U(X') = \lim_{k\in\mathcal{K}, k\to\infty} \eta^{m_k(X')}(X') \leq F^*.$$

It's clear to see that $F(z') = F^*$. In addition, we have $F(z^k) = \eta^{m_k(X^k)}(X^k)$, then

$$\lim_{k\in\mathcal{K}, k\to\infty} F(z^k) = \lim_{k\in\mathcal{K}, k\to\infty} \eta^{m_k(X^k)}(X^k) = \lim_{k\in\mathcal{K}, k\to\infty} \eta^{m_k(X')}(X') = F^*.$$

It holds for all $X' \in \{X^k, k \geq k_0\}$, therefore the conclusion follows. □

Clearly, in (VSRA-BB), at each node we solve a continuous relaxation problem for the lower bound, i.e., for all $X^p \subset X$, we construct

$$L(X^p) := \min_{z\in Y^p \times R \cap D} F(z)$$
$$U(X^p) := \begin{cases} L(X^p) & \text{if } X^p \text{ is a singleton} \\ +\infty & \text{otherwise,} \end{cases}$$

where $Y^p \in \mathbb{R}_+^{p_1+p_2}$ is the relaxation set associated with X^p. For a partition $\mathcal{P}$, $\{Y^p\}_{p\in\mathcal{P}}$ is an admissible disjunction with set X. Thus, we have $L(X^p) \leq F^*(X^p) \leq U(X^p)$ and $L(X^p) =$

$F^*(X^p) = U(X^p)$ when X^p is a singleton. Hence, Assumption 11 is satisfied. For Assumption 12, we first show the continuous relaxation problem is a monotone inclusion and (VRSA) can solve this problem and generate a feasible sequence $\{z^k\}$ which converges to the optimal solution almost surely. Here k denotes the number of iterations of the scheme. Then the estimates for $L(\cdot)$ and $U(\cdot)$ are as follows

$$\xi^k(X^p) := F(z^k : z^k \in (Y^p \times R \cap D)$$
$$\eta^k(X^p) := \begin{cases} \xi^k(X^p) & \text{if } X^p \text{ is a singleton} \\ +\infty & \text{otherwise.} \end{cases}$$

The relaxation problem we consider is

$$\min_{x \in S \cap \mathbb{R}^{n_1}_+} f(x) + \mathbb{E}[\mathcal{Q}(x, \omega)]$$

where $Q(x, \omega)$ is the optimal value of the second-stage problem

$$\min_{y \in \mathbb{R}^{n_2}_+} q(y, \omega)$$
$$\text{subject to } W(y, \omega) + T(x, \omega) \preceq_C 0.$$

The Lagrangian dual problem of the second stage can be written in the form

$$\max_{\pi \succeq_{C^0}} \pi^T T(x, \omega) + \inf_{y \in \mathbb{R}^{n_2}_+} L(y, \pi, \omega),$$

where $L(y, \pi, \omega) := q(y, \omega) + \pi^T W(y, \omega)$. Let us denote by $\Lambda(x, \omega)$ the set of optimal solutions of the dual problem. By leveraging the properties of the recourse function (cf. [164]), we recall $\mathcal{Q}(x, \omega)$, $\partial_x \mathbb{E}[\mathcal{Q}(x, \omega)] = \mathbb{E}[\partial_x \mathcal{Q}(x, \omega)]$ where $\nabla T(x, \omega)^T \Lambda(x, \omega) = \partial_x \mathcal{Q}(x, \omega)$. Consequently, the optimality conditions of the first stage problem is given by

$$0 \in \nabla f(x) + \underbrace{\mathbb{E}[\nabla T(x, \omega)^T \Lambda(x, \omega)]}_{\triangleq G(x)} + \mathcal{N}_{S \cap \mathbb{R}^{n_1}_+}(x) := V(x).$$

We observe that $\nabla f(\cdot)$ is a monotone map while $G(x)$ is the expectation of subdifferentials, implying that G is also a monotone map. Furthermore, $\mathcal{N}_{\mathcal{X}}$ is a normal cone of a convex set, it is also a monotone map.

Another important implementational and theoretical issue is the stopping criterion. Clearly, because of the stochastic nature of the bounds, a solution obtained after a finite number of observations is, in general, an approximation. We have the following lemma regarding this.

Lemma 23. *Assume that the algorithm stops at iteration s and that we can build for all $X^s \in \mathcal{P}_s$ confidence intervals $[\underline{\xi}(X^s), +\infty)$ for $L(X^s)$ such that*

$$\mathsf{P}\left\{\forall X^s \subset \mathcal{P}_s,\ \underline{\xi}(X^s) \leq L(X^s)\right\} \geq 1 - \delta.$$

Then, with probability at least $1 - \delta$,

$$F(z^s) - F^* \leq F(z^s) - \min_{X^s \in P_s} \underline{\xi}(X^s).$$

Proof. With probability not smaller than $1 - \delta$, $F^* \geq \min_{X^s \in P_s} L(X^s) \geq \min_{X^s \in P_s} \underline{\xi}(X^s)$. □

Although the construction of the confidence interval may not be done in an explicit way, we can choose $\xi(\cdot)$ as an estimator of $\underline{\xi}(\cdot)$ and employ it in (VSRA-BB).

Algorithm 9 VRSA-BB with tolerance

initialization: $f_u \leftarrow \infty$, $f_l \leftarrow -\infty$; $X_0 \leftarrow \emptyset$, $Y_{0k} \leftarrow \emptyset$, $\forall k \in \{1, \cdots, |\Omega|\}$; $(x^*, y^*) \leftarrow 0$; $\varepsilon > 0$;
`node_list`$\leftarrow \emptyset$, `candidate_list`$\leftarrow \emptyset$;
Node 0 $:= (f_u, f_l, X_0, Y_0 := (Y_{0k})_{k=1}^{|\Omega|})$, add node 0 to `node_list`; $i \leftarrow 1$; Set number of iterations T;
while `node_list` is not empty **do**
    Choose a node n with smallest f_{n_l} from `node_list`; $f_s \leftarrow \infty$, $id \leftarrow 0$;
    Solve the subproblem by `VRSA`(X_n, Y_n, T), get $(f_m, (x_m, y_m)), X_m, Y_m$;
    Remove node n from `node_list`;
    if $n \in$ `candidate_list` **then**
        Remove node n from `candidate_list`;
    end if
    if no available partitions **then**
        Set $f_{n_l} \leftarrow f_m$, $f_{n_u} \leftarrow f_m$, $X_n \leftarrow X_m$, $Y_n \leftarrow Y_m$;
        Add node n to `node_list`;
        Add node n to `candidate_list`;
        if $f_m < f_u$ **then**
            $(x^*, y^*) \leftarrow (x_m, y_m)$; $f_u \leftarrow f_m$;
        end if
    else
        if integer constraints are all satisfied **then**
            $id \leftarrow 1$;
        end if
        Choose a variable $v_m \in (x_m, y_m)$ to branch; Let a_m be the value of v_m.
        if $v_m \in x_m$ **then**
            Add cut to X_m and generate X_{m1} and X_{m2}, respectively;
            for $w \in \{1, 2\}$ **do**
                Check X_{mw} feasibility;
                if feasible **then**
                    if $id = 1$ and $\{x_j = a_m\} \subset X_{mw}$ **then**
                        Set $i \leftarrow i + 1$, $f_{i_l} \leftarrow f_m$, $f_{i_u} \leftarrow f_m$, $X_i \leftarrow X_{mw}$, $Y_i \leftarrow Y_m$;
                        Add node i to `candidate_list`;
                        if $f_m < f_u$ **then**
                            $(x^*, y^*) \leftarrow (x_m, y_m)$; $f_u \leftarrow f_m$;
                        end if
                    else
                        Set $i \leftarrow i + 1$, $f_{i_l} \leftarrow f_m$, $f_{i_u} \leftarrow f_s$, $X_i \leftarrow X_{mw}$, $Y_i \leftarrow Y_m$;
                    end if
                    Add node i to `node_list`;
                end if
            end for
        else
            $v_m \in y_m$, find the corresponding scenario k;
            Add cut to Y_{mk} and generate Y_{mk1} and Y_{mk2}, respectively;
            for $w \in \{1, 2\}$ **do**
                Check Y_{mkw} feasibility
                if feasible **then**
                    Determine the feasible set of x; Add cut to X_{mw};
                    if $id = 1$ and $\{y_j = a_m\} \subset Y_{mkw}$ **then**
                        Set $i \leftarrow i + 1$, $f_{i_l} \leftarrow f_m$, $f_{i_u} \leftarrow f_m$, $X_i \leftarrow X_{mw}$, $Y_i \leftarrow Y_{mw}$;
                        Add node i to `candidate_list`;

50: **if** $f_m < f_u$ **then**
51: $(x^*, y^*) \leftarrow (x_m, y_m)$; $f_u \leftarrow f_m$;
52: **end if**
53: **else**
54: Set $i \leftarrow i+1$, $f_{i_l} \leftarrow f_m$, $f_{i_u} \leftarrow f_s$, $X_i \leftarrow X_{mw}$, $Y_i \leftarrow Y_{mw}$;
55: **end if**
56: Add node i to `node_list`;
57: **end if**
58: **end for**
59: **end if**
60: **end if**
61: Set $f_l \leftarrow \min\{f_{n_l} : n \in$ `node_list`$\}$;
62: **if** $f_u - f_l < \varepsilon$ **then**
63: Terminate;
64: **end if**
65: $T_i \leftarrow T_i + 1$;
66: **end while**

4.3 Numerical implemenation

In this section, we report our computational experience with the proposed algorithm on instances of two-stage stochastic integer programs from the literature. The test problems involves pure-integer first-stage variables and mixed-binary second-stage variables and is inspired from Ahmed, Tawarmalani and Sahinidis [142]. The test problems are generated from the following basic model:

$$\begin{aligned} &\min\ 0.5x_1^2 + 0.5x_2^2 - 1.5x_1 - 4x_2 + \mathbb{E}[\mathcal{Q}(x_1, x_2, \omega_1, \omega_2)] \\ &\text{subject to } x_1, x_2 \in [0,5] \cap \mathbb{Z}_+, \end{aligned}$$

where

$$\begin{aligned} \mathcal{Q}(x_1, x_2, \omega_1, \omega_2) := \min\ & 0.5y_1^2 + 0.5y_2^2 + 0.5y_3^2 + 0.5y_4^2 - 16y_1 - 19y_2 - 23y_3 - 28y_4 \\ \text{subject to } & 2y_1 + 3y_2 + 4y_3 + 5y_4 \le \omega_1 - x_1 \\ & 6y_1 + y_2 + 3y_3 + 2y_4 \le \omega_2 - x_2 \\ & y_1, y_2, y_3, y_4 \in \{0,1\}, \end{aligned}$$

where (ω_1, ω_2) is uniformly distributed on $\Omega \subseteq [5,15] \times [5,15]$. Hence, both stages are quadratic programs.
In the implementation, we use several acceleration approach.

(1) *Mixed-integer rounding cuts.* Suppose the constraints in the second-stage are linear with mixed-integer variables with the form

$$\left\{ v \in \mathbb{R}_+^{|C|}, y \in \mathbb{Z}_+^{|I|} : \sum_{j \in C} c_j v_j + \sum_{j \in I} a_j y_j \ge b \right\},$$

then the MIR cut is given as

$$\sum_{c_j > 0} c_j v_j + \sum_{\hat{a}_j < \hat{b}_j} \hat{a}_j y_j + \hat{b} \left(\sum_{\hat{a}_j \ge \hat{b}_j} y_j + \sum_{j \in I} \lfloor a_j \rfloor y_j \right) \ge \hat{b} \lceil b \rceil,$$

where $a = \hat{a} + \lfloor a \rfloor$, $b = \hat{b} + \lfloor b \rfloor$. Before line 11 of **VRSA**, we could add MIR cuts to enforce more variables of y_{ns} to be integer. It could significantly reduce the number of nodes.

(2) *Multiprocessing scheme with best-first strategy.* We could construct a decent upper bound after line 11 of **VRSA** by rounding the values of first-stage integer variables. Thus, the best-first strategy could help improve the lower bound at each iteration. Furthermore, it can be combined with the multiprocessing scheme which deals with multiple best possible values simultaneously, and it reduces the running time notably.

(3) *Automatic second-stage cuts.* Consider the second-stage constraint

$$Wy_\omega + Tx \leq h_\omega,$$

where

$$W = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 6 & 1 & 3 & 2 \end{bmatrix} \quad T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad h_\omega = \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix}.$$

When a variable $y_{\omega j}$ is branched and a cut $y_{\omega j} \leq c$ is added, it can be noticed that if the second-stage problem is feasible after adding the cut, then the scenario ω' with the same cut is still feasible once $h_{\omega'} \geq h_\omega$. The reason is that all elements in W are positive. Similar arguments can be made depending on the sign of coefficients associated with each second-stage variable.

The proposed optimization routine was implemented in Python. In the (VRSA-BB) algorithm, we employed the CVXOPT solver to solve the projection problem. The termination tolerance for the scheme is set at 0.01. All computations are performed on a PC with 16GB RAM and 6-Core Intel Core i7 processor. In Table 15, we compare the cpu seconds required by the VRSA-BB algorithm against that required by the Gurobi 9.1 mixed-integer quadratic programming solver for solving the two-stage problem for sample sizes $N = 50$ to 400.

Scenario#	Integer#	LB	UB	Time	Gurobi Solution	Gurobi Time
50	52	-51.84	-51.39	19.4	-51.53	0.7
100	102	-57.90	-57.36	51.3	-57.51	1.3
200	202	-61.29	-60.70	176.1	-60.96	15.6
300	302	-60.38	-59.80	418.4	-60.02	110.0
400	402	-59.71	-59.12	575.2	-59.31	2312.8

Table 15: QP with 2 integers in 1st-stage and 1 binary in each 2nd-stage

5 Zeroth-order schemes for Stochastic MPECs

In this section, we consider the resolution of variants and stochastic generalizations of the mathematical program with equilibrium constraints[8] (MPEC), given by

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & f(\mathbf{x},\mathbf{y}) \\ \text{subject to} \quad & \mathbf{y} \in \text{SOL}(\mathcal{Y}, F(\mathbf{x},\bullet)), \\ & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{MPEC}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a real-valued function, $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$, $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ denote closed and convex sets, and $\text{SOL}(\mathcal{Y}, F(\mathbf{x},\bullet))$ denotes the solution set of the parametrized variational inequality problem $\text{VI}(\mathcal{Y}, F(\mathbf{x},\bullet))$, given an upper-level decision $\mathbf{x}$. Recall that the variational inequality problem $\text{VI}(\mathcal{Y}, F(\mathbf{x},\bullet))$ requires a vector $\mathbf{y}$ in the set $\mathcal{Y}$ such that

$$(\tilde{\mathbf{y}} - \mathbf{y})^T F(\mathbf{x},\mathbf{y}) \ \geq \ 0, \qquad \forall \ \tilde{\mathbf{y}} \ \in \ \mathcal{Y}. \tag{VI}(\mathcal{Y}, F(\mathbf{x},\bullet)))$$

MPECs have a broad range of applications arising in hierarchical optimization, frictional contact problems, power systems [166], traffic equilibrium problems [167], and Stackelberg equilibrium problems [168]. A comprehensive survey of models, analysis, and algorithms can be found in [169] while a subsequent monograph emphasized the implicit framework [170].

The MPEC is an ill-posed generalization of a nonconvex and nonlinear program, an observation that follows from considering the setting where $\mathcal{Y}$ is a convex cone in $\mathbb{R}^m$. In such an instance, (MPEC) reduces to a mathematical program with complementarity constraints (MPCC) since $\mathbf{y}$ solves $\text{VI}(\mathcal{Y}, F(\mathbf{x},\bullet))$ if and only if $\mathbf{x}$ solves $\text{CP}(\mathcal{Y}, F(\mathbf{x},\bullet))$, defined as the problem of finding a vector $\mathbf{y}$ such that

$$\mathcal{Y} \ni \mathbf{y} \ \perp \ F(\mathbf{x},\mathbf{y}) \in \mathcal{Y}^*, \tag{CP}(\mathcal{Y}, F(\mathbf{x},\bullet)))$$

where $\mathcal{Y}^* \triangleq \{u \mid y^T u \geq 0, y \in \mathcal{Y}\}$. When $\mathcal{Y}$ is the nonnegative orthant, then (MPEC) reduces to the following MPCC.

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & f(\mathbf{x},\mathbf{y}) \\ \text{subject to} \quad & 0 \leq \mathbf{y} \perp F(\mathbf{x},\mathbf{y}) \geq 0, \\ & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{MPCC}$$

Ill-posedness of (MPCC) arises from noting that standard constraint qualifications (such as the Mangasarian-Fromovitz constraint qualification) fail to hold at any feasible point. This has led to a concerted effort in developing weaker stationarity conditions for MPECs [171] as well as a host of regularization [172–176] and penalization [177] schemes.

Yet an enduring gap persists in the development of algorithms for such problems. Despite a wealth of developments in the field of zeroth and first-order algorithms for deterministic and stochastic convex and nonconvex optimization, there are no available non-asymptotic rate guarantees for either zeroth or first-order schemes for MPECs or their stochastic variants. In particular, our interest lies in two distinct stochastic variants presented as follows.

[8] This section has been adapted from [165].

5.1 Problems of interest

We focus on the problem (MPEC) where the lower-level map $F(\mathbf{x}, \bullet)$ is strongly monotone over $\mathcal{Y}$ uniformly in $\mathbf{x}$. This ensures that the solution of $\mathrm{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet))$ is a singleton for every $\mathbf{x} \in \mathcal{X}$. We consider two generalizations of the lower-level problem.

(i) *Mathematical programs with equilibrium constraints characterized by expectation-valued maps.* In many settings, the lower-level equilibrium constraint is cast as a stochastic variational inequality problem, i.e. a variational inequality problem with expectation-valued maps. Such problems assume relevance in modeling a range of stochastic equilibrium problems; more specifically, such problems represent the necessary and sufficient equilibrium conditions of smooth stochastic convex optimization problems and smooth stochastic convex Nash equilibrium problems [178, 179]. More formally, suppose the variational inequality problem $\mathrm{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet))$ is characterized by a map F whose components are expectation-valued, i.e.

$$F(\mathbf{x}, \mathbf{y}) \triangleq \begin{pmatrix} \mathbb{E}[G_1(\mathbf{x}, \mathbf{y}, \xi(\omega))] \\ \vdots \\ \mathbb{E}[G_m(\mathbf{x}, \mathbf{y}, \xi(\omega))] \end{pmatrix}, \tag{86}$$

where $\xi : \Omega \to \mathbb{R}^d$ and $G_i : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$. For the ease of presentation, throughout the paper, we refer to the integrand $G_i(\mathbf{x}, \mathbf{y}, \xi(\omega))$ by $G_i(\mathbf{x}, \mathbf{y}, \omega)$. In effect, the lower-level problem is a stochastic variational inequality problem [178, 180]. In addition, the objective may also be expectation-valued and the resulting problem is defined as follows.

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & \mathbb{E}[f(\mathbf{x}, \mathbf{y}, \omega)] \\ \text{subject to} \quad & \mathbf{y} \in \mathrm{SOL}(\mathcal{Y}, \mathbb{E}[G(\mathbf{x}, \bullet, \omega)]), \\ & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{SMPEC$^{\mathrm{exp}}$}$$

An instance of the above formulation for stochastic mathematical program with equilibrium constraints is the case where the lower-level equilibrium problem captures the equilibrium conditions of a convex stochastic optimization problem given by

$$\min_{\mathbf{y} \in \mathcal{Y}} \mathbb{E}[h(\mathbf{x}, \mathbf{y}, \omega)], \tag{87}$$

where $F(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{E}[\nabla_{\mathbf{y}} h(\mathbf{x}, \mathbf{y}, \omega)]$. A more general instance is when a solution to the lower-level equilibrium problem is a Nash equilibrium of a noncooperative game with expectation-valued objectives, as given by

$$\min_{\mathbf{y}_i \in \mathcal{Y}_i} \mathbb{E}[h_i(\mathbf{x}, (\mathbf{y}_i; \mathbf{y}_{-i}), \omega)], \tag{88}$$

where $i \in \{1, \ldots, N\}$, N denotes the number of players, $\mathbf{y}_i \in \mathcal{Y}_i$ and $h_i(\mathbf{x}, (\bullet; \mathbf{y}_{-i}), \omega)$ denote the strategy set and the cost function of player $i \in \{1, \ldots, N\}$, respectively, and $\mathbf{y}_{-i}$ denotes the strategies of the other players than player i. Under some mild conditions, it is known that the equilibrium conditions of the aforementioned game can be characterized as $\mathrm{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet))$ where $\mathcal{Y} \triangleq \prod_{i=1}^N \mathcal{Y}_i$ and $F(\mathbf{x}, \mathbf{y}) \triangleq \prod_{i=1}^N \mathbb{E}[\nabla_{\mathbf{y}_i} h_i(\mathbf{x}, (\mathbf{y}_i; \mathbf{y}_{-i}), \omega)]$ (cf. Chap. 1 in [63]).
An alternate approach for modeling uncertainty in MPECs is provided in the next model, where the lower-level problem constraints are imposed in an almost sure (a.s.) sense [181].

(ii) *MPECs with almost sure equilibrium constraints.* Consider a leader-follower game where the follower makes decision $\mathbf{y}$ contingent on the leader's decision $\mathbf{x}$ and the realization of uncertainty is denoted by ω. Consequently, the leader's problem requires minimizing her expected cost

$\mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)]$ where $\mathbf{y}(\mathbf{x}, \omega)$ represents follower's decision, given $\mathbf{x}$ and ω. Such a problem can be compactly represented as (SMPEC$^{\text{as}}$), defined next.

$$\begin{array}{rl} \min\limits_{\mathbf{x}, \mathbf{y}(\omega)} & \mathbb{E}[f(\mathbf{x}, \mathbf{y}(\omega), \omega)] \\ \text{subject to} & \mathbf{y}(\omega) \in \text{SOL}(\mathcal{Y}(\mathbf{x}, \omega), G(\mathbf{x}, \bullet, \omega)), \text{ for almost every } \omega \in \Omega \\ & \mathbf{x} \in \mathcal{X}. \end{array} \tag{SMPEC$^{\text{as}}$}$$

In regimes where $\text{VI}(\mathcal{Y}(\mathbf{x}, \omega), G(\mathbf{x}, \bullet, \omega))$ has a unique solution for any $\mathbf{x} \in \mathcal{X}$ and any $\omega \in \Omega$, we may recast (SMPEC$^{\text{as}}$) as the following *implicit* stochastic optimization problem where $\mathbf{y} : \mathcal{X} \times \Omega \to \mathbb{R}^m$ denotes a single-valued solution map.

$$\begin{array}{rl} \min\limits_{\mathbf{x}} & \mathbb{E}[f(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)] \\ \text{subject to} & \mathbf{x} \in \mathcal{X}. \end{array} \tag{SMPEC$^{\text{as}}$}$$

5.2 Background

Throughout this paper, we assume that in the case of (MPEC) and (SMPEC$^{\text{exp}}$), the set $\mathcal{Y}$ is closed and convex in $\mathbb{R}^m$ and the parametrized map $F(\mathbf{x}, \bullet)$ is strongly monotone on $\mathcal{Y}$ uniformly in $\mathbf{x}$. An analogous assumption for (SMPEC$^{\text{as}}$) requires that $G(\mathbf{x}, \bullet, \omega)$ is strongly monotone on $\mathcal{Y}$ for every $\omega \in \Omega$. Since lower-level problem is strongly monotone, the solution map of the lower-level problem is single-valued. Consequently, we may recast (SMPEC$^{\text{as}}$) as the following implicit program in $\mathbf{x}$.

$$\min_{\mathbf{x} \in \mathcal{X}} \; f(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)], \tag{MPEC$^{\text{imp,as}}$}$$

where $f(\bullet)$ is assumed to be Lipschitz continuous on a closed and convex set $\mathcal{X}$. In the case of (SMPEC$^{\text{exp}}$), the implicit problem reduces to

$$\min_{\mathbf{x} \in \mathcal{X}} \; f(\mathbf{x}, \mathbf{y}(\mathbf{x})), \tag{MPEC$^{\text{imp,exp}}$}$$

where $\mathbf{y}(\mathbf{x})$ represents the solution to a variational inequality problem $\text{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet))$. Note that this problem subsumes (SMPEC$^{\text{exp}}$) by suppressing the expectation in the upper-level. We now formalize the assumptions on the problems of interest.

Assumption 13 (Properties of $f, F, \mathcal{X}, \mathcal{Y}$). Consider the problem (MPEC$^{\text{imp,as}}$) or (MPEC$^{\text{imp,exp}}$).
(a.i) $f(\bullet, \mathbf{y}(\bullet))$ is L_0-Lipschitz continuous on $\mathcal{X} + \eta_0 \mathbb{B}$ for some $\eta_0 > 0$. $f(\mathbf{x}, \bullet)$ is Lipschitz with the parameter $\tilde{L}_0 > 0$ for all $\mathbf{x} \in \mathcal{X} + \eta_0 \mathbb{B}$ for some $\eta_0 > 0$.
(a.ii) $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are nonempty, closed, and convex sets.
(a.iii) $F(\mathbf{x}, \bullet)$ is a μ_F-strongly monotone and L_F-Lipschitz continuous map on $\mathcal{Y}$ uniformly in $\mathbf{x} \in \mathcal{X}$.

Consider the problem (MPEC$^{\text{imp,as}}$).
(b.i) $\tilde{f}(\bullet, \mathbf{y}(\bullet, \omega), \omega)$ is L_0-Lipschitz continuous on $\mathcal{X} + \eta_0 \mathbb{B}$ for every $\omega \in \Omega$ and for some $\eta_0 > 0$. $f(\mathbf{x}, \bullet)$ be Lipschitz with the parameter $\tilde{L}_0 > 0$ for all $\mathbf{x} \in \mathcal{X} + \eta_0 \mathbb{B}$ for some $\eta_0 > 0$.
(b.ii) $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are closed and convex sets.
(b.iii) $G(\mathbf{x}, \bullet, \omega)$ is a μ_F-strongly monotone and L_F-Lipschitz continuous map on $\mathcal{Y}$ uniformly in $\mathbf{x} \in \mathcal{X}$ for every $\omega \in \Omega$. □

We observe that the requirement that f is Lipschitz continuous on $\mathcal{X} + \eta_0\mathbb{B}$ (rather than $\mathcal{X}$) is a consequence of employing a smoothed approximation of f in our algorithm development. A natural question is whether the Lipschitz continuity of the objective f over $\mathcal{X}$ in the implicit problem follows under reasonable conditions. The next result addresses precisely such a concern.

Proposition 7. Consider the problem (SMPEC$^{\text{as}}$). Suppose Assumption 13 (a.ii, a.iii) hold. Suppose $\tilde{f}(\bullet, \bullet, \omega)$ is continuously differentiable on $\mathcal{C} \times \mathbb{R}^m$ where $\mathcal{C}$ is an open set containing $\mathcal{X}$, and $\mathcal{X}$ is bounded. Then the function f, defined as $f(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)]$, is Lipschitz and directionally differentiable on $\mathcal{X}$.

Proof. This result follows from invoking [182, Cor. 4.3] together with the compactness of $\mathcal{X}$. □

Naturally, when Ω reduces to a singleton, one may obtain a similar claim for (MPEC) and (MPEC$^{\text{imp,exp}}$). In a subset of regimes, $f(\bullet, \mathbf{y}(\bullet))$ is captured by the next assumption.

Assumption 14 (Convexity of f in implicit problem). Consider any of the implicit problems (MPEC$^{\text{imp,as}}$) or (MPEC$^{\text{imp,exp}}$). Suppose $f(\bullet, \mathbf{y}(\bullet))$ is convex on $\mathcal{X}$ in (MPEC$^{\text{imp,exp}}$) or $f(\bullet)$ is convex on $\mathcal{X}$ in (MPEC$^{\text{imp,as}}$).

We note that there has been extensive study of conditions under which the implicit function $f(\bullet, \mathbf{y}(\bullet))$ is indeed convex (for example, see [181–183]).

5.2.1 Stationarity conditions

While f can be shown to be convex in some select settings, the function f is Lipschitz continuous on $\mathcal{X}$ in more general settings. Consequently, the problem can be compactly stated as

$$\min_{\mathbf{x} \in \mathcal{X}} \; h(\mathbf{x}) \triangleq f(\mathbf{x}, \mathbf{y}(\mathbf{x})). \tag{89}$$

We observe that h is a nonsmooth and possibly nonconvex function on $\mathcal{X}$. In the remainder of this subsection, we recap some of the concepts of Clarke's nonsmooth calculus that will facilitate the development of stationarity conditions. We begin by defining the directional derivative, a key object necessary in addressing nonsmooth and possibly nonconvex optimization problems.

Definition 4 (cf. [184])**.** The directional derivative of h at $\mathbf{x}$ in a direction v is defined as

$$h^\circ(\mathbf{x}, v) \triangleq \limsup_{\mathbf{y} \to \mathbf{x}, t \downarrow 0} \left(\frac{h(\mathbf{y} + tv) - h(\mathbf{y})}{t} \right). \tag{90}$$

The Clarke generalized gradient at $\mathbf{x}$ can then be defined as

$$\partial h(\mathbf{x}) \triangleq \{\xi \in \mathbb{R}^n \mid h^\circ(\mathbf{x}, v) \geq \langle \xi, v \rangle, \quad \forall v \in \mathbb{R}^n\}. \tag{91}$$

In other words, $h^\circ(\mathbf{x}, v) = \sup_{g \in \partial h(\mathbf{x})} \langle g, v \rangle$. □

If h is continuously differentiable at $\mathbf{x}$, we have that the Clarke generalized gradient reduces to the standard gradient, i.e. $\partial h(\mathbf{x}) = \nabla_{\mathbf{x}} h(\mathbf{x})$. If $\mathbf{x}$ is a minimal point of h, then we have that $0 \in \partial h(\mathbf{x})$. For purposes of completeness, we recap some properties of $\partial h(\mathbf{x})$. Recall that if h is locally Lipschitz on an open set $\mathcal{C}$ containing $\mathcal{X}$, then h is differentiable almost everywhere on $\mathcal{C}$ by Rademacher's theorem [184]. Suppose $\mathcal{C}_h$ denotes the set of points where h is not differentiable. We may then recall some properties of Clarke generalized gradients.

Proposition 8 (Properties of Clarke generalized gradients [184])**.** Suppose h is Lipschitz continuous on $\mathbb{R}^n$. Then the following hold.

(i) $\partial h(\mathbf{x})$ is a nonempty, convex, and compact set and $\|g\| \leq L$ for any $g \in \partial h(\mathbf{x})$.

(ii) h is differentiable almost everywhere.

(iii) $\partial h(\mathbf{x})$ is an upper semicontinuous map defined as

$$\partial h(\mathbf{x}) = \operatorname{conv}\left\{ g \mid g = \lim_{k\to\infty} \nabla_{\mathbf{x}} h(\mathbf{x}_k), \mathcal{C}_h \not\ni \mathbf{x}_k \to \mathbf{x} \right\}.$$

We may also define the ϵ-generalized gradient [185] as

$$\partial_\epsilon h(\mathbf{x}) \triangleq \operatorname{conv}\left\{ \xi : \xi \in \partial h(\mathbf{y}), \|\mathbf{x} - \mathbf{y}\| \leq \epsilon \right\}. \tag{92}$$

Under the assumption that h is globally bounded from below and Lipschitz continuous on $\mathcal{X}$, our interest in the nonconvex regimes lies in developing techniques for computing an *approximate* stationary point. For instance, when h is L-smooth, then computing an approximate stationary point in unconstrained regimes such that $\|\nabla_{\mathbf{x}} h(\mathbf{x})\| \leq \epsilon$ requires at most $\mathcal{O}(1/\epsilon^2)$ gradient steps. Much of the prior work in the computation of stationary points of nonconvex and nonsmooth functions is either asymptotic [186, 187] or relies on some structure [59] where the nonconvex part is smooth while the convex part may be closed, convex, and proper. However, the question of computing approximate stationary points for functions that are both nonconvex and nonsmooth has been less studied.

5.2.2 Properties of spherical smoothing of f

We consider an iterative smoothing approach in this paper where a smoothed approximation of h is minimized and the smoothing parameter is progressively reduced. This avenue has a long history, beginning with the efforts by Steklov [188] leading to significant efforts in both convex [189–191] and nonconvex [192] regimes. In this paper, we consider the following smoothing of h, given by h_η where

$$h_\eta(\mathbf{x}) \triangleq \mathbb{E}_{u\in\mathbb{B}}[h(\mathbf{x} + \eta u)], \tag{93}$$

where u is a random vector in the unit ball $\mathbb{B}$, defined as $\mathbb{B} \triangleq \{u \in \mathbb{R}^n \mid \|u\| \leq 1\}$. Throughout, we let $\mathbb{S}$ denote the surface of the ball $\mathbb{B}$, i.e., $\mathbb{S} \triangleq \{v \in \mathbb{R}^n \mid \|v\| = 1\}$. We also let $\eta\mathbb{B}$ and $\eta\mathbb{S}$ denote the ball with radius η and its surface, respectively. Recall that if h is locally Lipschitz over a compact set $\mathcal{X}$, it is globally Lipschitz on $\mathcal{X}$. We may derive the following properties on h_η.

Lemma 24 (**Properties of spherical smoothing**[9])**.** Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is a continuous function and $\eta > 0$ is a given scalar. Let h_η be defined as (93). Then the following hold.
(i) The smoothed function h_η is continuously differentiable over $\mathcal{X}$. In particular, for any $\mathbf{x} \in \mathcal{X}$, we have that

$$\nabla_{\mathbf{x}} h_\eta(\mathbf{x}) = \left(\frac{n}{\eta}\right) \mathbb{E}_{v\in\eta\mathbb{S}}\left[h(\mathbf{x} + v) \frac{v}{\|v\|} \right]. \tag{94}$$

Suppose $h \in C^{0,0}(\mathcal{X}_\eta)$ with parameter L_0. For any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have that (ii) – (iv) hold.

[9]We note that while spherical smoothing have apparently been studied in [193], we did not have access to this text. Part (i) of our Lemma is inspired by Flaxman et al. [194] while other parts either follow in a fashion similar to Gaussian smoothing [192] or are directly proven.

(ii) $|h_\eta(\mathbf{x}) - h_\eta(\mathbf{y})| \leq L_0\|\mathbf{x}-\mathbf{y}\|$.

(iii) $|h_\eta(\mathbf{x}) - h(\mathbf{x})| \leq L_0\eta$.

(iv) $\|\nabla_{\mathbf{x}} h_\eta(\mathbf{x}) - \nabla_{\mathbf{x}} h_\eta(\mathbf{y})\| \leq \frac{L_0 n}{\eta}\|\mathbf{x}-\mathbf{y}\|$.

(v) If h is convex and $h \in C^{0,0}(\mathcal{X}_\eta)$ with parameter L_0, then h_η is convex and satisfies the following for any $\mathbf{x} \in \mathcal{X}$.

$$h(\mathbf{x}) \leq h_\eta(\mathbf{x}) \leq h(\mathbf{x}) + \eta L_0. \tag{95}$$

(vi) If h is convex and $h \in C^{0,0}(\mathcal{X}_\eta)$ with parameter L_0, then $\nabla_x h_\eta(\mathbf{x}) \in \partial_\epsilon h(\mathbf{x})$ where $\epsilon \triangleq \eta L_0$.

(vii) If $h \in C^{1,1}(\mathcal{X}_\eta)$ with constant L_1, then $\|\nabla_x h_\eta(\mathbf{x}) - \nabla_x h(\mathbf{x})\| \leq \eta L_1 n$.

(viii) Suppose $h \in C^{0,0}(\mathcal{X}_\eta)$ with parameter L_0. Let us define for $v \in \eta\mathbb{S}$

$$g_\eta(\mathbf{x}, v) \triangleq \left(\frac{n}{\eta}\right)\frac{(h(\mathbf{x}+v)-h(\mathbf{x}))v}{\|v\|}.$$

Then, for any $\mathbf{x} \in \mathcal{X}$, we have that $\mathbb{E}_{v\in\eta\mathbb{S}}[\|g_\eta(\mathbf{x}, v)\|^2] \leq L_0^2 n^2$.

Proof. (i) We elaborate on the proof sketch provided in [194]. By definition, we have that

$$h_\eta(\mathbf{x}) = \mathbb{E}_{u\in\eta\mathbb{B}}[h(\mathbf{x}+u)] = \int_{\eta\mathbb{B}} h(\mathbf{x}+u)p(u)du.$$

Let $p(u)$ denote the probability density function of u. Since u is uniformly distributed in the ball $\eta\mathbb{B}$, we have that $p(u) = \frac{1}{\mathrm{Vol}(\eta\mathbb{B})}$ for any $u \in \eta\mathbb{B}$. Consequently,

$$h_\eta(\mathbf{x}) = \int_{\eta\mathbb{B}} h(\mathbf{x}+u)p(u)du = \frac{\int_{\eta\mathbb{B}} h(\mathbf{x}+u)du}{\mathrm{Vol}_n(\eta\mathbb{B})}.$$

We may then compute the derivative $\nabla_{\mathbf{x}} h_\eta(\mathbf{x})$ by leveraging Stoke's theorem and by defining $\tilde{p}(v) = \frac{1}{\mathrm{Vol}_{n-1}(\eta\mathbb{S})}$ for all v.

$$\begin{aligned}\nabla_{\mathbf{x}} h_\eta(\mathbf{x}) &= \nabla_{\mathbf{x}}\left[\frac{\int_{\eta\mathbb{B}} h(\mathbf{x}+u)du}{\mathrm{Vol}_n(\eta\mathbb{B})}\right] \overset{\text{Stoke's theorem}}{=} \left[\frac{\int_{\eta\mathbb{S}} h(\mathbf{x}+v)\frac{v}{\|v\|}dv}{\mathrm{Vol}_n(\eta\mathbb{B})}\right] = \left[\frac{\int_{\eta\mathbb{S}} h(\mathbf{x}+v)\frac{v}{\|v\|}dv}{\mathrm{Vol}_n(\eta\mathbb{B})}\right]\frac{\mathrm{Vol}_{n-1}(\eta\mathbb{S})}{\mathrm{Vol}_{n-1}(\eta\mathbb{S})} \\ &= \left[\frac{\int_{\eta\mathbb{S}} h(\mathbf{x}+v)\frac{v}{\|v\|}dv}{\mathrm{Vol}_{n-1}(\eta\mathbb{S})}\right]\frac{\mathrm{Vol}_{n-1}(\eta\mathbb{S})}{\mathrm{Vol}_n(\eta\mathbb{B})} = \left[\int_{\eta\mathbb{S}} h(\mathbf{x}+v)\frac{v}{\|v\|}\tilde{p}(v)dv\right]\frac{n}{\eta} = \frac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[h(\mathbf{x}+v)\frac{v}{\|v\|}\right].\end{aligned}$$

(ii) We have

$$\begin{aligned}|h_\eta(\mathbf{x}) - h_\eta(\mathbf{y})| &= |\mathbb{E}_{u\in\mathbb{B}}[h(\mathbf{x}+\eta u)] - \mathbb{E}_{u\in\mathbb{B}}[h(\mathbf{y}+\eta u)]| \overset{\text{Jensen's ineq.}}{\leq} \mathbb{E}_{u\in\mathbb{B}}[|h(\mathbf{x}+\eta u) - h(\mathbf{y}+\eta u)|] \\ &\overset{h\in C^{0,0}(\mathcal{X}_\eta)}{\leq} \mathbb{E}_{u\in\mathbb{B}}[L_0\|\mathbf{x}-\mathbf{y}\|] = L_0\|\mathbf{x}-\mathbf{y}\|.\end{aligned}$$

(iii) Next, we show that $|h_\eta(\mathbf{x}) - h(\mathbf{x})|$ can be bounded in terms of η and L_0.

$$|h_\eta(\mathbf{x}) - h(\mathbf{x})| = \left|\int_{\eta\mathbb{B}} (h(\mathbf{x}+u) - h(\mathbf{x}))p(u)du\right|$$

$$\leq \int_{\eta\mathbb{B}} |(h(\mathbf{x}+u) - h(\mathbf{x}))|\, p(u)du$$
$$\leq L_0 \int_{\eta\mathbb{B}} \|u\| p(u)du \leq L_0\eta \int_{\eta\mathbb{B}} p(u)du = L_0\eta.$$

(iv) Note that we have $\mathcal{X} + \eta\mathbb{S} \subseteq \mathcal{X} + \eta\mathbb{B}$. Thus, from the definition of $\mathcal{X}_\eta$ and $h \in C^{0,0}(\mathcal{X}_\eta)$, we have $h \in C^{0,0}(\mathcal{X} + \eta\mathbb{S})$. As such, we have

$$\begin{aligned}\|\nabla_{\mathbf{x}} h_\eta(\mathbf{x}) - \nabla_{\mathbf{x}} h_\eta(\mathbf{y})\| &= \left\| \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[h(\mathbf{x}+v)\tfrac{v}{\|v\|}\right] - \tfrac{n}{\eta}\mathbb{E}_{v\in\mathbb{S}}\left[h(\mathbf{y}+v)\tfrac{v}{\|v\|}\right]\right\| \\ &\leq \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\left\|(h(\mathbf{x}+v) - h(\mathbf{y}+v))\tfrac{v}{\|v\|}\right\|\right] \\ &\leq \tfrac{L_0 n}{\eta}\|\mathbf{x}-\mathbf{y}\|\mathbb{E}_{v\in\eta\mathbb{S}}\left[\tfrac{\|v\|}{\|v\|}\right] = \tfrac{L_0 n}{\eta}\|\mathbf{x}-\mathbf{y}\|.\end{aligned}$$

(v) First, note that from $h \in C^{0,0}(\mathcal{X}_\eta)$, we have that $h \in C^{0,0}(\text{int}(\mathcal{X}_\eta))$. Noting that $\text{int}(\mathcal{X}_\eta)$ is an open set, from part (b) of Theorem 3.61 in [59], we have that $\|\tilde{g}\| \leq L_0$ for all $\mathbf{x} \in \text{int}(\mathcal{X}_\eta)$ and $\tilde{g} \in \partial h(\mathbf{x})$. The desired statements then follow from part (a) and part (b) of Lemma 2 [195].
(vi) From part (v), function h_η is convex and $h(\mathbf{y}) + \eta L_0 \geq h_\eta(\mathbf{y})$ for any $\mathbf{y} \in \mathcal{X}$. Thus, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ we have

$$h(\mathbf{y}) + \eta L_0 \geq h_\eta(\mathbf{y}) \geq h_\eta(\mathbf{x}) + \nabla h_\eta(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \geq h(\mathbf{x}) + \nabla h_\eta(\mathbf{x})^T(\mathbf{y}-\mathbf{x}).$$

(vii) Note that we can show that $\int_{\eta\mathbb{S}} vv^T p_v(v)dv = \frac{\eta^2}{n}\mathbf{I}$. We may then express $\nabla_x h(x)$ as

$$\begin{aligned}\nabla_x h(\mathbf{x}) &= \tfrac{n}{\eta^2}\left(\int_{\eta\mathbb{S}} vv^T p_v(v)dv\right)\nabla_x h(\mathbf{x}) = \tfrac{n}{\eta^2}\left(\int_{\eta\mathbb{S}} v^T\nabla_x h(\mathbf{x}) v p_v(v)dv\right) \\ &= \tfrac{n}{\eta}\left(\int_{\eta\mathbb{S}} v^T\nabla_x h(\mathbf{x})\tfrac{v}{\|v\|}p_v(v)dv\right) = \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\left(\nabla_x h(\mathbf{x})^T v\right)\tfrac{v}{\|v\|}\right],\end{aligned}$$

where the third inequality follows from $\|v\| = \eta$ for $v \in \eta\mathbb{S}$. From this relation, part (i), and by recalling that $\frac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[h(\mathbf{x})\frac{v}{\|v\|}\right] = 0$, we can write

$$\begin{aligned}\|\nabla_x h_\eta(\mathbf{x}) - \nabla_x h(\mathbf{x})\| &= \left\|\tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[(h(\mathbf{x}+v) - h(\mathbf{x}))\tfrac{v}{\|v\|}\right] - \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\left(\nabla h(\mathbf{x})^T v\right)\tfrac{v}{\|v\|}\right]\right\| \\ &\leq \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\left|h(\mathbf{x}+v) - h(\mathbf{x}) - \nabla h(\mathbf{x})^T v\right|\tfrac{\|v\|}{\|v\|}\right] \\ &\leq \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[L_1\|v\|^2\right] = n\eta L_1.\end{aligned}$$

(viii) We observe that for any $\mathbf{x}$, $\mathbb{E}_{v\in\eta\mathbb{S}}[\|g_\eta(\mathbf{x},v)\|^2]$ may be bounded as follows.

$$\begin{aligned}\mathbb{E}_{v\in\eta\mathbb{S}}[\|g_\eta(\mathbf{x},v)\|^2] &= \frac{n^2}{\eta^2}\int_{\eta\mathbb{S}} \frac{\|(h(\mathbf{x}+v) - h(\mathbf{x}))v\|^2}{\|v\|^2} p_v(v)dv \\ &\leq \frac{n^2}{\eta^2}\int_{\eta\mathbb{S}} L_0^2\|v\|^2 p_v(v)dv \leq n^2\int_{\eta\mathbb{S}} p_v(v)dv = n^2L_0^2.\end{aligned}$$

□

Remark 5. *Local vs global smoothing: Gaussian smoothing as employed in [192] allows for unbounded random variables as part of the smoothing process. However, this precludes contending with compact regimes which we may require to impose Lipschitzian assumptions. Furthermore, in many settings, the domain of the function is compact and Gaussian smoothing cannot be adopted. Instead, local smoothing requires that the smoothing random variable have compact support. In [190, 195], we examine smoothings based on random variables defined on a cube or a sphere. However, most of the results of the previous Lemma are novel with respect to [195].*

We intend to develop schemes for computing approximate stationary points of (89) by an iterative smoothing scheme. However, this needs formalizing the relationship between the original problem and its smoothed counterpart. Before proceeding, we define ϵ-Clarke generalized gradient of h, denoted by $\partial_\epsilon h(\mathbf{x})$ at $\mathbf{x}$, as follows [185].

$$\partial_\epsilon h(\mathbf{x}) \triangleq \operatorname{conv}\{\xi \mid \xi \in \partial h(\mathbf{y}), \|\mathbf{y}-\mathbf{x}\| \leq \epsilon\}. \tag{96}$$

It was first shown by Goldstein [185] that $\partial_\epsilon h(\mathbf{x})$ is nonempty, compact, and convex set.

Proposition 9. Consider the problem (89) where h is a locally Lipschitz continuous function and $\mathcal{X}$ is a closed, convex, and bounded set in $\mathbb{R}^n$.
(i) For any $\eta > 0$ and any $\mathbf{x} \in \mathbb{R}^n$, $\nabla h_\eta(\mathbf{x}) \in \partial_{2\eta} h(\mathbf{x})$. Furthermore, if $0 \notin \partial h(\mathbf{x})$, then there exists an η such that $\nabla_{\mathbf{x}} h_{\tilde{\eta}}(\mathbf{x}) \neq 0$ for $\tilde{\eta} \in (0, \eta]$.
(ii) For any $\eta > 0$ and any $\mathbf{x} \in \mathcal{X}$,

$$[0 \in \nabla_{\mathbf{x}} h_\eta(\mathbf{x}) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})] \implies [0 \in \partial_{2\eta} h(\mathbf{x}) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})]. \tag{97}$$

Proof. (i) and (ii) represent a constrained counterparts of [196, Prop. 2.2 and Cor. 2.1]. □

Lemma 24 (v) provides a statement that relates the true objective to its smoothed counterpart in convex regimes. This provides an avenue for developing finite-time schemes for computing approximate solutions to the *original problem*. Prop. 9 (iii) provides a relationship in settings where h is locally Lipschitz; in particular, it is shown that if $\mathbf{x}$ satisfies stationarity of the η-smoothed problem, it satisfies a suitable $2\eta-$stationarity property for the original problem.

5.3 An implicit zeroth-order framework

In this section, we present and analyze a zeroth-order framework for contending with (MPEC$^{\text{imp,exp}}$) and (MPEC$^{\text{imp,as}}$). The remainder of this section is organized as follows. In Section 5.3.1, we introduce an implicit zeroth-order scheme that can allow for progressive reduction of the smoothing parameter and can accommodate inexact solutions of the lower-level problem. To address settings where the implicit problem is convex, we derive rate and complexity guarantees for an iteratively smoothed gradient framework in Section 5.3.2 when the lower-level problem is either inexactly or exactly resolved. We extend this framework to accommodate a smoothed accelerated scheme in Section 5.3.5.

5.3.1 An implicit zeroth-order scheme

Since the function $f(\bullet, \mathbf{y}(\bullet))$ is merely Lipschitz continuous, we employ a zeroth-order framework that relies on computing a zeroth-order approximation of the gradient. Given the function $f(\mathbf{x}, \mathbf{y}(\mathbf{x}))$ and a scalar η, we consider a spherical smoothing of $f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x}))$ based on (93), defined as follows:

$$f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x})) \triangleq \mathbb{E}_{u \in \mathbb{B}}[f(\mathbf{x} + \eta u, \mathbf{y}(\mathbf{x} + \eta u))], \tag{G-Smooth}$$

where u is uniformly distributed in the unit ball $\mathbb{B}$. Let $g_\eta(\mathbf{x})$ denote a zeroth-order approximation of the gradient of $f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x}))$. Invoking Lemma 24, one choice for g_η is given by the following for any $\mathbf{x}$.

$$g_\eta(\mathbf{x}) = \left(\frac{n}{\eta}\right) \mathbb{E}_{v \in \eta \mathbb{S}} \left[\frac{(f(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v)) - f(\mathbf{x}, \mathbf{y}(\mathbf{x})))\, v}{\|v\|} \right]. \tag{98}$$

Naturally, $g_\eta(\mathbf{x})$ is challenging to evaluate and a common approach has been in utilizing an unbiased estimate given by $g_\eta(\mathbf{x}, v)$ defined as

$$g_{\eta,N}(\mathbf{x}, v) \triangleq \left(\frac{n}{\eta}\right) \left[\frac{(f(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v)) - f(\mathbf{x}, \mathbf{y}(\mathbf{x})))\, v}{\|v\|} \right]. \tag{99}$$

Given a vector $\mathbf{x}_0 \in \mathcal{X}$, we may employ (99) in constructing a sequence $\{\mathbf{x}_k\}$ where $\mathbf{x}_k$ satisfies the following projected stochastic gradient update.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}} \left[\mathbf{x}_k - \gamma_k g_\eta(\mathbf{x}_k, v_k) \right]. \tag{100}$$

The scheme (100) has been studied for addressing nonsmooth convex and nonconvex optimization problems [192] while unconstrained nonconvex regimes were also examined in [197]. In particular, in the work by Nesterov and Spokoiny [192], zeroth-order randomized smoothing gradient schemes are proposed under a single sample with a fixed smoothing parameter η with the assumption that the smoothing random variable v has a Gaussian distribution. Importantly, a direct adoption of such smoothing schemes to address the hierarchical problems studied in this work is afflicted by several challenges.

(i) *Lack of asymptotic guarantees.* When $\eta > 0$, the scheme generates a sequence that is convergent to an approximate solution, at best. In addition, the choice of η is contingent on accurate estimates of other problem parameters (such as L_0), in the absence of which, η may be chosen to be extremely small. This often afflicts the practical behavior of the scheme. Moreover, employing a fixed η precludes asymptotic convergence to the true counterpart. Instead, in most of our schemes, we employ a mini-batch approximation of $g_\eta(\mathbf{x})$, denoted by $g_{\eta,N}(\mathbf{x})$ and defined as

$$g_{\eta,N}(\mathbf{x}) \triangleq \frac{\sum_{j=1}^{N} g_\eta(\mathbf{x}, v_j)}{N}. \tag{101}$$

Furthermore, we replace a fixed η by a diminishing sequence $\{\eta_k\}$, the resulting iterative smoothing scheme being articulated as follows.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}} \left[\mathbf{x}_k - \gamma_k g_{\eta_k, N_k}(\mathbf{x}_k) \right]. \tag{102}$$

(ii) *Unavailability of exact solutions of* $\mathbf{y}(\mathbf{x})$. Even if $\mathbf{y}(\bullet)$ is a single-valued map requiring the solution of a strongly monotone lower-level problem, computing a solution to this problem is not necessarily cheap. As a consequence, our scheme needs to account for errors in the computation of $g_{\eta_k}(\mathbf{x}_k)$, denoted by $\tilde{\epsilon}_k$. As a consequence, the resulting scheme is defined as follows.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}} \left[\mathbf{x}_k - \gamma_k (g_{\eta_k, N_k}(\mathbf{x}_k) + \tilde{\epsilon}_k) \right]. \tag{103}$$

For instance, when considering problems of the form (SMPEC$^{\text{exp}}$), exact solutions of $\mathbf{y}(\mathbf{x}_k)$ are generally unavailable in finite time. Instead, one can take j_k steps of a standard projection scheme.

$$\mathbf{y}_{j+1} := \Pi_{\mathcal{Y}} \left[\mathbf{y}_j - \beta_j \bar{F}(\mathbf{x}_k, \mathbf{y}_j) \right], \qquad j = 1, \cdots, j_k, \tag{104}$$

where $\bar{F}(\mathbf{x}_k, \mathbf{y}_j) \triangleq \frac{\sum_{\ell=1}^{M_j} G(\mathbf{x}_k, \mathbf{y}_j, \omega_j)}{M_j}$. When using such a variance-reduced scheme, $\log(1/\epsilon_k)$ steps of the above scheme are required to obtain an ϵ_k-solution of $\mathbf{y}_k$.

(iii) *Bias in* $\tilde{\epsilon}_k$. A key issue that arises from (ii) emerges in the form of bias. In particular, $g_{\eta_k,N_k}(\mathbf{x}_k)+\tilde{\epsilon}_k$ is not necessarily an unbiased estimator of $g_{\eta_k}(\mathbf{x}_k)$. Further, it remains unclear how the bias and variance of $g_{\eta_k,N_k}(\mathbf{x}_k)+\tilde{\epsilon}_k$ propagate through this framework (103)-(104) as γ_k, η_k, and N_k are updated iteratively in the outer loop (103). Consequently, in the development of the inexact smoothing scheme (103)-(104), it remains critical to design prescribed stepsize, smoothing, and sample-size sequences to control the accuracy of the estimator $g_{\eta_k,N_k}(\mathbf{x}_k)+\tilde{\epsilon}_k$ and consequently, ascertain the convergence of the generated iterate to an optimal solution of the underlying MPEC. This concern will be examined in detail in the subsequent sections.

5.3.2 Convex regimes

In this subsection, we consider resolving the implicit formulations when the implicit function is convex. As we pointed out earlier, the convexity of the implicit problem often holds in practice (cf. [181–183]). We first consider the inexact case where the exact value of $\mathbf{y}(\bullet)$ is not necessarily available. We then specialize our statements to settings where exact solutions of lower-level problems can be employed. The next result is used in deriving the convergence rate results.

Lemma 25 (cf. Lemma 10 in [180] and Lemma 2.14 in [198])**.** Let ℓ and N be arbitrary integers where $0 \leq \ell \leq N-1$. The following hold.

(a) $\ln\left(\frac{N+1}{\ell+1}\right) \leq \sum_{k=\ell}^{N-1} \frac{1}{k+1} \leq \frac{1}{\ell+1} + \ln\left(\frac{N}{\ell+1}\right)$.

(b) If $0 \leq \alpha < 1$, then for any $N \geq 2^{\frac{1}{1-\alpha}} - 1$, we have $\frac{(N+1)^{1-\alpha}}{2(1-\alpha)} \leq \sum_{k=0}^{N} \frac{1}{(k+1)^{\alpha}} \leq \frac{(N+1)^{1-\alpha}}{1-\alpha}$.

5.3.3 An inexact zeroth-order scheme

We now delve into developing and analyzing an inexact zeroth-order method for resolving the implicit variant of (SMPEC$^{\text{exp}}$), i.e. (MPEC$^{\text{imp,exp}}$) where the lower-level problem is complicated by the presence of expectation-valued maps, i.e., F is defined as (86) and satisfies Assumption 13 (a.iii). In such an instance, obtaining $\mathbf{y}(\mathbf{x})$ is impossible in finite time unless the expectation can be tractably resolved. Instead, by employing stochastic approximation (SA) methods for addressing the lover-level problem, we consider the case where we have access to an approximate solution $\mathbf{y}_{\epsilon_k}(\mathbf{x}_k)$ such that

$$\mathbb{E}[\|\mathbf{y}_{\epsilon_k}(\mathbf{x}_k) - \mathbf{y}(\mathbf{x}_k)\|^2 \mid \mathbf{x}_k] \leq \epsilon_k, \text{ where } \mathbf{y}(\mathbf{x}_k) \in \text{SOL}(\mathcal{Y}, F(\mathbf{x}_k, \bullet)). \tag{105}$$

As a consequence, we may define an inexact zeroth-order gradient mapping $g_{\eta,\epsilon}(\mathbf{x}, v)$ as follows.

$$g_{\eta,\epsilon}(\mathbf{x}, v) := \frac{n(f(\mathbf{x}+v, \mathbf{y}_{\epsilon}(\mathbf{x}+v)) - f(\mathbf{x}, \mathbf{y}_{\epsilon}(\mathbf{x}))v}{\|v\|\eta}, \tag{106}$$

where $v \in \eta\mathbb{S}$ and $\mathbf{y}_{\epsilon_k}(\mathbf{x}_k)$ is an output of a variance-reduced stochastic approximation scheme. The outline of the proposed zeroth-order solver (`ZSOL`) is presented in Algorithm 10 while an inexact solution of $\mathbf{y}(\mathbf{x})$ is computed by Algorithm 11. We impose the following assumptions on v_k and the lower-level evaluations $G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t})$.

Assumption 15. Given a sequence $\{\eta_k\}$, let $v_k \in \mathbb{R}^n$ be generated randomly and independently from $\eta_k\mathbb{S}$ for all $k \geq 0$.

Assumption 16. Let the following hold for all $k \geq 0$, $t \geq 0$, and $1 \leq \ell \leq M_t$.
(a) The random realizations $\omega_{\ell,t}$ are independent and identically distributed.
(b) $\mathbb{E}[G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t}) \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] = F(\hat{\mathbf{x}}_k, \mathbf{y}_t)$.
(c) $\mathbb{E}[\|G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t}) - F(\mathbf{x}_k, \mathbf{y}_t)\|^2 \mid \mathbf{x}_k, \mathbf{y}_t] \leq \nu_{\mathbf{y}}^2 \|\mathbf{y}_t\|^2 + \nu_G^2$ for some $\nu_{\mathbf{y}}, \nu_G > 0$.

Algorithm 10 `ZSOL`: Inexact zeroth-order method for the convex case

input: Given $\mathbf{x}_0 \in \mathcal{X}$, $\bar{\mathbf{x}}_0 := \mathbf{x}_0$, stepsize sequence $\{\gamma_k\}$, smoothing parameter sequence $\{\eta_k\}$, inexactness sequence $\{\epsilon_k\}$, $r \in [0, 1)$, and $S_0 := \gamma_0^r$
for $k = 0, 1, \ldots, K-1$ **do**
  Call Algorithm 11 to obtain $\mathbf{y}_{\epsilon_k}(\mathbf{x}_k) := \mathbf{y}_{t_k}$
  Generate a random realization $v_k \in \eta_k \mathbb{S}$
  Call Algorithm 11 to obtain $\mathbf{y}_{\epsilon_k}(\mathbf{x}_k + v_k) := \mathbf{y}_{t_k}$
  Evaluate the inexact zeroth-order gradient approximation as follows.

$$g_{\eta_k,\epsilon_k}(\mathbf{x}_k, v_k) := \frac{n\left(f(\mathbf{x}_k + v_k, \mathbf{y}_{\epsilon_k}(\mathbf{x}_k + v_k)) - f(\mathbf{x}_k, \mathbf{y}_{\epsilon_k}(\mathbf{x}_k))\right) v_k}{\|v_k\| \eta_k}$$

  Update $\mathbf{x}_k$ as follows. $\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k,\epsilon_k}(\mathbf{x}_k, v_k)\right]$
  Update the averaged iterate as follows.

$$S_{k+1} := S_k + \gamma_{k+1}^r, \qquad \bar{x}_{k+1} := \frac{S_k \bar{x}_k + \gamma_{k+1}^r \mathbf{x}_{k+1}}{S_{k+1}} \tag{107}$$

end for

Algorithm 11 Variance-reduced SA method for stochastic VI in the lower-level

input: An arbitrary $\mathbf{y}_0 \in \mathcal{Y}$, vector $\hat{\mathbf{x}}_k$, scalar $\rho \in (0, 1)$, stepsize $\alpha > 0$, mini-batch sequence $\{M_t\}$, integer k, and scalar $\tau > 0$
Compute $t_k := \lceil \tau \ln(k+1) \rceil$
for $t = 0, 1, \ldots, t_k - 1$ **do**
  Generate random realizations of the stochastic mapping $G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t})$ for $\ell = 1, \ldots, M_t$
  Update $\mathbf{y}_t$ as follows.

$$\mathbf{y}_{t+1} := \Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \alpha \frac{\sum_{\ell=1}^{M_t} G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t})}{M_t}\right]$$

end for
Return $\mathbf{y}_{t_k}$

Before analyzing `ZSOL`, we review the properties of the exact zeroth-order gradient denoted by $g_\eta(\mathbf{x}, v)$ and show that the exact zeroth-order gradient is an unbiased estimator of the gradient of the smoothed implicit function. We then derive a bound on the second moment of this stochastic gradient under the assumption that the implicit function is Lipschitz.

Lemma 26. Consider (MPEC$^{\text{imp,exp}}$) and suppose Assumption 13 holds. Given $\mathbf{x} \in \mathcal{X}$ and $\eta > 0$, consider the following zeroth-order mapping for $v \in \eta \mathbb{S}$ and $k \geq 0$.

$$g_\eta(\mathbf{x}, v) \triangleq \left(\frac{n}{\eta}\right) \frac{\left(f(\mathbf{x} + v, \mathbf{y}(\mathbf{x} + v)) - f(\mathbf{x}, \mathbf{y}(\mathbf{x}))\right) v}{\|v\|}. \tag{108}$$

Then, $\nabla f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x})) = \mathbb{E}[g_\eta(\mathbf{x}, v) \mid \mathbf{x}]$ and $\mathbb{E}[\|g_\eta(\mathbf{x}, v)\|^2 \mid \mathbf{x}] \leq L_0^2 n^2$ almost surely for all $k \geq 0$.

Proof. The two results follow from Lemma 24 (i) and (viii), respectively. □

We are now ready to present the properties of the inexact zeroth-order gradient mapping.

Lemma 27 (Properties of the inexact zeroth-order gradient). Suppose Assumption 13 holds. Let $g_{\eta,\epsilon}(\mathbf{x}, v)$ be defined as (106) for $v \in \eta\mathbb{S}$ for $\eta, \epsilon > 0$. Suppose $\mathbb{E}[\|\mathbf{y}_\epsilon(\mathbf{x}) - \mathbf{y}(\mathbf{x})\|^2 \mid \mathbf{x}] \leq \epsilon$ almost surely for any $\mathbf{x} \in \mathcal{X}$. Then the following hold.
(a) $\mathbb{E}[\|g_{\eta,\epsilon}(\mathbf{x}, v)\|^2 \mid \mathbf{x}] \leq 3n^2 \left(\frac{2\tilde{L}_0^2 \epsilon}{\eta^2} + L_0^2\right)$, almost surely.
(b) $\mathbb{E}\left[\|g_{\eta,\epsilon}(\mathbf{x}, v) - g_\eta(\mathbf{x}, v)\|^2 \mid \mathbf{x}\right] \leq \frac{4\tilde{L}_0^2 n^2 \epsilon}{\eta^2}$, almost surely.

Proof. **(a)** Adding and subtracting $g_\eta(\mathbf{x}, v)$ defined by (101), we obtain from (106)

$$
\begin{aligned}
&\|g_{\eta,\epsilon}(\mathbf{x}, v)\| \\
&= \left\| \left(\frac{n(f(\mathbf{x}+v, \mathbf{y}_\epsilon(\mathbf{x}+v)) - f(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v))v}{\|v\|\eta} \right) + g_\eta(\mathbf{x}, v) + \left(\frac{n(f(\mathbf{x}, \mathbf{y}(\mathbf{x})) - f(\mathbf{x}, \mathbf{y}_\epsilon(\mathbf{x})))v}{\|v\|\eta} \right) \right\| \\
&\leq \left\| \left(\frac{n(f(\mathbf{x}+v, \mathbf{y}_\epsilon(\mathbf{x}+v)) - f(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v)))v}{\|v\|\eta} \right) \right\| + \|g_\eta(\mathbf{x}, v)\| \\
&+ \left\| \left(\frac{n(f(\mathbf{x}, \mathbf{y}(\mathbf{x})) - f(\mathbf{x}, \mathbf{y}_\epsilon(\mathbf{x}))v}{\|v\|\eta} \right) \right\| \\
&\leq \left(\frac{\|f(\mathbf{x}+v, \mathbf{y}_\epsilon(\mathbf{x}+v)) - f(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v))\|n\|v\|}{\|v\|\eta} \right) + \|g_\eta(\mathbf{x}, v)\| \\
&+ \left(\frac{\|f(\mathbf{x}, \mathbf{y}(\mathbf{x})) - f(\mathbf{x}, \mathbf{y}_\epsilon(\mathbf{x}))\|n\|v\|}{\eta\|v\|} \right) \\
&\leq \left(\frac{\tilde{L}_0 \|\mathbf{y}_\epsilon(\mathbf{x}+v) - \mathbf{y}(\mathbf{x}+v)\|n}{\eta} \right) + \|g_\eta(\mathbf{x}, v)\| + \left(\frac{\tilde{L}_0 \|\mathbf{y}_\epsilon(\mathbf{x}) - \mathbf{y}(\mathbf{x})\|n}{\eta} \right).
\end{aligned}
$$

Invoking Lemma 26, we may then bound the second moment of $\|g_{\eta,\epsilon}(\mathbf{x}, u)\|$ as follows.

$$
\begin{aligned}
\mathbb{E}[\|g_{\eta,\epsilon}(\mathbf{x}, v)\|^2 \mid \mathbf{x}] &\leq 3\mathbb{E}\left[\left(\frac{\tilde{L}_0^2 n^2 \|\mathbf{y}_\epsilon(\mathbf{x}+v) - \mathbf{y}(\mathbf{x}+v)\|^2}{\eta^2} \right) \mid \mathbf{x} \right] + 3\mathbb{E}\left[\|g_\eta(\mathbf{x}, v)\|^2 \mid \mathbf{x} \right] \\
&\quad + 3\mathbb{E}\left[\left(\frac{\tilde{L}_0^2 n^2 \|\mathbf{y}_\epsilon(\mathbf{x}) - \mathbf{y}(\mathbf{x})\|^2}{\eta^2} \right) \mid \mathbf{x} \right] \leq 6\left(\frac{\tilde{L}_0^2 n^2 \epsilon}{\eta^2} \right) + 3L_0^2 n^2, \text{ a.s.}
\end{aligned}
$$

(b) Next, we derive a bound on $\|g_{\eta,\epsilon}(\mathbf{x}, v) - g_\eta(\mathbf{x}, v)\|$.

$$
\begin{aligned}
&\|g_{\eta,\epsilon}(\mathbf{x}, v) - g_\eta(\mathbf{x}, v)\| \\
&= \left\| \left(\frac{n(f(\mathbf{x}+v, \mathbf{y}_\epsilon(\mathbf{x}+v)) - f(\mathbf{x}, \mathbf{y}_\epsilon(\mathbf{x})))v}{\|v\|\eta} \right) - \left(\frac{n(f(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v)) - f(\mathbf{x}, \mathbf{y}(\mathbf{x}))v}{\|v\|\eta} \right) \right\| \\
&\leq \left\| \left(\frac{n(f(\mathbf{x}+v, \mathbf{y}_\epsilon(\mathbf{x}+v)) - f(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v)))v}{\|v\|\eta} \right) \right\| + \left\| \left(\frac{f(\mathbf{x}, \mathbf{y}_\epsilon(\mathbf{x})) - f(\mathbf{x}, \mathbf{y}(\mathbf{x}))}{\|v\|\eta} \right) v \right\| \\
&\leq \left(\frac{\tilde{L}_0 n \|\mathbf{y}_\epsilon(\mathbf{x}+v) - \mathbf{y}(\mathbf{x}+v)\|}{\eta} \right) + \left(\frac{\tilde{L}_0 n \|\mathbf{y}_\epsilon(\mathbf{x}) - \mathbf{y}(\mathbf{x})\|}{\eta} \right).
\end{aligned}
$$

It follows that $\mathbb{E}\left[\|g_{\eta,\epsilon}(\mathbf{x}, v) - g_\eta(\mathbf{x}, v)\|^2 \mid \mathbf{x}\right] \leq \frac{4\tilde{L}_0^2 n^2 \epsilon}{\eta^2}$. □

We make use of the following results in the convergence and rate analysis.

Lemma 28 (Lemma 2.11 in [198])**.** Let $\{\bar{\mathbf{x}}_k\}$ be generated by Algorithm 10. Let us define the weights $\alpha_{k,N} \triangleq \frac{\gamma_k^r}{\sum_{j=0}^N \gamma_j^r}$ for $k \in \{0,\ldots,N\}$ and $N \geq 0$. Then, for any $N \geq 0$, we have $\bar{\mathbf{x}}_N = \sum_{k=0}^N \alpha_{k,N}\mathbf{x}_k$. Furthermore, when $\mathcal{X}$ is a convex set, we have $\bar{\mathbf{x}}_N \in \mathcal{X}$.

Lemma 29 (Theorem 6, page 75 in [199])**.** Let $\{u_t\} \subset \mathbb{R}^n$ denote a sequence of vectors where $\lim_{t\to\infty} u_t = \hat{u}$. Also, let $\{\alpha_k\}$ denote a sequence of strictly positive scalars such that $\sum_{k=0}^\infty \alpha_k = \infty$. Suppose $v_k \in \mathbb{R}^n$ is defined by $v_k \triangleq \frac{\sum_{t=0}^k \alpha_t u_t}{\sum_{t=0}^k \alpha_t}$ for all $k \geq 0$. Then, $\lim\limits_{k\to\infty} v_k = \hat{u}$.

We are now in a position to develop rate and complexity statements for Algorithms 10–11. The algorithm parameters for both schemes are defined next.

Definition 5 (Parameters for Algorithms 10–11)**.** Let the stepsize and smoothing sequence in Algorithm 10 be given by $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ and $\eta_k := \frac{\eta_0}{\sqrt{k+1}}$, respectively for all $k \geq 0$ where γ_0 and η_0 are strictly positive. In Algorithm 11, suppose $\alpha \leq \frac{\mu_F}{2L_F^2}$, $M_t := \lceil M_0\rho^{-t}\rceil$ for $t \geq 0$ for some $0 < \rho < 1$ where $M_0 \geq \frac{2\nu_{\mathbf{y}}^2}{L_F^2}$. Let $t_k := \lceil \tau \ln(k+1)\rceil$ where $\tau \geq \frac{-2}{\ln(\max\{1-\mu_F\alpha,\rho\})}$. Finally, suppose $r \in [0,1)$ is an arbitrary scalar.

Theorem 30 (**Rate statements and complexity results for Algorithms 10–11).** Consider the sequence $\{\bar{\mathbf{x}}_k\}$ generated by applying Algorithm 10 on (MPEC$^{\text{imp,exp}}$). Suppose Assumptions 13– 16 hold and algorithm parameters are defined by Def. 5.
(a) Suppose $\hat{\mathbf{x}}_k \in \mathcal{X}$ and let $\{\mathbf{y}_{t_k}\}$ be the sequence generated by Algorithm 11. Then for suitably defined $\tilde{d} < 1$ and $B > 0$, the following holds for $t_k \geq 1$.

$$\mathbb{E}[\|\mathbf{y}_{t_k} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \leq \epsilon_k \triangleq B\tilde{d}^{t_k}.$$

(b) For all $K \geq 2^{\frac{1}{1-r}} - 1$, we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K, \mathbf{y}(\bar{\mathbf{x}}_K))\right] - f^* \leq (2-r)\left(\frac{D_{\mathcal{X}}}{\gamma_0} + \frac{2D_{\mathcal{X}}\gamma_0 + \frac{(4+6\gamma_0^2)B}{\eta_0^2\gamma_0} + 2\eta_0 L_0 + 3n^2L_0^2\gamma_0}{1-r}\right)\frac{1}{\sqrt{K+1}}.$$

(c) Suppose $\gamma_0 = \mathcal{O}(\frac{1}{nL_0})$ and $r = 0$. Let $\epsilon > 0$ be an arbitrary scalar and K_ϵ be such that $\mathbb{E}\left[f(\bar{\mathbf{x}}_{K_\epsilon}, \mathbf{y}(\bar{\mathbf{x}}_{K_\epsilon}))\right] - f^* \leq \epsilon$. Then,
(c-1) The total number of upper-level projection steps on $\mathcal{X}$ is $K_\epsilon = \mathcal{O}\left(n^2L_0^2\epsilon^{-2}\right)$.
(c-2) The total sample complexity of upper-level evaluations of $\mathbf{y}(\bullet)$ is $\mathcal{O}\left(n^2L_0^2\epsilon^{-2}\right)$.
(c-3) The total number of lower-level projection steps on $\mathcal{Y}$ is $\mathcal{O}\left(n^2L_0^2\epsilon^{-2}\ln\left(nL_0\epsilon^{-1}\right)\right)$.
(c-4) The total sample complexity of lower-level is $\mathcal{O}\left(n^{2\bar{\tau}}L_0^{2\bar{\tau}}\epsilon^{-2\bar{\tau}}\right)$ where $\bar{\tau} \geq 1 - \tau\ln(\rho)$.

Proof. **(a)** We denote the history generated by Algorithm 11 by $\mathcal{F}_t \triangleq \cup_{j=0}^{t-1}\cup_{\ell=1}^{M_j}\{\omega_{\ell,j}\}$ for $t \geq 1$, and $\mathcal{F}_0 \triangleq \cup_{\ell=1}^{M_0}\{\omega_{\ell,0}\}$. Let us define $\bar{F}(\hat{\mathbf{x}}_k, \mathbf{y}_t) \triangleq \frac{\sum_{\ell=1}^{M_t} G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t})}{M_t}$ for $t \geq 0$ and $k \geq 0$. We also define the errors $\Delta_t \triangleq \bar{F}(\hat{\mathbf{x}}_k, \mathbf{y}_t) - F(\hat{\mathbf{x}}_k, \mathbf{y}_t)$ for $t \geq 0$. Next, we estimate a bound on the term $\mathbb{E}[\|\Delta_t\|^2 \mid \mathcal{F}_t]$. From Assumption 16 we have

$$\mathbb{E}[\|\Delta_t\|^2 \mid \mathcal{F}_t] = \mathbb{E}\left[\left\|\frac{\sum_{\ell=1}^{M_t}(G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t}) - F(\hat{\mathbf{x}}_k, y_t))}{M_t}\right\|^2 \mid \mathcal{F}_t\right]$$

$$= \tfrac{1}{M_t^2}\mathbb{E}\left[\sum_{\ell=1}^{M_t}\|G(\hat{\mathbf{x}}_k,\mathbf{y}_t,\omega_{\ell,t})-F(\hat{\mathbf{x}}_k,y_t)\|^2 \mid \mathcal{F}_t\right] \leq \tfrac{\nu_{\mathbf{y}}^2\|\mathbf{y}_t\|^2+\nu_G^2}{M_t}. \tag{109}$$

From $\mathbf{y}(\hat{\mathbf{x}}_k)\in \mathrm{SOL}(\mathcal{Y},F(\hat{\mathbf{x}}_k,\bullet))$, we have $\mathbf{y}(\hat{\mathbf{x}}_k)=\Pi_{\mathcal{Y}}\left[\mathbf{y}(\hat{\mathbf{x}}_k)-\alpha F(\hat{\mathbf{x}}_k,\mathbf{y}(\hat{\mathbf{x}}_k))\right]$ for any $\alpha>0$. We have

$$\begin{aligned}
&\|\mathbf{y}_{t+1}-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2 = \|\Pi_{\mathcal{Y}}\left[\mathbf{y}_t-\alpha\bar{F}(\hat{\mathbf{x}}_k,\mathbf{y}_t)\right]-\Pi_{\mathcal{Y}}\left[\mathbf{y}(\hat{\mathbf{x}}_k)-\alpha F(\hat{\mathbf{x}}_k,\mathbf{y}(\hat{\mathbf{x}}_k))\right]\|^2\\
&\leq \|\mathbf{y}_t-\alpha\bar{F}(\hat{\mathbf{x}}_k,\mathbf{y}_t)-\mathbf{y}(\hat{\mathbf{x}}_k)+\alpha F(\hat{\mathbf{x}}_k,\mathbf{y}(\hat{\mathbf{x}}_k))\|^2\\
&= \|\mathbf{y}_t-\alpha F(\hat{\mathbf{x}}_k,\mathbf{y}_t)-\alpha\Delta_t-\mathbf{y}(\hat{\mathbf{x}}_k)+\alpha F(\hat{\mathbf{x}}_k,\mathbf{y}(\hat{\mathbf{x}}_k))\|^2\\
&= \|\mathbf{y}_t-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\alpha^2\|F(\hat{\mathbf{x}}_k,\mathbf{y}_t)-F(\hat{\mathbf{x}}_k,\mathbf{y}(\hat{\mathbf{x}}_k))\|^2+\alpha^2\|\Delta_t\|^2\\
&-2\alpha(\mathbf{y}_t-\mathbf{y}(\hat{\mathbf{x}}_k))^T(F(\hat{\mathbf{x}}_k,\mathbf{y}_t)-F(\hat{\mathbf{x}}_k,\mathbf{y}(\hat{\mathbf{x}}_k)))\\
&-2\alpha(\mathbf{y}_t-\mathbf{y}(\hat{\mathbf{x}}_k)-\alpha F(\hat{\mathbf{x}}_k,\mathbf{y}_t)+\alpha F(\hat{\mathbf{x}}_k,\mathbf{y}(\hat{\mathbf{x}}_k)))^T\Delta_t.
\end{aligned}$$

Taking conditional expectations in the preceding relation, using (109), and invoking the strong monotonicity and Lipschitzian property of the mapping F in Assumption 13, we obtain

$$\mathbb{E}[\|\mathbf{y}_{t+1}-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2\mid\mathcal{F}_t]\leq\left(1-2\mu_F\alpha+\alpha^2L_F^2\right)\|\mathbf{y}_t-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\tfrac{\nu_{\mathbf{y}}^2\|\mathbf{y}_t\|^2+\nu_G^2}{M_t}\alpha^2.$$

Taking expectations on both sides, we obtain

$$\begin{aligned}
\mathbb{E}[\|\mathbf{y}_{t+1}-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2] &\leq \left(1-2\mu_F\alpha+\alpha^2L_F^2\right)\mathbb{E}[\|\mathbf{y}_t-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2]+\tfrac{\nu_{\mathbf{y}}^2\mathbb{E}[\|\mathbf{y}_t-\mathbf{y}(\hat{\mathbf{x}}_k)+\mathbf{y}(\hat{\mathbf{x}}_k)\|^2]+\nu_G^2}{M_t}\alpha^2\\
&\leq \left(1-2\mu_F\alpha+\alpha^2L_F^2+\tfrac{2\nu_{\mathbf{y}}^2}{M_0}\alpha^2\right)\mathbb{E}[\|\mathbf{y}_t-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2]+\tfrac{2\nu_{\mathbf{y}}^2\|\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\nu_G^2}{M_t}\alpha^2.
\end{aligned}$$

Let $\lambda\triangleq 1-2\mu_F\alpha+\alpha^2L_F^2+\frac{2\nu_{\mathbf{y}}^2}{M_0}\alpha^2$ and recall that $\Lambda_t\triangleq\frac{2\nu_{\mathbf{y}}^2\|\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\nu_G^2}{M_t}\alpha^2$ for $t\geq 0$. Note that since $M_0\geq\frac{2\nu_{\mathbf{y}}^2}{L_F^2}$ and that $\alpha\leq\frac{\mu_F}{2L_F^2}$, we have $\lambda\leq 1-\mu_F\alpha<1$. We obtain for any $t\geq 0$

$$\begin{aligned}
\mathbb{E}[\|\mathbf{y}_{t+1}-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2] &\leq \lambda^{t+1}\|\mathbf{y}_0-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\sum_{j=0}^{t}\lambda^{t-j}\Lambda_j\\
&\leq \lambda^{t+1}\|\mathbf{y}_0-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\Lambda_0(\max\{\lambda,\rho\})^t\sum_{j=0}^{t}\left(\tfrac{\min\{\lambda,\rho\}}{\max\{\lambda,\rho\}}\right)^{t-j}\\
&\leq \lambda^{t+1}\|\mathbf{y}_0-\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\tfrac{\Lambda_0(\max\{\lambda,\rho\})^t}{1-(\min\{\lambda,\rho\}/\max\{\lambda,\rho\})}\leq B\tilde{d}^{t+1}.
\end{aligned}$$

where $\tilde{d}\triangleq\max\{\lambda,\rho\}$ and $B\triangleq\sup_{\mathbf{y}\in\mathcal{Y}}\|\mathbf{y}-\mathbf{y}_0\|^2+\frac{\Lambda_0}{\max\{\lambda,\rho\}-\min\{\lambda,\rho\}}$. Note that without loss of generality, we assume $\rho\neq\lambda$.

(b) Note that from the compactness of the set $\mathcal{X}$ and the continuity of the implicit function, the set $\mathcal{X}^*$ is nonempty. Let $\mathbf{x}^*\in\mathcal{X}$ be an arbitrary optimal solution. We have that

$$\begin{aligned}
\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 &= \|\Pi_{\mathcal{X}}\left[\mathbf{x}_k-\gamma_k g_{\eta_k,\epsilon_k}(\mathbf{x}_k,v_k)\right]-\Pi_{\mathcal{X}}\left[\mathbf{x}^*\right]\|^2\leq\|\mathbf{x}_k-\gamma_k g_{\eta_k,\epsilon_k}(\mathbf{x}_k,v_k)-\mathbf{x}^*\|^2\\
&= \|\mathbf{x}_k-\mathbf{x}^*\|^2-2\gamma_k(\mathbf{x}_k-\mathbf{x}^*)^Tg_{\eta_k,\epsilon_k}(\mathbf{x}_k,v_k)+\gamma_k^2\|g_{\eta_k,\epsilon_k}(\mathbf{x}_k,v_k)\|^2\\
&= \|\mathbf{x}_k-\mathbf{x}^*\|^2-2\gamma_k(\mathbf{x}_k-\mathbf{x}^*)^T(g_{\eta_k}(\mathbf{x}_k,v_k)+w_k)+\gamma_k^2\|g_{\eta_k,\epsilon_k}(\mathbf{x}_k,v_k)\|^2,
\end{aligned}$$

where we define $w_k\triangleq g_{\eta_k,\epsilon_k}(\mathbf{x}_k,v_k)-g_{\eta_k}(\mathbf{x}_k,v_k)$. Taking conditional expectations on the both sides, and invoking Lemma 26 and Lemma 27 (a), we obtain

$$\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2\mid\mathcal{F}_k\right]\leq\|\mathbf{x}_k-\mathbf{x}^*\|^2-2\gamma_k(\mathbf{x}_k-\mathbf{x}^*)^T\nabla f_{\eta_k}(\mathbf{x}_k,\mathbf{y}(\mathbf{x}_k))$$

$$- 2\gamma_k \mathbb{E}\left[(\mathbf{x}_k - \mathbf{x}^*)^T w_k \mid \mathcal{F}_k\right] + 3n^2\gamma_k^2 \left(\tfrac{2\tilde{L}_0^2 \epsilon_k}{\eta_k^2} + L_0^2\right).$$

Invoking the convexity of f_{η_k}, bounding $-2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T w_k$, and rearranging the terms, we obtain

$$\begin{aligned} 2\gamma_k \left(f_{\eta_k}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) - f_{\eta_k}(\mathbf{x}^*, \mathbf{y}(\mathbf{x}^*))\right) &\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathcal{F}_k\right] \\ &\quad + \gamma_k^2 \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \mathbb{E}\left[\|w_k\|^2 \mid \mathcal{F}_k\right] + 3n^2\gamma_k^2 \left(\tfrac{2\tilde{L}_0^2 \epsilon_k}{\eta_k^2} + L_0^2\right). \end{aligned}$$

From Lemma 27 (b) we obtain

$$\begin{aligned} 2\gamma_k \left(f_{\eta_k}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) - f_{\eta_k}(\mathbf{x}^*, \mathbf{y}(\mathbf{x}^*))\right) &\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathcal{F}_k\right] \\ &\quad + \gamma_k^2 \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \tfrac{4\tilde{L}_0^2 n^2 \epsilon_k}{\eta_k^2} + 3n^2\gamma_k^2 \left(\tfrac{2\tilde{L}_0^2 \epsilon_k}{\eta_k^2} + L_0^2\right). \end{aligned}$$

From Lemma 24 (v) we have that $f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) \leq f_{\eta_k}(\mathbf{x}^*, \mathbf{y}(\mathbf{x}^*)) \leq f^* + \eta_k L_0$. From the preceding two inequalities we obtain

$$\begin{aligned} &2\gamma_k \left(\mathbb{E}\left[f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right] - f^*\right) \\ &\leq \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] + \gamma_k^2 \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2] + (4 + 6\gamma_0^2)\tfrac{\tilde{L}_0^2 n^2 \epsilon_k}{\eta_k^2} + 2\gamma_k \eta_k L_0 + 3n^2 L_0^2 \gamma_k^2. \end{aligned}$$

Next, we derive a bound on $\frac{\epsilon_k}{\eta_k^2}$. From part (a) and the update rule of η_k we have

$$\tfrac{\epsilon_k}{\eta_k^2} = \left(\tfrac{\epsilon_k}{\eta_k^2 \gamma_k^2}\right)\gamma_k^2 = \left(\tfrac{(\max\{\lambda,\rho\})^{t_k} B(k+1)^2}{\eta_0^2 \gamma_0^2}\right)\gamma_k^2. \tag{110}$$

Note that from $\alpha \leq \frac{\mu_F}{2L_F^2}$ and $M_0 \geq \frac{2\nu_{\mathbf{y}}^2}{L_F^2}$, we have $\lambda \leq 1 - \mu_F \alpha$. Thus, we have $\tau \geq \frac{-2}{\ln(\max\{1-\mu_F\alpha, \rho\})} \geq \frac{-2}{\ln(\max\{\lambda,\rho\})}$. From $t_k := \lceil \tau \ln(k+1) \rceil \geq \tau \ln(k+1)$ and $\tau \geq \frac{-2}{\ln(\max\{\lambda,\rho\})}$ we have that

$$(\max\{\lambda,\rho\})^{t_k} (k+1)^2 \leq \left((\max\{\lambda,\rho\})^{\tau} \mathrm{e}^2\right)^{\ln(k+1)} \leq (\max\{\lambda,\rho\})^{\tau} \mathrm{e}^2 \leq 1.$$

This relation and (110) imply that $\frac{\epsilon_k}{\eta_k^2} \leq \left(\frac{B}{\eta_0^2\gamma_0^2}\right)\gamma_k^2$. Also, note that since $\mathcal{X}$ is bounded, there exists a scalar $D_{\mathcal{X}} \triangleq \frac{1}{2}\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}^*\|^2$. Therefore, we obtain

$$2\gamma_k \left(\mathbb{E}\left[f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right] - f^*\right) \leq \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] + 2\gamma_k^2 \theta_0,$$

where $\theta_0 \triangleq D_{\mathcal{X}} + \frac{(2+3\gamma_0^2)B}{\eta_0^2\gamma_0^2} + \frac{\eta_0 L_0}{\gamma_0} + 1.5n^2 L_0^2$. Multiplying both sides by $\frac{\gamma_k^{r-1}}{2}$, we have that

$$\gamma_k^r \left(\mathbb{E}\left[f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right] - f^*\right) \leq \frac{\gamma_k^{r-1}}{2}\left(\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right]\right) + \gamma_k^{1+r}\theta_0. \tag{111}$$

Adding and subtracting the term $\frac{\gamma_{k-1}^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right]$, we obtain

$$\begin{aligned} &\gamma_k^r \left(\mathbb{E}\left[f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right] - f^*\right) \\ &\leq \frac{\gamma_{k-1}^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \frac{\gamma_k^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] + \left(\gamma_k^{r-1} - \gamma_{k-1}^{r-1}\right)\frac{\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right]}{2} + \gamma_k^{1+r}\theta_0 \\ &\leq \frac{\gamma_{k-1}^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \frac{\gamma_k^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] + \left(\gamma_k^{r-1} - \gamma_{k-1}^{r-1}\right) D_{\mathcal{X}} + \gamma_k^{1+r}\theta_0. \end{aligned}$$

Summing both sides from $k = 1, \ldots, K$ we obtain

$$\sum_{k=1}^{K} \gamma_k^r \left(\mathbb{E}\left[f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) \right] - f^* \right) \leq \frac{\gamma_0^{r-1}}{2} \mathbb{E}\left[\|\mathbf{x}_1 - \mathbf{x}^*\|^2 \right] + \left(\gamma_K^{r-1} - \gamma_0^{r-1} \right) D_{\mathcal{X}} + \theta_0 \sum_{k=1}^{K} \gamma_k^{1+r}.$$

Writing (111) for $k := 0$ we have

$$\gamma_0^r \left(\mathbb{E}\left[f(\mathbf{x}_0, \mathbf{y}(\mathbf{x}_0)) \right] - f^* \right) \leq \frac{\gamma_0^{r-1}}{2} \left(\mathbb{E}\left[\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \right] - \mathbb{E}\left[\|\mathbf{x}_1 - \mathbf{x}^*\|^2 \right] \right) + \gamma_0^{1+r} \theta_0.$$

Adding the preceding two relations together and using the definition of $D_{\mathcal{X}}$, we obtain

$$\sum_{k=0}^{K} \gamma_k^r \left(\mathbb{E}\left[f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) \right] - f^* \right) \leq D_{\mathcal{X}} \gamma_K^{r-1} + \theta_0 \sum_{k=0}^{K} \gamma_k^{1+r}.$$

From the definition $\bar{\mathbf{x}}_K \triangleq \sum_{k=0}^{K} \alpha_{k,K} \mathbf{x}_k$ in Lemma 28 and applying the convexity of the implicit function, we obtain

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K, \mathbf{y}(\bar{\mathbf{x}}_N)) \right] - f^* \leq \frac{D_{\mathcal{X}} \gamma_K^{r-1} + \theta_0 \sum_{k=0}^{K} \gamma_k^{1+r}}{\sum_{k=0}^{K} \gamma_k^r}.$$

Substituting $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ and invoking Lemma 25, we obtain

$$\begin{aligned}
\mathbb{E}\left[f(\bar{\mathbf{x}}_K, \mathbf{y}(\bar{\mathbf{x}}_K)) \right] - f^* &\leq \frac{D_{\mathcal{X}} \gamma_0^{r-1} (K+1)^{0.5(1-r)} + \theta_0 \gamma_0^{1+r} \frac{(K+1)^{1-0.5(1+r)}}{1-0.5(1+r)}}{\gamma_0^r \frac{(K+1)^{1-0.5r}}{2-r}} \\
&\leq (2-r) \left(\frac{D_{\mathcal{X}}}{\gamma_0} + \frac{2(D_{\mathcal{X}} + \frac{(2+3\gamma_0^2)B}{\eta_0^2 \gamma_0^2} + \frac{\eta_0 L_0}{\gamma_0} + 1.5 n^2 L_0^2) \gamma_0}{1-r} \right) \frac{1}{\sqrt{K+1}} \\
&= (2-r) \left(\frac{D_{\mathcal{X}}}{\gamma_0} + \frac{2(D_{\mathcal{X}} \gamma_0 + \frac{(2+3\gamma_0^2)B}{\eta_0^2 \gamma_0} + \eta_0 L_0 + 1.5 n^2 L_0^2 \gamma_0)}{1-r} \right) \frac{1}{\sqrt{K+1}}.
\end{aligned}$$

(c) The results in (c-1) and (c-2) follow directly from part (b) by substituting γ_0 and r. To show part (c-3), note that in Algorithm 10, we have $t_k := \lceil \tau \ln(k+1) \rceil$. From part (b), we require the following total number of iterations of the SA scheme.

$$\begin{aligned}
2 \sum_{k=0}^{K_\epsilon} t_k = 2 \sum_{k=0}^{K_\epsilon} \lceil \tau \ln(k+1) \rceil &\leq 2 (K_\epsilon + 1) + 2\tau \sum_{k=2}^{K_\epsilon + 1} \ln(k) \\
&\leq 2 (K_\epsilon + 1) + 2\tau \int_2^{K_\epsilon + 1} \ln(x) dx \leq 2 (K_\epsilon + 1) + 2\tau (K_\epsilon + 2) \ln (K_\epsilon + 2) \\
&\leq 4 \max\{\tau, 1\} (K_\epsilon + 2) \ln (K_\epsilon + 2).
\end{aligned}$$

The bound in (c-3) follows from the preceding inequality and the bound on K_ϵ in (c-1). To show (c-4), note that the total samples used in the lower-level is as follows.

$$2 \sum_{k=0}^{K_\epsilon} \sum_{t=0}^{t_k} M_t = 2 \sum_{k=0}^{K_\epsilon} \sum_{t=0}^{t_k} \lceil M_0 \rho^{-t} \rceil \leq 4 M_0 \sum_{k=0}^{K_\epsilon} \sum_{t=0}^{t_k} \rho^{-t} = \mathcal{O}\left(\sum_{k=0}^{K_\epsilon} \frac{\rho^{-t_k}}{\ln(\frac{1}{\rho})} \right) = \mathcal{O}\left(\sum_{k=0}^{K_\epsilon} \frac{\rho^{-\tau \ln(k+1)}}{\ln(\frac{1}{\rho})} \right)$$

$$\leq \mathcal{O}\left(\sum_{k=0}^{K_\epsilon} \frac{e^{(\bar{\tau}-1)\ln(k+1)}}{\ln(\frac{1}{\rho})}\right) = \mathcal{O}\left(\sum_{k=0}^{K_\epsilon} \frac{(k+1)^{\bar{\tau}-1}}{\ln(\frac{1}{\rho})}\right) = \mathcal{O}\left(\frac{K_\epsilon^{\bar{\tau}}}{\ln(\frac{1}{\rho})}\right),$$

where $\bar{\tau} \geq 1 + \tau \ln(\frac{1}{\rho})$. The bound in (c-4) follows from the preceding inequality and the bound on K_ϵ in (c-1). □

5.3.4 A zeroth-order scheme for addressing the exact regime

In this subsection, we consider the case where an exact solution of the lower-level problem is available. We develop a zeroth-order method where the gradient mapping is approximated using two evaluations of the implicit function. Similar to the inexact setting, we allow for iterative smoothing and provide the convergence analysis in addressing the original implicit problem. the outline of the scheme is provided by Algorithm 12.

Algorithm 12 Zeroth-order method for exact regimes

1: **input:** Given $\mathbf{x}_0 \in \mathcal{X}$, $\bar{\mathbf{x}}_0 := \mathbf{x}_0$, stepsize sequence $\{\gamma_k\}$, smoothing parameter sequence $\{\eta_k\}$, $r \in [0, 1)$, and $S_0 := \gamma_0^r$
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Generate a random realization $v_k \in \eta_k \mathbb{S}$
4: Evaluate $\mathbf{y}(\mathbf{x}_k)$ and $\mathbf{y}(\mathbf{x}_k + v_k)$
5: Evaluate the zeroth-order gradient approximation as follows.

$$g_{\eta_k}(\mathbf{x}_k, v_k) := \frac{n(f(\mathbf{x}_k + v_k, \mathbf{y}(\mathbf{x}_k + v_k)) - f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))v_k}{\|v_k\|\eta_k}$$

6: Update $\mathbf{x}_k$ as follows. $\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k}(\mathbf{x}_k, v_k)\right]$
7: Update the averaged iterate as follows.

$$S_{k+1} := S_k + \gamma_{k+1}^r, \qquad \bar{x}_{k+1} := \frac{S_k \bar{x}_k + \gamma_{k+1}^r \mathbf{x}_{k+1}}{S_{k+1}}$$

8: **end for**

We make use of the following lemma in the convergence analysis in this subsection (cf. [68]).

Lemma 31. Let v_k, u_k, α_k, and β_k be nonnegative random variables, and let the following relations hold almost surely:

$$\mathsf{E}\Big[v_{k+1} \mid \tilde{\mathcal{F}}_k\Big] \leq (1 + \alpha_k)v_k - u_k + \beta_k \quad \text{for all } k, \qquad \sum_{k=0}^{\infty} \alpha_k < \infty, \qquad \sum_{k=0}^{\infty} \beta_k < \infty,$$

where $\tilde{\mathcal{F}}_k$ denotes the collection $v_0, \ldots, v_k$, $u_0, \ldots, u_k$, $\alpha_0, \ldots, \alpha_k$, $\beta_0, \ldots, \beta_k$. Then, we have almost surely $\lim_{k\to\infty} v_k = v$ and $\sum_{k=0}^{\infty} u_k < \infty$, where $v \geq 0$ is some random variable.

In the following, we derive non-asymptotic convergence rate statements and also, show an almost sure convergence result for the proposed zeroth-order method in the exact regimes.

Theorem 32 (Rate statement and complexity result for Algorithm 12). Consider the problem (MPEC$^{\text{imp,exp}}$). Suppose Assumptions 13– 15 hold. Suppose $\{\bar{\mathbf{x}}_k\}$ denotes the sequence generated by Algorithm 12 in which the stepsize and smoothing sequences are defined as $\gamma_k :=$

$\frac{\gamma_0}{(k+1)^a}$ and $\eta_k := \frac{\eta_0}{(k+1)^b}$, respectively, for all $k \geq 0$ where γ_0 and η_0 are strictly positive. Then, the following statements hold.

(a) Let $a = 0.5$ and $b \in [0.5, 1)$ and $0 \leq r < 2(1-b)$. Then, for all $K \geq 2^{\frac{1}{1-r}} - 1$ we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K, \mathbf{y}(\bar{\mathbf{x}}_K))\right] - f^* \leq (2-r)\left(\frac{D_{\mathcal{X}}}{\gamma_0} + \frac{L_0^2 n^2 \gamma_0}{1-r}\right)\frac{1}{\sqrt{K+1}} + (2-r)\left(\frac{\eta_0 L_0}{1-0.5r-b}\right)\frac{1}{(K+1)^b}.$$

In particular, when $b := 1 - \delta$ and $r = 0$, where $\delta > 0$ is a small scalar, we have for all $K \geq 1$

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K, \mathbf{y}(\bar{\mathbf{x}}_K))\right] - f^* \leq 2\left(\frac{D_{\mathcal{X}}}{\gamma_0} + L_0^2 n^2 \gamma_0\right)\frac{1}{\sqrt{K+1}} + \left(\frac{2\eta_0 L_0}{\delta}\right)\frac{1}{(K+1)^{1-\delta}}.$$

(b) Let $a := 0.5$, $b = 0.5$, $r = 0$, $\gamma_0 := \frac{\sqrt{D_{\mathcal{X}}}}{nL_0}$, and $\eta_0 \leq \sqrt{D_{\mathcal{X}}}n$. Then, the iteration complexity in projection steps on $\mathcal{X}$ for achieving $\mathbb{E}\left[f(\bar{\mathbf{x}}_{K_\epsilon}, \mathbf{y}(\bar{\mathbf{x}}_{K_\epsilon}))\right] - f^* \leq \epsilon$ for some $\epsilon > 0$ is bounded as follows.

$$K_\epsilon \geq \frac{64 n^2 L_0^2 D_{\mathcal{X}}}{\epsilon^2}.$$

(c) For any $a \in (0.5, 1]$ and $b > 1 - a$, there exists $\mathbf{x}^* \in \mathcal{X}^*$ such that $\lim_{k\to\infty} \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 = 0$ almost surely.

Proof. **(a)** Let $\mathbf{x}^* \in \mathcal{X}^*$ be an arbitrary optimal solution. We can write:

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \|\Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k}(\mathbf{x}_k, v_k)\right] - \Pi_{\mathcal{X}}\left[\mathbf{x}^*\right]\|^2 \leq \|\mathbf{x}_k - \gamma_k g_{\eta_k}(\mathbf{x}_k, v_k) - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T g_{\eta_k}(\mathbf{x}_k, v_k) + \gamma_k^2 \|g_{\eta_k}(\mathbf{x}_k, v_k)\|^2.\end{aligned}$$

Let the history of the method be denoted by $\mathcal{F}_k \triangleq \{v_0, v_1, \ldots, v_{k-1}\}$ for $k \geq 1$ where $\mathcal{F}_0 \triangleq \{v_0\}$. Taking conditional expectations on the both sides and invoking Lemma 26, we obtain

$$\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathcal{F}_k\right] \leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T \nabla f_{\eta_k}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + \gamma_k^2 L_0^2 n^2.$$

Invoking the convexity of f_{η_k}, we obtain

$$\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathcal{F}_k\right] \leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k\left(f_{\eta_k}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) - f_{\eta_k}(x^*, \mathbf{y}(x^*))\right) + \gamma_k^2 L_0^2 n^2. \tag{112}$$

Taking expectations from both sides of the preceding relation and rearranging the terms, we obtain

$$2\gamma_k\left(\mathbb{E}\left[f_{\eta_k}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right] - f_{\eta_k}(\mathbf{x}^*, \mathbf{y}(\mathbf{x}^*))\right) \leq \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] + \gamma_k^2 L_0^2 n^2.$$

From the Lipschitzian property of the implicit function and Lemma 24 (v), we have that

$$f_{\eta_k}(\mathbf{x}^*, \mathbf{y}(\mathbf{x}^*)) \leq f(\mathbf{x}^*, \mathbf{y}(x^*)) + \eta_k L_0. \tag{113}$$

From the preceding two inequalities and that $f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) \leq f_{\eta_k}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))$, we obtain

$$2\gamma_k\left(\mathbb{E}\left[f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right] - f^*\right) \leq \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] + \gamma_k^2 L_0^2 n^2 + 2\gamma_k \eta_k L_0.$$

Multiplying both sides by $\frac{\gamma_k^{r-1}}{2}$, we have that

$$\gamma_k^r\left(\mathbb{E}\left[f(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right] - f^*\right) \leq \frac{\gamma_k^{r-1}}{2}\left(\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right]\right) + 0.5\gamma_k^{1+r} L_0^2 n^2 + \gamma_k^r \eta_k L_0. \tag{114}$$

Adding and subtracting the term $\frac{\gamma_{k-1}^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_k-\mathbf{x}^*\|^2\right]$, we obtain

$$\begin{aligned}&\gamma_k^r\left(\mathbb{E}\left[f(\mathbf{x}_k,\mathbf{y}(\mathbf{x}_k))\right]-f^*\right)\\&\leq\frac{\gamma_{k-1}^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_k-\mathbf{x}^*\|^2\right]-\frac{\gamma_k^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2\right]+\left(\gamma_k^{r-1}-\gamma_{k-1}^{r-1}\right)D_{\mathcal{X}}+0.5\gamma_k^{1+r}L_0^2n^2+\gamma_k^r\eta_kL_0.\end{aligned}$$

Summing both sides from $k=1,\ldots,K$ we obtain

$$\begin{aligned}\sum_{k=1}^K\gamma_k^r\left(\mathbb{E}\left[f(\mathbf{x}_k,\mathbf{y}(\mathbf{x}_k))\right]-f^*\right)&\leq\frac{\gamma_0^{r-1}}{2}\mathbb{E}\left[\|\mathbf{x}_1-\mathbf{x}^*\|^2\right]+\left(\gamma_K^{r-1}-\gamma_0^{r-1}\right)D_{\mathcal{X}}\\&\quad+0.5L_0^2n^2\sum_{k=1}^K\gamma_k^{1+r}+L_0\sum_{k=1}^K\gamma_k^r\eta_k.\end{aligned}$$

Writing (114) for $k:=0$ we have

$$\gamma_0^r\left(\mathbb{E}\left[f(\mathbf{x}_0,\mathbf{y}(\mathbf{x}_0))\right]-f^*\right)\leq\frac{\gamma_0^{r-1}}{2}\left(\mathbb{E}\left[\|\mathbf{x}_0-\mathbf{x}^*\|^2\right]-\mathbb{E}\left[\|\mathbf{x}_1-\mathbf{x}^*\|^2\right]\right)+0.5\gamma_0^{1+r}L_0^2n^2+\gamma_0^r\eta_0L_0.$$

Adding the preceding two relations together and using the definition of $D_{\mathcal{X}}$, we obtain

$$\sum_{k=0}^K\gamma_k^r\left(\mathbb{E}\left[f(\mathbf{x}_k,\mathbf{y}(\mathbf{x}_k))\right]-f^*\right)\leq D_{\mathcal{X}}\gamma_K^{r-1}+0.5L_0^2n^2\sum_{k=0}^K\gamma_k^{1+r}+L_0\sum_{k=0}^K\gamma_k^r\eta_k.$$

From the definition $\bar{\mathbf{x}}_K\triangleq\sum_{k=0}^K\alpha_{k,K}\mathbf{x}_k$ in Lemma 28 and applying the convexity of the implicit function, for all $K\geq2^{\frac{1}{1-r}}-1$ we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K,\mathbf{y}(\bar{\mathbf{x}}_N))\right]-f^*\leq\frac{D_{\mathcal{X}}\gamma_K^{r-1}+0.5L_0^2n^2\sum_{k=0}^K\gamma_k^{1+r}+L_0\sum_{k=0}^K\gamma_k^r\eta_k}{\sum_{k=0}^K\gamma_k^r}.$$

Substituting $\gamma_k:=\frac{\gamma_0}{\sqrt{k+1}}$ and $\eta_k:=\frac{\eta_0}{(k+1)^b}$, and invoking Lemma 25, we obtain

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{x}}_K,\mathbf{y}(\bar{\mathbf{x}}_K))\right]-f^*&\leq\frac{D_{\mathcal{X}}\gamma_0^{r-1}(K+1)^{0.5(1-r)}+0.5L_0^2n^2\gamma_0^{1+r}\frac{(K+1)^{1-0.5(1+r)}}{1-0.5(1+r)}+\gamma_0^r\eta_0L_0\frac{(K+1)^{1-0.5r-b}}{1-0.5r-b}}{\gamma_0^r\frac{(K+1)^{1-0.5r}}{2-r}}\\&\leq(2-r)\left(\frac{D_{\mathcal{X}}}{\gamma_0}+\frac{L_0^2n^2\gamma_0}{1-r}\right)\frac{1}{\sqrt{K+1}}+(2-r)\left(\frac{\eta_0L_0}{1-0.5r-b}\right)\frac{1}{(K+1)^b}.\end{aligned}$$

(b) Under the specified setting, from part (a) we have

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{x}}_K,\mathbf{y}(\bar{\mathbf{x}}_K))\right]-f^*&\leq2\left(\frac{D_{\mathcal{X}}}{\gamma_0}+L_0^2n^2\gamma_0\right)\frac{1}{\sqrt{K+1}}+\left(\frac{2\eta_0L_0}{0.5}\right)\frac{1}{\sqrt{K+1}}\\&=2(nL_0\sqrt{D_{\mathcal{X}}}+nL_0\sqrt{D_{\mathcal{X}}})\frac{1}{\sqrt{K+1}}+\left(4nL_0\sqrt{D_{\mathcal{X}}}\right)\frac{1}{\sqrt{K+1}}\\&=\frac{8nL_0\sqrt{D_{\mathcal{X}}}}{\sqrt{K+1}}\leq\epsilon.\end{aligned}$$

This implies the desired bound.

(c) Consider relation (112). Invoking (113), for all $k\geq0$ we have

$$\mathbb{E}[\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2\mid\mathcal{F}_k]\leq\|\mathbf{x}_k-\mathbf{x}^*\|^2-2\gamma_k(f(\mathbf{x}_k,\mathbf{y}(\mathbf{x}_k))-f^*)+2\gamma_k\eta_kL_0+\gamma_k^2L_0^2n^2.$$

Note that we have $\sum_{k=0}^{\infty}\gamma_k^2<\infty$ and $\sum_{k=0}^{\infty}\gamma_k\eta_k<\infty$ since $b>0.5$. Thus, in view of Lemma 31, we have that $\{\|\mathbf{x}_k-\mathbf{x}^*\|^2\}$ is a convergent sequence in an almost sure sense and $\sum_{k=0}^{\infty}\gamma_k(f(\mathbf{x}_k,\mathbf{y}(\mathbf{x}_k))-f^*)<\infty$ almost surely. The former statement implies that $\{\mathbf{x}_k\}$ is a bounded sequence. Further, the latter statement and $\sum_{k=0}^{\infty}\gamma_k=\infty$ imply that $\liminf_{k\to\infty}f(\mathbf{x}_k,\mathbf{y}(\mathbf{x}_k))=f^*$. Thus, from continuity of the implicit function, there is a subsequence of $\{\mathbf{x}_k\}$ for $k\in\mathcal{K}$ with its limit point denoted by $\hat{\mathbf{x}}$ such that $\hat{\mathbf{x}}\in\mathcal{X}^*$. Since $\{\|\mathbf{x}_k-\mathbf{x}^*\|^2\}$ is a convergent sequence for all $\mathbf{x}^*\in\mathcal{X}^*$, we have $\{\|\mathbf{x}_k-\hat{\mathbf{x}}\|^2\}$ is a convergent sequence. But already have that $\lim_{k\to\infty,\ k\in\mathcal{K}}\|\mathbf{x}_k-\hat{\mathbf{x}}\|^2=0$ almost surely. Hence $\lim_{k\to\infty}\|\mathbf{x}_k-\hat{\mathbf{x}}\|^2=0$ almost surely where $\hat{\mathbf{x}}\in\mathcal{X}^*$. Next, we show that $\lim_{k\to\infty}\|\bar{\mathbf{x}}_k-\hat{\mathbf{x}}\|^2=0$. In view of Lemmas 28 and 29, it suffices to have $\sum_{k=0}^{\infty}\gamma_k^r=\infty$ or equivalently, we must have $ar\leq 1$. This is already satisfied as a consequence of $a\in(0.5,1]$ and $r\in[0,1)$. □

5.3.5 Accelerated schemes

In this subsection, we consider an accelerated scheme for resolving the problem (SMPEC$^{\text{as}}$), whose implicit form is defined as

$$\min_{\mathbf{x}\in\mathcal{X}}\ f(\mathbf{x})\ \triangleq\mathbb{E}[\tilde{f}(\mathbf{x},\mathbf{y}(\mathbf{x},\omega))] \tag{115}$$

where $\mathbf{y}(\mathbf{x},\omega)$ is the unique solution of an ω-specific strongly monotone variational inequality problem parametrized by $\mathbf{x}$. Note that to ease the exposition, we consider the slightly simplified version of (SMPEC$^{\text{as}}$) in which the uncertainty only arises through the lower-level decision $\mathbf{y}(\mathbf{x},\omega)$. The deterministic counterpart of this problem is the standard MPEC in which the lower-level problem is a parametrized strongly monotone variational inequality problem. While the previous subsection has considered a standard gradient-based framework, we consider an accelerated counterpart motivated by Nesterov's celebrated accelerated gradient method [36] that produces a non-asymptotic rate of $\mathcal{O}(1/k^2)$ in terms of suboptimality for smooth convex optimization problems. In [192], Nesterov and Spokoiny develop an accelerated zeroth-order scheme for the unconstrained minimization of a smooth function. Instead, we present an accelerated gradient-free scheme for a nonsmooth function by leveraging the smoothing architecture. Notably, this scheme can contend with MPECs with convex implicit functions. In this subsection, we assume that $\mathbf{y}(\mathbf{x},\omega)$ can be generated by invoking a suitable variational inequality problem solver.

We provide convergence theory for `acc-ZSOL` by appealing to related work on smoothed accelerated schemes for nonsmooth stochastic convex optimization [26]. There are two key differences between the framework presented here and that of our prior work.

(a) *Smoothing.* In [26], we employ a deterministic smoothing technique [59] while in this paper, we consider a locally randomized smoothing technique in a zeroth-order regime. Notably, the latter leads to similar (but not identical) smoothness properties with related relationships (but not identical) between the smoothed function and its original counterpart.

(b) *Zeroth-order gradient approximation.* In [26], a sampled gradient of the smoothed function is available. However, faced by the need to resolve hierarchical problems, we do not have such access in this paper. Instead, we utilize an increasingly accurate zeroth-order approximation of the gradient by raising the sample-size N_k in constructing this approximation.

If $g_{\eta_k}(\mathbf{x}_k)$ is defined as

$$g_{\eta_k}(\mathbf{x}_k)=\mathbb{E}_{v\in\eta\mathbb{S}}\left[\mathbb{E}\left[\tilde{f}(\mathbf{x}+v,\mathbf{y}(\mathbf{x}+v,\omega))\mid v\right]\right], \tag{118}$$

we may define $\bar{w}_{k,N_k}$ as follows.

$$\bar{w}_{k,N_k}=g_{\eta_k,N_k}(\mathbf{x}_k)-g_{\eta_k}(\mathbf{x}_k). \tag{119}$$

Algorithm 13 `acc-ZSOL`: Variance-reduced accelerated zeroth-order method

1: **input:** Given $\mathbf{x}_0 \in \mathcal{X}$, $\bar{\mathbf{x}}_0 := \mathbf{x}_0$, $\lambda_0 = 1$, stepsize sequence $\{\gamma_k\}$, smoothing parameter sequence $\{\eta_k\}$, sample-size $\{N_k\}$

2: **for** $k = 0, 1, \ldots, K-1$ **do**

3: Generate a random realization $v_{j,k} \in \eta_k \mathbb{S}$ and $\mathbf{y}(\mathbf{x}_k + v_{j,k}, \omega_{j,k})$ for $j = 1, \cdots, N_k$

4: Evaluate the variance-reduced zeroth-order gradient approximation as follows.

$$g_{\eta_k, N_k}(\mathbf{x}_k) := \sum_{j=1}^{N_k} \frac{n\left(\tilde{f}(\mathbf{x}_k + v_{j,k}, \mathbf{y}(\mathbf{x}_k + v_{j,k}, \omega_{j,k})) - \tilde{f}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k, \omega_{j,k}))\right) v_{j,k}}{\|v_{j,k}\| N_k \eta_k}. \tag{116}$$

5: Update $\mathbf{x}_k$ as follows.

$$\begin{aligned} \mathbf{z}_{k+1} &:= \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k, N_k}(\mathbf{x}_k, v_k)\right] \\ \lambda_{k+1} &:= \tfrac{1+\sqrt{1+4\lambda_k^2}}{2} \\ \mathbf{x}_{k+1} &= \mathbf{z}_{k+1} + \tfrac{(\lambda_k - 1)}{\lambda_{k+1}}\left(\mathbf{z}_{k+1} - \mathbf{z}_k\right). \end{aligned} \tag{117}$$

6: **end for**

The following claims can be made about $\bar{w}_{k,N_k}$ obtained by generating N_K independent realizations given by $\{v_{j,k}\}_{j=1}^{N_k}$ and $\{\mathbf{y}(\mathbf{x}^k, \omega_{j,k})\}_{j=1}^{N_k}$.

Assumption 17. Let the following hold and for all $k \geq 0$ and $1 \leq \ell \leq N_k$.
(a) The random realizations $\{\omega_{k,\ell}\}$ are independent and identically distributed.
(b) $\mathbb{E}_{\omega,v}[\bar{w}_{k,N_k} \mid \mathbf{x}_k] = 0$, almost surely.
(c) $\mathbb{E}_{\omega,v}[\|\bar{w}_{k,N_k}\|^2 \mid \mathbf{x}_k] \leq \frac{\nu^2}{N_k}$ for some $\nu > 0$.

Lemma 33. [26, Lemma 4] Consider the problem (5.3.5). Suppose Assumptions 13– 15, 17 hold. Suppose $\{\mathbf{x}_k, \mathbf{z}_k\}$ denote the sequence generated by Algorithm 13 in which the stepsize and smoothing sequences are defined as $\eta_k = \frac{1}{k}$ and $\gamma_k = \frac{1}{2k}$, and $N_k = \lfloor k^a \rfloor$ for $k \geq 1$. Suppose $\mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^*\|] \leq C$. Then the following holds for $a = 1 + \delta$.

$$\mathbb{E}\left[f_{\eta_K}(\mathbf{z}_K) - f_{\eta_K}(\mathbf{x}^*)\right] \leq \frac{2}{\gamma_{K-1}(K-1)^2} \sum_{k=1}^{K-1} \frac{\gamma_k^2 k^2 \nu^2}{N_{k-1}} + \frac{2C^2}{\gamma_{K-1}(K-1)^2}. \tag{120}$$

We may now provide the main rate statement for the smoothed accelerated scheme by adapting [26, Thm. 5].

Proposition 10 (Rate statement for Algorithm 13). Consider the problem (5.3.5). Suppose Assumptions 13– 15, 17 hold. Suppose $\{\mathbf{x}_k, \mathbf{z}_k\}$ denote the sequence generated by Algorithm 13 in which the stepsize and smoothing sequences are defined as $\eta_k = \frac{1}{k}$ and $\gamma_k = \frac{1}{2k}$, and $N_k = \lfloor k^a \rfloor$ for $k \geq 1$. Suppose $\mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^*\|] \leq C$. Then the following hold for $a = 1 + \delta$.
(a) For $\mathbb{E}[f(\mathbf{z}_k) - f(\mathbf{x}^*)] \leq \mathcal{O}\left(\frac{1}{k}\right)$. Then the iteration complexity in zeroth-order gradient steps is $\mathcal{O}(1/\epsilon)$.
(b) Suppose K^ϵ is such that $\mathbb{E}[f(\mathbf{z}_k) - f(\mathbf{x}^*)] \leq \epsilon$. Then $\sum_{k=1}^{K_\epsilon} N_k \leq \mathcal{O}(1/\epsilon^{2+\delta})$ implying that the iteration complexity in terms of lower-level calls to the VI solver is $\mathcal{O}(1/\epsilon^{2+\delta})$.

Proof. **(a)** From Lemma 33, we have that

$$\mathbb{E}\left[f_{\eta_K}(\mathbf{z}_K) - f_{\eta_K}(\mathbf{x}^*)\right] \leq \frac{2}{\gamma_{K-1}(K-1)^2}\sum_{k=1}^{K-1}\frac{\gamma_k^2 k^2\nu^2}{N_{k-1}} + \frac{2C^2}{\gamma_{k-1}(K-1)^2}. \tag{121}$$

From Lemma 24 (v), we have that $f(\mathbf{x}) \leq f_\eta(\mathbf{x}) \leq f(\mathbf{x}) + \eta L_0$. Consequently, we have that

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{z}_K) - f(\mathbf{x}^*)\right] &\leq \mathbb{E}\left[f_{\eta_K}(\mathbf{z}_K) - f_{\eta_K}(\mathbf{x}^*)\right] + \eta_K L_0 \\ &\leq \frac{2}{\gamma_{K-1}(K-1)^2}\sum_{k=1}^{K-1}\frac{\gamma_k^2 k^2\nu^2}{N_{k-1}} + \frac{2C^2}{\gamma_{K-1}(K-1)^2} + \eta_K L_0 \leq \mathcal{O}(\tfrac{1}{K}),\end{aligned}$$

by invoking $\gamma_k = 1/(2k)$, $\eta_k = 1/k$, and $N_k = \lfloor k^a \rfloor$ where $a = 1 + \delta$.
(b) Omitted.

□

Remark 6. *Several points deserve emphasis. (i) The proposed scheme employs diminishing smoothing sequences rather than fixed, leading to asymptotic convergence guarantees, a key distinction from the scheme proposed in [192]. (ii) By adapting the framework employed for the inexact oracles, one may consider similar extensions to the accelerated framework. However, this would lead to bias in the gradient approximation and one would expect this to adversely affect the rate. This remains a goal of future study.*

5.4 Nonconvex settings

In this section, in addressing (MPEC$^{\text{imp,exp}}$) in the nonconvex case, we consider a smoothed implicit problem given by the following.

$$\begin{aligned}\min \quad & f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x})) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X},\end{aligned} \tag{122}$$

where f_η is defined by (G-Smooth) for a given $\eta > 0$. The outline of the proposed zeroth-order scheme is given by Algorithms 14–15. We make the following assumptions.

Assumption 18. Given a mini-batch sequence $\{N_k\}$ and a smoothing parameter $\eta > 0$, let $v_{j,k} \in \mathbb{R}^n$, for $j = 1, \ldots, N_k$ and $k \geq 0$ be generated randomly and independently, from $\eta\mathbb{S}$ for all $k \geq 0$.

Assumption 19. Let the following hold and for all $k \geq 0$ and $t \geq 0$.
(a) The random realizations ω_t for $t \geq 0$ are independent and identically distributed.
(b) $\mathbb{E}[G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t}) \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] = F(\hat{\mathbf{x}}_k, \mathbf{y}_t)$.
(c) $\mathbb{E}[\|G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t}) - F(\mathbf{x}_k, \mathbf{y}_t)\|^2 \mid \mathbf{x}_k, \mathbf{y}_t] \leq \nu_G^2$ for some $\nu_G > 0$.

We utilize the following definition and lemma in the analysis in this subsection.

Definition 6 (The residual mappings)**.** *Suppose Assumption 13 holds. Given a scalar $\beta > 0$ and a smoothing parameter $\eta > 0$, for any $\mathbf{x} \in \mathbb{R}^n$, let the residual mappings $G_{\eta,\beta}(\mathbf{x})$ and $\tilde{G}_{\eta,\beta}(\mathbf{x})$ be defined as follows.*

$$G_{\eta,\beta}(\mathbf{x}) \triangleq \beta\Big(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}\nabla_x f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x}))\right]\Big), \tag{123}$$

$$\tilde{G}_{\eta,\beta}(\mathbf{x}) \triangleq \beta\left(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x})) + \tilde{e})\right]\right), \tag{124}$$

where $\tilde{e} \in \mathbb{R}^n$ is an arbitrary given vector.

Algorithm 14 VR-ZSOL: Variance reduced inexact zeroth-order method for the nonconvex case

input: Given $\mathbf{x}_0 \in \mathcal{X}$, $\bar{\mathbf{x}}_0 := \mathbf{x}_0$, stepsize $\gamma > 0$, smoothing parameter $\eta > 0$, mini-batch sequence $\{N_k\}$ such that $N_k := k+1$, an integer K, a scalar $\lambda \in (0,1)$, and an integer R randomly selected from $\{\lceil \lambda K \rceil, \ldots, K\}$ using a uniform distribution
for $k = 0, 1, \ldots, K-1$ **do**
  Call Algorithm 11 to obtain $\mathbf{y}_{\epsilon_k}(\mathbf{x}_k) := \mathbf{y}_{t_k}$
  for $j = 1, \ldots, N_k$ **do**
    Generate a random batch $v_{j,k} \in \eta\mathbb{S}$ for $j = 1, \ldots, N_k$
    Call Algorithm 11 to obtain $\mathbf{y}_{\epsilon_k}(\mathbf{x}_k + v_{j,k}) := \mathbf{y}_{t_k}$
    Compute a sample zeroth-order gradient approximation as follows.

$$g_{\eta,\epsilon_k}(\mathbf{x}_k, v_{j,k}) := \frac{n\left(f(\mathbf{x}_k + v_{j,k}, \mathbf{y}_{\epsilon_k}(\mathbf{x}_k + v_{j,k})) - f(\mathbf{x}_k, \mathbf{y}_{\epsilon_k}(\mathbf{x}_k))\right) v_{j,k}}{\|v_{j,k}\|\eta}$$

  end for
  Evaluate the mini-batch inexact zeroth-order gradient. $g_{\eta,N_k,\epsilon_k}(\mathbf{x}_k) = \frac{\sum_{j=1}^{N_k} g_{\eta,\epsilon_k}(\mathbf{x}_k, v_{j,k})}{N_k}$
  Update $\mathbf{x}_k$ as follows. $\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma g_{\eta,N_k,\epsilon_k}(\mathbf{x}_k)\right]$
end for
Return $\mathbf{x}_R$

Algorithm 15 SA method for stochastic VI in the lower-level

input: An arbitrary $\mathbf{y}_0 \in \mathcal{Y}$, vector $\hat{\mathbf{x}}_k$, and initial stepsize $\alpha_0 > \frac{1}{2\mu_F}$
Set $t_k := k+1$
for $t = 0, 1, \ldots, t_k - 1$ **do**
  Generate a random realization of the stochastic mapping $G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t)$
  Update $\mathbf{y}_t$ as follows. $\mathbf{y}_{t+1} := \Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \alpha_t G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t)\right]$
  Update the stepsize using $\alpha_{t+1} := \frac{\alpha}{t+\Gamma}$
end for
Return $\mathbf{y}_{t_k}$

It may be observed that $G_{\eta,\beta}$ is a residual for stationarity for the minimization of smooth nonconvex objectives over convex sets (cf. [59]). In fact, the first part of (125) is a consequence of the well known result relating the residual function $G_{\eta,\beta}(\mathbf{x})$ to the standard stationarity condition (cf. [200, Thm. 9.10]) while the second implication in (125) is Prop. 9.

Lemma 34. Consider the problem (122). Then the following holds for any $\eta, \beta > 0$.

$$[G_{\eta,\beta}(\mathbf{x}) = 0] \iff [0 \in \nabla_{\mathbf{x}} f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x})) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})] \implies [0 \in \partial_{2\eta} f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x})) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})]. \tag{125}$$

Consequently, a zero of the residual of the η-smoothed problem satisfies an η-approximate stationarity property for the original problem. The residual $\tilde{G}_{\eta,\beta}$ represents the counterpart of $G_{\eta,\beta}$ when employing an error-afflicted estimate of the gradient. We now derive a bound on $G_{\eta,\beta}$ in terms of $\tilde{G}_{\eta,\beta}$ and $\tilde{e}$, the error in the gradient.

Lemma 35. Let Assumption 13 hold. Then the following holds for any $\beta > 0$, $\eta > 0$, and $\mathbf{x} \in \mathbb{R}^n$.

$$\|G_{\eta,\beta}(\mathbf{x})\|^2 \leq 2\|\tilde{G}_{\eta,\beta}(\mathbf{x})\|^2 + 2\|\tilde{e}\|^2.$$

Proof. From Definition 6, we may bound $G_{\eta,\beta}(\mathbf{x})$ as follows.

$$\begin{aligned}
\|G_{\eta,\beta}(\mathbf{x})\|^2 &= \left\|\beta\left(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}\nabla_x f_\eta(\mathbf{x},\mathbf{y}(\mathbf{x}))\right]\right)\right\|^2 \\
&= \left\|\beta\left(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta(\mathbf{x},\mathbf{y}(\mathbf{x})) + \tilde{e})\right]\right)\right. \\
&\quad \left. + \beta\Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta(\mathbf{x},\mathbf{y}(\mathbf{x})) + \tilde{e})\right] - \beta\Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}\nabla_x f_\eta(\mathbf{x},\mathbf{y}(\mathbf{x}))\right]\right\|^2 \\
&\leq 2\left\|\beta\left(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta(\mathbf{x},\mathbf{y}(\mathbf{x})) + \tilde{e})\right]\right)\right\|^2 \\
&\quad + 2\left\|\beta\Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta(\mathbf{x},\mathbf{y}(\mathbf{x})) + \tilde{e})\right] - \beta\Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}\nabla_x f_\eta(\mathbf{x},\mathbf{y}(\mathbf{x}))\right]\right\|^2 \\
&\leq 2\|\tilde{G}_{\eta,\beta}(\mathbf{x})\|^2 + 2\|\tilde{e}\|^2,
\end{aligned}$$

where the last inequality is a consequence of the non-expansivity of the Euclidean projector. □

The proposed scheme can be compactly represented as follows.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma\left(\nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + e_k\right)\right], \tag{126}$$

where we define the stochastic errors $e_k \triangleq g_{\eta,N_k,\epsilon_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))$ for all $k \geq 0$. We make use of the following result in the convergence analysis.

Lemma 36. Let Assumption 13. Suppose $\mathbf{x}_k$ is generated by Algorithm 14 in which $\gamma \in (0, \frac{\eta}{nL_0})$ for a given $\eta > 0$. Then, we have for any k,

$$f_\eta(\mathbf{x}_{k+1}, \mathbf{y}(\mathbf{x}_{k+1})) \leq f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + \left(-1 + \tfrac{nL_0\gamma}{\eta}\right)\tfrac{\gamma}{4}\|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma\|e_k\|^2.$$

Proof. Note that by Lemma 24 (iv), $\nabla f_\eta(\bullet, \mathbf{y}(\bullet))$ is Lipschitz with parameter $L \triangleq \frac{nL_0}{\eta}$. By the descent lemma, we have that

$$\begin{aligned}
f_\eta(\mathbf{x}_{k+1}, \mathbf{y}(\mathbf{x}_{k+1})) &\leq f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + \nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))^T(\mathbf{x}_{k+1} - \mathbf{x}_k) + \tfrac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + \left(\nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + e_k\right)^T(\mathbf{x}_{k+1} - \mathbf{x}_k) \\
&\quad - e_k^T(\mathbf{x}_{k+1} - \mathbf{x}_k) + \tfrac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.
\end{aligned}$$

From the properties of the Euclidean projection, we have that

$$\begin{aligned}
&(\mathbf{x}_k - \gamma(\nabla_x f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + e_k)) - \mathbf{x}_{k+1})^T(\mathbf{x}_k - \mathbf{x}_{k+1}) \leq 0 \\
\implies &(\nabla_x f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + e_k))^T(\mathbf{x}_{k+1} - \mathbf{x}_k) \leq -\tfrac{1}{\gamma}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.
\end{aligned}$$

In addition, for any $u, v \in \mathbb{R}^n$ we can write $u^T v \leq \frac{1}{2}\left(\gamma\|u\|^2 + \frac{\|v\|^2}{\gamma}\right)$. Thus, we have that

$$-e_k^T(\mathbf{x}_{k+1} - \mathbf{x}_k) \leq \tfrac{\gamma}{2}\|e_k\|^2 + \tfrac{1}{2\gamma}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.$$

Consequently, from the preceding three inequalities we have that

$$\begin{aligned}
f_\eta(\mathbf{x}_{k+1}, \mathbf{y}(\mathbf{x}_{k+1})) &\leq f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) - \tfrac{1}{\gamma}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 + \tfrac{1}{2\gamma}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \tfrac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + \left(-\tfrac{1}{2\gamma} + \tfrac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \tfrac{\gamma}{2}\|e_k\|^2.
\end{aligned}$$

From $\gamma < \frac{1}{L}$, we have

$$\begin{aligned} f_\eta(\mathbf{x}_{k+1}, \mathbf{y}(\mathbf{x}_{k+1})) &\leq f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + \left(-\tfrac{1}{2\gamma} + \tfrac{L}{2}\right) \|\mathbf{x}_{k+1} - \mathbf{x}\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 \\ &= f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + \left(-\tfrac{1}{2\gamma} + \tfrac{L}{2}\right) \gamma^2 \|\tilde{G}_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 \\ &= f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + (-1 + L\gamma)\, \tfrac{\gamma}{2} \|\tilde{G}_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 \\ &\overset{\text{Lemma 35}}{\leq} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + (-1 + L\gamma)\, \tfrac{\gamma}{4} \|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + (1 - L\gamma)\, \tfrac{\gamma}{2}\|e_k\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 \\ &= f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) + (-1 + L\gamma)\, \tfrac{\gamma}{4} \|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + \left(1 - \tfrac{L\gamma}{2}\right)\gamma \|e_k\|^2. \end{aligned}$$

Substituting $L := \frac{nL_0}{\eta}$ we obtain the desired inequality. □

We make use of the following result in the convergence analysis.

Lemma 37. Let $\{e_k\}$ be a non-negative sequence such that for an arbitrary non-negative sequence $\{\gamma_k\}$, the following relation is satisfied.

$$e_{k+1} \leq (1 - \alpha\gamma_k)e_k + \beta\gamma_k^2, \quad \text{for all } k \geq 0. \tag{127}$$

where α and β are positive scalars. Suppose $\gamma_k = \frac{\gamma}{k+\Gamma}$ for any $k \geq 0$, where $\gamma > \frac{1}{\alpha}$ and $\Gamma > 0$. Then, we have

$$e_k \leq \frac{\max\left\{\frac{\beta\gamma^2}{\alpha\gamma - 1}, \Gamma e_0\right\}}{k+\Gamma}, \qquad \text{for all } k \geq 0. \tag{128}$$

Next, we present the rate and complexity result for the proposed inexact method for addressing the nonconvex case.

Theorem 38 (Rate statements and complexity results for Algorithms 14–15). Consider Algorithms 14–15 for solving (MPEC$^{\text{imp,exp}}$) and suppose Assumptions 13, 18, and 19 hold.
(a) Given $\hat{\mathbf{x}}_k \in \mathcal{X}$, let $\mathbf{y}(\hat{\mathbf{x}}_k)$ denote the unique solution of VI$(\mathcal{Y}, F(\bullet, \hat{\mathbf{x}}_k))$. Let $\mathbf{y}_{t_k}$ be generated by Algorithm 15. Let us define $C_F \triangleq \max_{\mathbf{x}\in X,\ \mathbf{y}\in\mathcal{Y}} \|F(\mathbf{x}, \mathbf{y})\|$. Then for all $t_k \geq 0$, we have

$$\mathbb{E}[\|\mathbf{y}_{t_k} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \leq \epsilon_k \triangleq \frac{\max\left\{\frac{(C_F^2 + \nu_G^2)\alpha^2}{2\alpha\mu_F - 1}, \Gamma \sup_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{y} - \mathbf{y}_0\|^2\right\}}{t_k + \Gamma}.$$

(b) The following holds for any $\gamma < \frac{\eta}{nL_0}$, $\ell \triangleq \lceil \lambda K \rceil$, and all $K > \frac{2}{1-\lambda}$.

$$\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{n^2\gamma(1 - 2\ln(\lambda))\left(1 - \frac{nL_0\gamma}{2\eta}\right)\left(\frac{4\tilde{L}_0^2}{\eta^2} + L_0^2\right) + \mathbb{E}\left[f(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell))\right] - f^* + 2L_0\eta}{\left(1 - \frac{nL_0\gamma}{\eta}\right)\frac{\gamma}{4}(1 - \lambda)K}.$$

(c) Suppose $\gamma = \frac{\eta}{2nL_0}$ and $\eta = \frac{1}{L_0}$. Let $\epsilon > 0$ be an arbitrary scalar and K_ϵ be such that $\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \epsilon$. Then,
(c-1) The total number of upper-level projection steps on $\mathcal{X}$ is $K_\epsilon = \mathcal{O}\left(n^2 L_0^2 \epsilon^{-1}\right)$.
(c-2) The total sample complexity of upper-level is $\mathcal{O}\left(n^4 L_0^4 \epsilon^{-2}\right)$.
(c-3) The total number of lower-level projection steps on $\mathcal{Y}$ is $\mathcal{O}\left(n^6 L_0^6 \epsilon^{-3}\right)$.
(c-4) The total sample complexity of lower-level is $\mathcal{O}\left(n^6 L_0^6 \epsilon^{-3}\right)$.

Proof. **(a)** Let us define the errors $\Delta_t \triangleq G(\hat{\mathbf{x}}_k, y_t, \omega_t) - F(\hat{\mathbf{x}}_k, y_t)$ for $t \geq 0$. Also, let the history of Algorithm 11 be denoted by $\mathcal{F}_t \triangleq \{\omega_0, \ldots, \omega_{t-1}\}$ for $t \geq 1$, and $\mathcal{F}_0 \triangleq \{\omega_0\}$. We have

$$\begin{aligned}
&\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{x}_k)\|^2 = \|\Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \alpha_t G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t)\right] - \Pi_{\mathcal{Y}}\left[\mathbf{y}(\hat{x}_k)\right]\|^2 \leq \|\mathbf{y}_t - \alpha_t G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t) - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \\
&= \|\mathbf{y}_t - \alpha_t F(\hat{\mathbf{x}}_k, \mathbf{y}_t) - \alpha_t \Delta_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \\
&= \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \alpha_t^2 \|F(\hat{\mathbf{x}}_k, \mathbf{y}_t)\|^2 + \alpha_t^2 \|\Delta_t\|^2 - 2\alpha_t (\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T F(\hat{\mathbf{x}}_k, \mathbf{y}_t) \\
&- 2\alpha_t (\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k) - \alpha_t F(\hat{\mathbf{x}}_k, \mathbf{y}_t))^T \Delta_t.
\end{aligned}$$

Taking conditional expectations from the preceding relation and invoking Assumption 19, we obtain

$$\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \mid \mathcal{F}_t] \leq \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \alpha_t^2 (C_F^2 + \nu_G^2) - 2\alpha_t (\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T F(\hat{\mathbf{x}}_k, y_t).$$

From strong monotonicity of mapping $F(\hat{\mathbf{x}}_k, \bullet)$ uniformly in $\hat{\mathbf{x}}_k$ and the definition of $\mathbf{y}(\hat{x}_k)$, we have

$$(\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T F(\hat{\mathbf{x}}_k, y_t) \geq (\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T F(\mathbf{y}(\hat{x}_k), \hat{\mathbf{x}}_k) + \mu_F \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \geq \mu_F \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2.$$

From the preceding relations, we obtain

$$\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \mid \mathcal{F}_t] \leq (1 - 2\mu_F \alpha_t) \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \alpha_t^2 (C_F^2 + \nu_G^2).$$

Taking expectations from both sides, we have

$$\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \leq (1 - 2\mu_F \alpha_t) \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] + \alpha_t^2 (C_F^2 + \nu_G^2).$$

Noting that in Algorithm 15 we have $\alpha_0 > \frac{1}{2\mu_F}$, using Lemma 37, we obtain that

$$\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \leq \frac{\max\left\{\frac{(C_F^2 + \nu_G^2)\alpha^2}{2\alpha\mu_F - 1}, \Gamma \sup_{\mathbf{y} \in \mathcal{Y}} \|\mathbf{y} - \mathbf{y}_0\|^2\right\}}{t + \Gamma}, \qquad \text{for all } t \geq 0.$$

(b) Let the history of Algorithm 14 be denoted by $\mathcal{F}_k \triangleq \cup_{i=0}^{k-1} \cup_{j=1}^{N_i} \{v_{j,i}\}$ for $k \geq 1$, and $\mathcal{F}_0 \triangleq \cup_{j=1}^{N_0} \{v_{j,0}\}$. We can write

$$\begin{aligned}
&\mathbb{E}\left[\|e_k\|^2 \mid \mathcal{F}_k\right] = \mathbb{E}\left[\|g_{\eta, N_k, \epsilon_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\|^2 \mid \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\|\frac{\sum_{j=1}^{N_k} g_{\eta, \epsilon_k}(\mathbf{x}_k, v_{j,k})}{N_k} - \nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right\|^2 \mid \mathcal{F}_k\right] \\
&\leq 2\mathbb{E}\left[\left\|\frac{\sum_{j=1}^{N_k} g_{\eta, \epsilon_k}(\mathbf{x}_k, v_{j,k})}{N_k} - \frac{\sum_{j=1}^{N_k} g_\eta(\mathbf{x}_k, v_{j,k})}{N_k}\right\|^2 \mid \mathcal{F}_k\right] + 2\mathbb{E}\left[\left\|\frac{\sum_{j=1}^{N_k} g_\eta(\mathbf{x}_k, v_{j,k})}{N_k} - \nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right\|^2 \mid \mathcal{F}_k\right] \\
&\leq \frac{2\sum_{j=1}^{N_k} \mathbb{E}\left[\left\|g_{\eta, \epsilon_k}(\mathbf{x}_k, v_{j,k}) - g_\eta(\mathbf{x}_k, v_{j,k})\right\|^2 | \mathcal{F}_k\right]}{N_k} + \frac{2\sum_{j=1}^{N_k} \mathbb{E}\left[\left\|g_{\eta,}(\mathbf{x}_k, v_{j,k}) - \nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\right\|^2 | \mathcal{F}_k\right]}{N_k^2} \\
&\leq \frac{8\tilde{L}_0^2 n^2 \epsilon_k}{\eta^2} + \frac{2\sum_{j=1}^{N_k} \left(\mathbb{E}\left[\left\|g_{\eta,}(\mathbf{x}_k, v_{j,k})\right\|^2 | \mathcal{F}_k\right] - \|\nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))\|^2\right)}{N_k^2} \\
&\leq \frac{8\tilde{L}_0^2 n^2 \epsilon_k}{\eta^2} + \frac{2n^2 L_0^2}{N_k},
\end{aligned}$$

where in the second inequality, the first term is implied by the relation $\|\sum_{i=1}^m u_i\|^2 \leq m \sum_{i=1}^m \|u_i\|^2$ for any $u_i \in \mathbb{R}^n$ for all $i = 1, \ldots, m$. The second term in the second inequality is implied by noting

that from Lemma 26, $g_\eta(\mathbf{x}_k, v)$ is an unbiased estimator of $\nabla_{\mathbf{x}} f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k))$. From Lemma 35 we have

$$\left(1 - \tfrac{nL_0\gamma}{\eta}\right) \tfrac{\gamma}{4} \|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 \leq f_\eta(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k)) - f_\eta(\mathbf{x}_{k+1}, \mathbf{y}(\mathbf{x}_{k+1})) + \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \|e_k\|^2.$$

Let $f_\eta^* \triangleq \inf_{x\in\mathcal{X}} f_\eta(\mathbf{x}, \mathbf{y}(\mathbf{x}))$. Summing the preceding relation from $k = \ell, \ldots, K-1$ where $\ell \triangleq \lceil \lambda K \rceil$, we have that

$$\left(1 - \tfrac{nL_0\gamma}{\eta}\right) \tfrac{\gamma}{4} \sum_{k=\ell}^{K-1} \|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 \leq f_\eta(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell)) - f_\eta(\mathbf{x}_K, \mathbf{y}(\mathbf{x}_K)) + \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \sum_{k=\ell}^{K-1} \|e_k\|^2.$$

Taking expectations from the both sides, it follows that

$$\begin{aligned}
&\left(1 - \tfrac{nL_0\gamma}{\eta}\right) \tfrac{\gamma}{4}(K-\ell)\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \sum_{k=\ell}^{K-1} \mathbb{E}\left[\|e_k\|^2\right] + \mathbb{E}\left[f_\eta(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell))\right] - f_\eta^* \\
&\leq \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \sum_{k=\ell}^{K-1} \mathbb{E}\left[\|e_k\|^2\right] + \mathbb{E}\left[f(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell)) + f_\eta(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell)) - f(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell))\right] - f_\eta^* + f^* - f^* \\
&\leq \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \sum_{k=\ell}^{K-1} \mathbb{E}\left[\|e_k\|^2\right] + \mathbb{E}\left[f(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell))\right] - f^* + \mathbb{E}\left[|f_\eta(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell)) - f(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell))|\right] + \left|f^* - f_\eta^*\right| \\
&\leq \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \sum_{k=\ell}^{K-1} \left(\tfrac{8\tilde{L}_0^2 n^2 \epsilon_k}{\eta^2} + \tfrac{2n^2 L_0^2}{N_k}\right) + \mathbb{E}\left[f(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell))\right] - f^* + 2L_0\eta,
\end{aligned}$$

where the preceding relation is implied by invoking the bound on $\mathbb{E}\left[\|e_k\|^2\right]$ and Lemma 24 (iii). Note that from part (a), we have $\epsilon_k = \frac{2(C_F^2+\nu_G^2)}{\mu_F^2 t_k}$ where $t_k := k+1$. Also, $N_k := k+1$. Note that $K > \frac{2}{1-\lambda}$ implies $\ell \leq K-1$. From Lemma 25, using $\ell \geq 1$ we have $\sum_{k=\ell}^{K-1} \frac{1}{k+1} \leq \frac{1}{\ell+1} + \ln\left(\frac{K}{\ell+1}\right) \leq 0.5 + \ln\left(\frac{N}{\lambda N+1}\right) \leq 0.5 - \ln(\lambda)$. Also, $K - \ell \geq K - \lambda K = (1-\lambda)K$. Thus, we obtain

$$\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{\left(1 - \tfrac{nL_0\gamma}{2\eta}\right)2n^2\gamma\left(\tfrac{4\tilde{L}_0^2}{\eta^2} + L_0^2\right)(0.5 - \ln(\lambda)) + \mathbb{E}\left[f(\mathbf{x}_\ell, \mathbf{y}(\mathbf{x}_\ell))\right] - f^* + 2L_0\eta}{\left(1 - \tfrac{nL_0\gamma}{\eta}\right)\tfrac{\gamma}{4}(1-\lambda)K}.$$

(c) To show (c-1), using the relation in part (b) and substituting $\gamma = \frac{\eta}{2nL_0}$ we obtain

$$\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{6n^2(1 - 2\ln(\lambda))\left(\tfrac{4\tilde{L}_0^2}{\eta^2} + L_0^2\right) + \tfrac{16nL_0}{\eta}(\sup_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}(\mathbf{x})) - f^*) + 32nL_0^2}{(1-\lambda)K}.$$

Further, from $\eta = \frac{1}{L_0}$ we obtain

$$\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{6n^2L_0^2(1 - 2\ln(\lambda))\left(4\tilde{L}_0^2 + 1\right) + 16nL_0^2(\sup_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}(\mathbf{x})) - f^*) + 32nL_0^2}{(1-\lambda)K}.$$

This implies that $\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{\mathcal{O}\left(n^2L_0^2\right)}{K}$ and thus, we obtain $K_\epsilon = \mathcal{O}\left(n^2L_0^2\epsilon^{-1}\right)$. Next, we show (c-2). The total sample complexity of upper-level is as follows.

$$\sum_{k=0}^{K_\epsilon} N_k = \sum_{k=0}^{K_\epsilon}(k+1) = \mathcal{O}(K_\epsilon^2) = \mathcal{O}\left(n^4L_0^4\epsilon^{-2}\right).$$

To show (c-3), note that the total number of lower-level projection steps is given by

$$\sum_{k=0}^{K_\epsilon}(1+N_k)t_k = \sum_{k=0}^{K_\epsilon}(k+1)(k+2) = \mathcal{O}(K_\epsilon^3) = \mathcal{O}\left(n^6 L_0^6 \epsilon^{-3}\right).$$

Noting that at each iteration in Algorithm 15 a single sample is taken, we obtain the bound in (c-4). □

5.5 Numerical results

In this section, we demonstrate the proposed methodology by solving some instances of MPECs inspired from the literature. All of the schemes were implemented in `MATLAB` on a PC with 16GB RAM and 6-Core Intel Core i7 processor (2.6GHz).

5.5.1 Numerics for SMPEC$^{\text{as}}$

In this section, we apply the schemes on a stochastic Stackelberg-Nash-Cournot equilibrium problem. The deterministic setting of the problem is derived from [168]. Consider a market with N profit-maximizing by competing in Cournot (quantities) under the Cournot assumption that the remaining firms will hold their outputs at existing levels. In addition, there is a Stackelberg firm (leader), supplying the same product which sets production levels by explicitly considering the reaction of the other N firms to its output variations. We assume that the ith Cournot firm (follower) supplies q_i units of the product while $f_i(q_i)$ denotes the total cost. In a similar fashion, suppose x denotes the output of the Stackelberg firm and let $f(x)$ denote the total cost. Next, let $p(\cdot,\omega)$ represent the random inverse demand curve. The N Cournot firms have sufficient capacity installed and thus they can wait to observe the quantities supplied by the Stackelberg firm as well as the realized demand function before making a decision on their supply quantities. For a given $x \geq 0$, let $(q_1(x,\omega),\ldots,q_N(x,\omega))$ be a set of quantities for every $\omega \in \Omega$ and each $q_i(x,\omega)$ solve the following profit maximization problem assuming that $q_j(x,\omega)$, $j \neq i$ are fixed:

$$\max_{q_i \geq 0} \quad q_i p\left(q_i + x + \textstyle\sum_{j=1, j\neq i}^{N} q_j(x,\omega), \omega\right) - f_i(q_i). \tag{129}$$

Accordingly, let $Q(x,\omega) \triangleq \sum_{i=1}^{N} q_i(x,\omega)$. In addition, we assume there exits a capacity limit x^u for x. Then x^* is said to be a Stackelberg-Nash-Cournot equilibrium solution if x^* solves

$$\underset{0\leq x\leq x^u}{\text{maximize}}\ \mathbb{E}[xp(x+Q(x,\omega),\omega)] - f(x). \tag{130}$$

We consider the case of a linear demand curve with convex quadratic cost functions. Specifically, let $p(x,\omega) = a(\omega) - bx$ and let $f_i(q) = \frac{1}{2}cq^2$ for $i = 1,\cdots,N$, and $f(x) = \frac{1}{2}dx^2$. Under this condition, the follower's objective can be shown to be strictly concave in q^i [183]. Consequently, the concatenated necessary and sufficient equilibrium conditions of the follower-level game are given by the following conditions.

$$0 \leq q \perp \nabla_q F(q) - p(x+Q(x,\omega),\omega)\mathbf{1} - p'(x+Q(x,\omega)\omega)q \geq 0, \tag{131}$$

where $F(q) = [f_1(q_1),\cdots,f_N(q_N)]'$. We observe that (131) is a strongly monotone variational inequality problem for $x \geq 0$ and for every $\omega \in \Omega$. Consequently, $q : \mathbb{R}_+ \times \Omega \to \mathbb{R}_+^N$ is a single-valued map and is convex in its first argument for every ω if c_j is quadratic and convex [181,

Prop. 4.2]. In fact, it can be claimed that $q(\cdot,\omega)$ is a piecewise C^2 and non-increasing function with $\partial_x q(x,\omega) \subset (-1,0]$ for $X \geq 0$. Consider the leader's problem (130) Consequently, we have that

$$0 \ni x \perp \mathbb{E}\left[-p(x+Q(x,\omega),\omega) + (1+\partial_x Q(x,\omega))bx - a(\omega)\right] + \nabla_x f(x) \in 0.$$

This may be viewed as the following inclusion which has been shown to be monotone [181, Thm. 4.4].

$$0 \in \mathbb{E}[T(x,\omega)] + \mathcal{N}_{\mathbb{R}^+_n},$$
$$\text{where } T(x,\omega) \triangleq \mathbb{E}\left[-p(x+Q(x,\omega),\omega)\mathbf{1} - a(\omega)\mathbf{1}\right] + \nabla_x f(x) + \{\mathbb{E}[(1+\partial_x Q(x,\omega))bx]\}.$$

Problem parameters. Suppose there are $N = 10$ Cournot firms and $c = d = 0.1$. Furthermore, $b = 1$ and $a(\omega) \sim \mathcal{U}(7.5, 12.5)$ where $\mathcal{U}(l,u)$ denotes the uniform distribution on $[l,u]$.

Algorithm parameters. We choose $\gamma_k = \frac{1}{\sqrt{k}}$ and $\eta_k = \frac{1}{\sqrt{k}}$, $\forall k \geq 1$ in (ZSOL) and $\gamma_k = \frac{1}{2k}$ and $\eta_k = \frac{1}{k}$, $\forall k \geq 1$ in (acc-ZSOL). In addition, we choose sample size $N_k = \lfloor k^{1.01} \rfloor$.
We compare the performance of (ZSOL) and (acc-ZSOL) with Nesterov's fixed smoothing scheme under the same number of iterations in Fig. 7. Next we change the size and parameters of the original game to ascertain parametric sensitivity. In Table 16, we consider a set of 12 problems where the settings, the empirical errors, and elapsed time are shown in Table 16. Note that we have access to the true solution from [168] and this is employed for computing the sub-optimality metrics. In addition, to show the performance of our proposed schemes, we consider the (SAA) scheme (utilizing the average of 100 samples) used in [181]. Let $(\omega_k)_{k=1}^K$ denote independent identically distributed (i.i.d.) samples. Then, with (SAA) we solve the following formulation of problem:

$$\max_{0\leq x\leq x^u} \ \frac{1}{K}\sum_{k=1}^{K}\left[x\cdot(a(\omega_k) - b\cdot(x+Q(x,\omega_k)))\right] - \tfrac{1}{2}dx^2$$
$$\text{subject to } 0 \leq q_i \perp (c+2b)q_i - a(w_k) + b\cdot\left(x + \textstyle\sum_{j=1,j\neq i}^{N} q_j(x,\omega_k)\right) \geq 0, \ \forall i,k.$$

This problem allows for utilizing `NLPEC` [201] in GAMS to compute a solution.

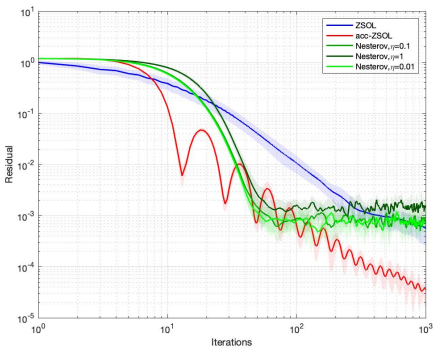

Figure 7: Trajectories for (ZSOL), (ac-ZSOL) and Nesterov on the convex SMPEC$^{\text{as}}$

Insights.

- *Scalability.* Both (ZSOL) and (acc-ZSOL) show far better scalability in terms of N with modest impact on accuracy and run-time. SAA schemes on the other hand grow by a factor of 10 when number of firms double. In fact, for $N = 20$, the SAA framework requires CPU time which is between 50 and 100 times greater than that required by the zeroth-order schemes.

Table 16: Errors and time comparison of the three schemes with different parameters

			(**ZSOL**)		(**acc-ZSOL**)		(**SAA**)	
			$f^* - f(\bar{x}_K)$	Time	$f^* - f(x_K)$	Time	$f^* - f(\hat{x})$	Time
$N = 10$	$b = 1$	$c = 0.05$	1.2e-3	0.1	6.6e-5	5.6	5.4e-4	130.2
		$c = 0.1$	8.2e-4	0.1	4.8e-5	5.4	4.2e-4	109.2
	$b = 0.5$	$c = 0.05$	1.7e-3	0.1	7.0e-5	5.4	3.8e-4	122.5
		$c = 0.1$	1.2e-3	0.1	6.3e-5	5.5	2.2e-4	116.8
$N = 20$	$b = 1$	$c = 0.05$	4.5e-4	0.1	2.6e-5	5.6	2.6e-4	426.7
		$c = 0.1$	4.0e-4	0.1	1.3e-5	5.7	5.7e-4	443.1
	$b = 0.5$	$c = 0.05$	6.3e-4	0.1	2.3e-5	5.7	4.8e-4	419.1
		$c = 0.1$	4.2e-4	0.1	2.9e-5	5.6	3.1e-4	450.0
$N = 100$	$b = 1$	$c = 0.05$	9.9e-5	0.2	3.2e-6	7.5	–	–
		$c = 0.1$	2.3e-5	0.2	1.3e-6	7.5	–	–
	$b = 0.5$	$c = 0.05$	2.6e-4	0.2	4.7e-6	7.9	–	–
		$c = 0.1$	2.5e-5	0.2	1.4e-6	7.5	–	–
$N = 1000$	$b = 1$	$c = 0.05$	2.2e-5	0.6	3.6e-7	27.9	–	–
		$c = 0.1$	1.7e-6	0.6	8.3e-8	28.8	–	–
	$b = 0.5$	$c = 0.05$	2.5e-5	0.6	3.1e-7	29.1	–	–
		$c = 0.1$	1.4e-6	0.6	8.9e-8	28.4	–	–
$N = 10000$	$b = 1$	$c = 0.05$	1.0e-5	4.6	5.2e-7	403.5	–	–
		$c = 0.1$	6.0e-6	4.5	3.8e-8	392.4	–	–
	$b = 0.5$	$c = 0.05$	1.1e-5	4.7	5.6e-8	334.2	–	–
		$c = 0.1$	7.1e-6	4.6	2.7e-8	399.7	–	–

The error and time pf (ZSOL) and (acc-ZSOL) are the average results of 20 runs ('–' means the running time is over 3600)

- *Accuracy.* The accelerated scheme provides nearly 10 times more accurate solutions than the unaccelerated scheme at a modest computational cost.
- *Comparison of accelerated schemes.* Figure 7 demonstrates the benefits of diminishing smoothing sequences as the scheme suggested in [192] degenerates for different values of the fixed smoothing parameter. Notably, (acc-ZSOL) shows no such degeneration and progressively improves in function value.

5.5.2 Numerics for SMPEC$^{\text{exp}}$

A convex implicit function Here, we consider a situation of the previous example when the lower level is SVI, which means we consider the following lower level problem for each q_i,

$$\max_{q_i \geq 0} \mathbb{E}[q_i(a(\omega) - b(q_i + x + \textstyle\sum_{j \neq i} q_j(x))] - \frac{1}{2} c q_i^2,$$

and accordingly, the upper level problem is as follows

$$\max_{0 \leq x \leq x^u} \mathbb{E}\left[x(a(\omega) - b(x + \textstyle\sum_{i=1}^N q_i(x)))\right] - \frac{1}{2} d x^2.$$

Similarly, this implicit function can be shown a convex function. We assume $b = 0.01$ and $c = 3$ here, other parameters are the same as in the previous section. It can be shown that $\mu_F = 3.01$ and $L_F = 3.11$. For the algorithm parameters, we suppose $\gamma_k = \frac{1}{\sqrt{k}}$ and $\eta_k = \frac{1}{\sqrt{k}}$ for (ZSOL). In (ZSOL) we run 10^3 iterations. In the lower level's variance-reduced stochastic approximation scheme, we choose steplength $\alpha = 0.15$, sampling rate $\rho = \frac{1}{1.5}$ and the sample size $M_t = \lceil 10^{-4} \cdot 1.5^t \rceil$. Thus we may calculate that $\tau \geq 4.9$ and then we choose $t_k = \lceil 5 \ln(k+1) \rceil$. In Fig. 8, we show the trajectories for (ZSOL) under various algorithm parameters.

Again, we compare the errors and time between (ZSOL) and (SAA) in Table 17. Here, with (SAA)

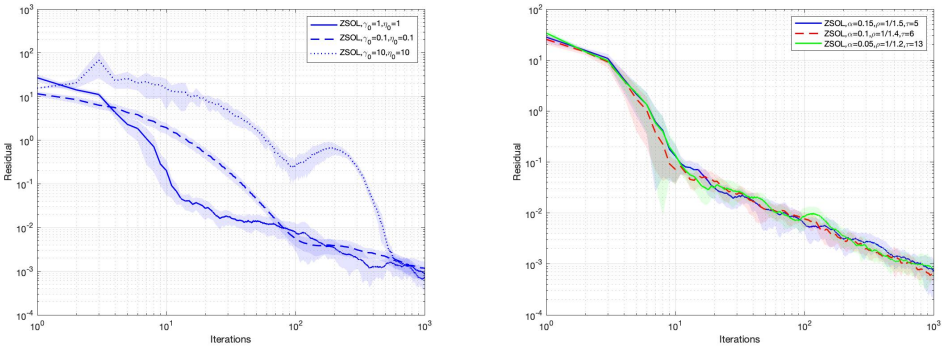

Figure 8: Trajectories for (ZSOL) on the convex SMPEC$^{\text{exp}}$

we solve the following optimization problem

$$\begin{aligned}&\underset{0\le x\le x^u}{\text{maximize}}\ \tfrac{1}{K}\sum_{k=1}^{K}[x\cdot(a(\omega_k)-b\cdot(x+Q(x)))]-\tfrac{1}{2}dx^2\\&\text{subject to } 0\le q_i\perp \tfrac{1}{L}\textstyle\sum_{\ell=1}^{L}\left[(c+2b)q_i-a(w_\ell)+b\cdot\left(x+\sum_{j=1,j\neq i}^{N}q_j(x)\right)\right]\ge 0,\ \forall i.\end{aligned}$$

In (SAA), we use 10^3 samples in both the upper and lower level problems.

Table 17: Errors and time comparison of (ZSOL) and (SAA) with various parameters

			(**ZSOL**)		(**SAA**)	
			$f^*-f(\bar{x}_K)$	Time	$f^*-f(\hat{x})$	Time
$N=10^2$	$b=0.01$	$c=3$	6.9e-4	0.1	2.2e-4	0.05
		$c=5$	3.7e-4	0.1	2.4e-4	0.05
	$b=0.02$	$c=3$	8.1e-4	0.1	7.3e-4	0.05
		$c=5$	3.5e-4	0.1	4.0e-4	0.05
$N=10^3$	$b=0.01$	$c=3$	7.0e-4	0.4	7.0e-4	1.2
		$c=5$	4.3e-4	0.4	5.0e-4	1.1
	$b=0.02$	$c=3$	8.0e-4	0.4	6.8e-4	1.2
		$c=5$	4.7e-4	0.4	4.2e-4	1.2
$N=10^4$	$b=0.01$	$c=3$	5.1e-4	5.8	7.3e-4	88.6
		$c=5$	2.5e-4	5.2	5.4e-4	85.7
	$b=0.02$	$c=3$	6.4e-4	5.6	4.3e-4	93.5
		$c=5$	3.1e-4	5.3	4.7e-4	87.3
$N=10^5$	$b=0.01$	$c=3$	8.7e-4	45.6	–	–
		$c=5$	6.5e-4	47.1	–	–
	$b=0.02$	$c=3$	9.7e-4	46.3	–	–
		$c=5$	7.5e-4	46.7	–	–

The error and time of (ZSOL) are the average results of 20 runs ('–' means the running time is over 3600)

Insights. Again we observe that the CPU times grow by a factor of 4 for the zeroth-order schemes; however the SAA schemes show a growth in time of more than 150 when N changes from 10 to 100. Both approaches provide similar accuracy but zeroth-order schemes require less than 1s in CPU time while the SAA framework requires approximately 15s for $N=100$. The accuracy of ZSOL is relatively robust to changing steplength and sampling rates at the lower level but does tend to be sensitive to changing the initial steplength at the upper level; however, as the scheme progresses, the impact of initial steplengths tends to be muted.

Table 18: Errors and confidence intervals for high dimensional problems from Table 16 and 17

			(**ZSOL**)		(**acc-ZSOL**)	
			$f^* - f(\bar{x}_K)$	CI	$f^* - f(x_K)$	CI
Table 16	$b = 1$	$c = 0.05$	1.0e-5	[0.9e-5,1.1e-5]	5.2e-7	[5.0e-7,5.4e-7]
		$c = 0.1$	6.0e-6	[5.9e-6,6.1e-6]	3.8e-8	[3.4e-8,4.2e-8]
$N = 10^4$	$b = 0.5$	$c = 0.05$	1.1e-5	[1.0e-5,1.2e-5]	5.6e-8	[5.2e-8,6.0e-8]
		$c = 0.1$	7.1e-6	[7.0e-6,7.2e-6]	2.7e-8	[2.4e-8,3.0e-8]
Table 17	$b = 0.01$	$c = 3$	8.7e-4	[7.5e-4,9.9e-4]	–	–
		$c = 5$	6.5e-4	[5.9e-4,7.1e-4]	–	–
$N = 10^5$	$b = 0.02$	$c = 3$	9.7e-4	[8.0e-4,1.1e-3]	–	–
		$c = 5$	7.5e-4	[6.4e-4,8.6e-4]	–	–

5.5.3 A nonconvex implicit function

The second example is inspired from [202] and is a bilevel problem with a strongly monotone mapping in the lower level. We add a stochastic component in the lower level to make the mapping expectation-valued. Formally, this problem is defined as follows.

$$\begin{aligned} \min_{x} \quad & -x_1^2 - 3x_2 - 4y_1 + y_2^2 \\ \text{subject to} \quad & x_1^2 + 2x_2 \leq 4 \\ & \leq x_1 \leq 1 \\ & 0 \leq x_2 \leq 2 \end{aligned}$$

where y is a solution to

$$\begin{aligned} \min_{y} \quad & \mathbb{E}[2x_1^2 + y_1^2 + y_2^2 - \xi(\omega)y_2] \\ \text{subject to} \quad & x_1^2 - 2x_1 + x_2^2 - 2y_1 + y_2 \geq -3 \\ & x_2 + 3y_1 - y_2 \geq 4 \\ & y_1 \geq 0, y_2 \geq 0, \end{aligned} \tag{132}$$

where we assume $\xi(\omega) \sim \mathcal{U}(4, 6)$. Based on the rule, we run 10^4 iterations and choose $\eta = 10^{-2}$, $\gamma = 10^{-3}$ in (VR-ZSOL). In addition, we choose $\alpha_0 = 1$ and $\alpha_t = \frac{\alpha_0}{t+0.01}$ for $t = 0, 1, \ldots, t_k - 1$ in the stochastic approximation method applied to the lower level. We compare the performance of (VR-ZSOL) on this problem in Fig. 9 for varying algorithm parameters, all of which suggest that the resulting sequences steadily converge to the global minimizer. To test the power of (ZSOL) on different problems, we change the objective function of upper level and lower level to $-ax_1^2 - bx_2^2 - 3x_2 - 4y_1 + y_2^2$ and $\mathbb{E}[2x_1^2 + cy_1^2 + dy_2^2 - \xi(\omega)y_2]$, respectively. Then we vary the values of a, b, c and d. For comparison, we also run each problem using solvers `NLPEC` and `BARON` [203, 204] on the NEOS Server [205–207]. We record the empirical errors of each scheme for 9 different settings, as shown in Table 17. In (VR-ZSOL), we use 10^4 samples in each test problem.

Insights. From Fig. 9, we observe that while all of the implementations perform well, large initial steplengths at the lower-level tend to lead relatively worse compared to more modest steplengths. Table 17 is instructive in that it shows that (VR-ZSOL) produces values close to the global minimum as obtained by `BARON` for all nine problem instances. Notably, solvers such as `NLPEC` are equipped with convergence guarantees to first-order points and tend to provide somewhat poorer values upon termination.

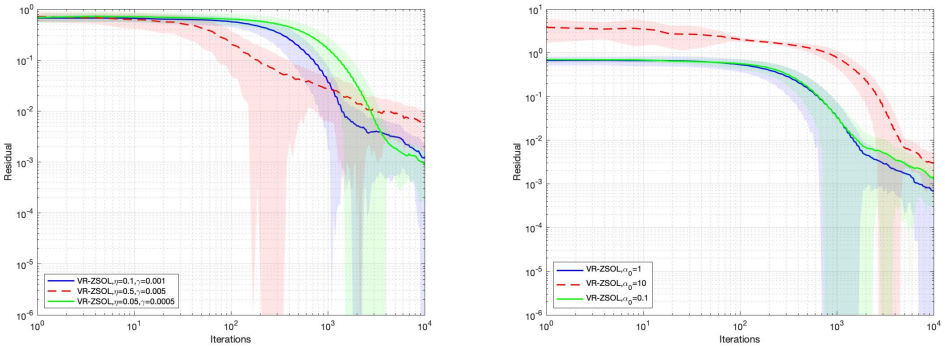

Figure 9: Trajectories for (VR-ZSOL) on the non-convex SMPEC$^{\text{exp}}$

Table 19: Errors comparison of the three schemes with different parameters

		ZSOL	NLPEC	BARON
		$f(x_K)$	local optimum	global optimum
$(a,b)=(1,0)$	$(c,d)=(1,1)$	-7.50	-7.20	-7.50
	$(c,d)=(2,2)$	-9.23	-9.04	-9.23
	$(c,d)=(3,3)$	-9.25	-9.10	-9.25
$(a,b)=(5,0)$	$(c,d)=(1,1)$	-11.50	-7.20	-11.50
	$(c,d)=(2,2)$	-13.23	-9.04	-13.23
	$(c,d)=(3,3)$	-13.25	-9.10	-13.25
$(a,b)=(10,0)$	$(c,d)=(1,1)$	-16.48	-7.20	-16.50
	$(c,d)=(2,2)$	-18.20	-9.04	-18.23
	$(c,d)=(3,3)$	-18.23	-9.10	-18.25

The error of (ZSOL) in the table is the average results of 20 runs

5.5.4 More academic problems

We test our schemes on several academic MPEC problems from the literature. In all the test problems, the VI is strongly monotone which means the lower-level decision is uniquely determined by a $\mathbf{x} \in \mathcal{X}$. We use the same algorithm parameters as in 5.5.3.

Problem 1. This problem is described in [170].

$$f(\mathbf{x},\mathbf{y}) = r_1(x) - xp(x+y_1+y_2+y_3+y_4),$$

where $r_i(v) = c_i v + \frac{\beta_i}{\beta_i+1} K_i^{1/\beta_i} v^{(1+\beta_i)/\beta_i}$, $p(Q) = 5000^{1/\gamma} Q^{-1/\gamma}$, c_i, β_i, K_i, $i = 1, \cdots, 5$ are given positive parameters in Table 21, γ is a positive parameter, $Q = x + y_1 + y_2 + y_3 + y_4$.

$$\mathcal{X} = \{0 \le x \le L\}.$$

$$F(\mathbf{x},\mathbf{y}) = \begin{pmatrix} \nabla r_2(y_1) - p(Q) - y_1 \nabla p(Q) \\ \vdots \\ \nabla r_5(y_4) - p(Q) - y_4 \nabla p(Q) \end{pmatrix}.$$

$$\mathcal{Y} = \{0 \le y_j \le L, \quad j = 1,2,3,4\}.$$

Based on different values of L and γ, we show the optimal value and optimal solution found by (**ZSOL**).
The following three examples were tested in [170].

Table 20: Parameter specification for problem 1

i	1	2	3	4	5
c_i	10	8	6	4	2
K_i	5	5	5	5	5
β_i	1.2	1.1	1.0	0.9	0.8

Table 21: Results comparison with solutions from the literature

Problem		**ZSOL**		Literature	
		f^*	x^*	f^*	x^*
Problem 1	$L = 150$, $\gamma = 1.0$	-343.35	55.57	-343.35	55.55
	$L = 150$, $\gamma = 1.1$	-203.15	42.57	-203.15	42.54
	$L = 150$, $\gamma = 1.3$	-68.14	24.19	-68.14	24.14
Problem 2		-1.00	(0.50,0.50)	-1.00	(0.50,0.50)
Problem 3		0.01	(0.00,0.00)	0.01	(0.00,0.00)
Problem 4		0.00	(5.00,8.99)	0.00	(5.00,9.00)
Problem 5	$0.5((y_1-3)^2+(y_2-4)^2)$	3.20	4.06	3.20	4.06
	$0.5((y_1-3)^2+(y_2-4)^2+(y_3-1)^2)$	3.45	5.13	3.45	5.15
	$0.5((y_1-3)^2+(y_2-4)^2+10y_4^2)$	4.60	2.39	4.60	2.39

Problem 2.

$$f(\mathbf{x},\mathbf{y}) = x_1^2 - 2x_1 + x_2^2 - 2x_2 + y_1^2 + y_2^2.$$
$$\mathcal{X} = \{0 \le x_i \le 2, \quad i = 1,2\}.$$
$$F(\mathbf{x},\mathbf{y}) = \begin{pmatrix} 2y_1 - 2x_1 \\ 2y_2 - 2x_2 \end{pmatrix}.$$
$$\mathcal{Y} = \{(y_j - 1)^2 \le 0.25, \quad j = 1,2\}.$$

Problem 3.

$$f(\mathbf{x},\mathbf{y}) = 2x_1 + 2x_2 - 3y_1 - 3y_2 - 60 + R[\max\{0, x_1 + x_2 + y_1 - 2y_2 - 40\}]^2.$$
$$\mathcal{X} = \{0 \le x_i \le 50, \quad i = 1,2\}.$$
$$F(\mathbf{x},\mathbf{y}) = \begin{pmatrix} 2y_1 - 2x_1 + 40 \\ 2y_2 - 2x_2 + 40 \end{pmatrix}.$$
$$\mathcal{Y} = \{-10 \le y_j \le 20,\ x_j - 2y_j - 10 \ge 0, \quad j = 1,2\}.$$

Problem 4.

$$f(\mathbf{x},\mathbf{y}) = \tfrac{1}{2}((x_1 - y_1)^2 + (x_2 - y_2)^2).$$
$$\mathcal{X} = \{0 \le x_i \le 10, \quad i = 1,2\}.$$
$$F(\mathbf{x},\mathbf{y}) = \begin{pmatrix} -34 + 2y_1 + \frac{8}{3}y_2 \\ -24.25 + 1.25y_1 + 2y_2 \end{pmatrix}.$$
$$\mathcal{Y} = \{-x_{3-j} - y_j + 15 \ge 0, \quad j = 1,2\}.$$

The next problem is taken from [170]. In all tests, the only difference is the objective function.
Problem 5.

$$\mathcal{X} = \{0 \le x \le 10\}.$$

$$F(\mathbf{x},\mathbf{y}) = \begin{pmatrix} (1+0.2x)y_1 - (3+1.333x) - 0.333y_3 + 2y_1y_4 - y_5 \\ (1+0.1x)y_2 - x + y_3 + 2y_2y_4 - y_6 \\ 0.333y_1 - y_2 + 1 - 0.1x \\ 9 + 0.1x - y_1^2 - y_2^2 \\ y_1 \\ y_2 \end{pmatrix}.$$
$$\mathcal{Y} = \{y_j \geq 0, \quad j = 3,4,5,6\}.$$

Consider the stochastic N-dimensional counterpart of Problem 1, defined as follows.

$$f(\mathbf{x},\mathbf{y}) = \mathbb{E}\left[r_1(x) - xp\left(x + \sum_{i=1}^{n} y_i, \omega\right)\right],$$

where $r_i(v) = c_i v + \frac{\beta_i}{\beta_i+1} K_i^{1/\beta_i} v^{(1+\beta_i)/\beta_i}$, $p(Q,\omega) = 5000^{1/\gamma(\omega)} Q^{-1/\gamma(\omega)}$, $c_i = 6$, $\beta_i = 1$, $K_i = 5$, $i = 1, \cdots, 5$, $\gamma(\omega) \in U[0.9, 1.1]$ is a positive parameter, $Q = x + \sum_{i=1}^{N} y_i$.

$$\mathcal{X} = \{0 \leq x \leq L\}.$$
$$F(\mathbf{x},\mathbf{y},\omega) = \begin{pmatrix} \nabla r_2(y_1) - p(Q,\omega) - y_1 \nabla p(Q,\omega) \\ \vdots \\ \nabla r_n(y_n) - p(Q,\omega) - y_n \nabla p(Q,\omega) \end{pmatrix}.$$
$$\mathcal{Y} = \{0 \leq y_j \leq L, \quad j = 1, \cdots, n\}.$$

Table 22: Results of high-dimensional counterparts

Problem	N	(**ZSOL**)					(**SAA**)		
		$\hat{f}(x_K)$	CI	Time	$\underline{lb}$	CI	$\hat{f}(\hat{x})$	CI	Time
Problem 1	5	-462.6	[-463.1,-462.1]	0.8	-462.8	[-464.0,-461.5]	-461.9	[-463.1,-460.7]	5.3
	10	-174.4	[-174.6,-174.2]	0.9	-174.7	[-175.2,-174.2]	-174.2	[-174.8,-173.6]	23.3
	100	-5.101	[-5.105,-5.097]	1.3	–	–	–	–	–
	1000	-0.071		5.2	–	–	–	–	–
Problem 2	2	-0.882	[-0.883,-0.881]	0.6	-0.883	[-0.886,-0.880]	-0.882	[-0.886,-0.878]	4.2
	10	-4.408	[-4.410,-4.406]	0.9	-4.408	[-4.414,-4.402]	-4.406	[-4.414,-4.398]	29.6
	100	-44.0		5.5	-44.1		3544.4		
	1000	-416.7		98.1	–		–		

The stochastic N-dimensional counterpart of Problem 2.

$$\mathbb{E}[f(\mathbf{x},\mathbf{y}(\omega))], \text{ where } f(x,y(\omega)) = \|x - \mathbf{1}\|^2 + \|y(\omega)\|^2$$
$$\mathcal{X} = \{0 \leq x_i \leq 2, \quad i = 1, \cdots, n\}.$$
$$F(\mathbf{x},\mathbf{y},\omega) = \left(2y - 2x + \omega\right).$$
$$\mathcal{Y} = \{\|y - \mathbf{1}\|^2 \leq 0.25\}.$$

where $\omega \in U[-0.5, 0.5]$.

6 Commercial Development

Power systems operational problems have traditionally been modeled as linear and mixed-integer linear programs. Both can be resolved by commercial mixed-integer programming (MIP) solvers such as `cplex` and `gurobi`. Increasingly, the models that lead to such linear formulations do not suffice. The GO competition has introduced a formulation for managing resources but such problems are complicated by uncertainty, nonlinearity, nonconvexity, and discreteness. Existing solvers can at best process small instances of such problems, if at all. This requires a comprehensive development of new tools. In this section, we discuss the pathway towards commercial development, emphasizing both the barriers and opportunities.

(a) *Market.* Possibly, the first question is whether such a tool that can process large-scale electrical networks, a range of contingency instances, and allow for a broad range of operational models is even of value. We believe that in the face of climate goals, the penetration of renewables is expected to grow even further. This brings forth new challenges in managing power systems at the bulk level, paerticularly in terms of the nature of models employed and the need to contend with uncertainty. The latter concern is itself becoming far more relevant with the growth in climate change impacts and demand variability. We believe such tools would be of particular value to system operators, independent power producers, utilities, and financial participants. In short, almost all stake-holders in the power market would find it beneficial to have such a tool. More generally, the algorithmic underpinnings of such tools have significant benefits that extend well beyond the power systems arena, spanning a range of application regimes including manufacturing, energy and building systems, logistics and supply-chain management, amongst others.

(b) *Commercial tool.* The commercial tool of interest would have the following broad features:

(i) *Computational optimization engine.* An underlying computational engine that can process two-stage stochastic optimization problems complicated by discreteness, nonlinearity, nonsmoothness, risk, and nonconvexity (in some structured form). This engine will probably be coded in a lower level language and may need to employ some of the linear algebra subroutines (akin to `lapack` etc.).

(ii) *Parallel and asynchronous implementations.* The engine in (i) should be implementable in a multi-core multi-processing environment that takes advantage of parallelism and allow for asynchronous and distributed computation.

(iii) *Flexibility of modeling platform.* The framework should be flexible and user-friendly in terms of building and processing models that can then allow for being processed.

(iv) *Management of contingencies.* The tool should allow for managament of a large number of contingencies each of which requires data for the entire network (which could be of the order of $1e5$ nodes or so).

(v) *Bounds and sensitivity analysis.* Finally, the tool should provide upper and lower bounds to provide some notion of quality for the solution. In addition, it should allow for conducting sensitivity analysis to parameters.

(c) *Progress.* The considering the overall framework for resolving large-scale power systems operation problems, complicated by risk, uncertainty, nonlinearity, and discreteness. We believe that we have made significant progress in developing such a branching framework. Yet, there are at least

two significant questions that remain to be addressed for building the computational engine and the parallel implementations as part of the tool described above. First, any serious attempt at building such a tool requires contending with the AC power flow equations in a systematic fashion, that may include (but is not restricted to) modeling enhancements or the introduction of reformulations and cuts. Second, scalability concerns require building parallel implementations that can leverage the inherent parallelism in the sampling-based schemes employed. This will allow for building schemes whose serial computation time can allow for commercial implementations for massive networks and a large number of contingencies.

(d) *Barriers and opportunities* . **Barriers:** We envisage two key barriers to our progress in developing a computational engine: (i) A key barrier to progress has been the ability to get students to aid in the development in the face of COVID-19. Most of our student body is international and a significant portion has been affected both in terms of coming to initiate their studies as well as in their ability to carry out their program. (ii) A second challenge arises in obtaining high quality programmers who can take working prototypes in `Python` and develop counterparts in lower level languages. **Opportunities.** At Penn. State, we have a broad set of resources in terms of the Institute of Computational and Data Sciences (ICDS) that manages our computing facility and has resources for aiding in the computational development. In addition, we also have a large cadre of students in computer science and other engineering disciplines that we hope to engage, despite the impact of COVID-19.

7 Summary and future work

In this section, we summarize the main contributions of our work and discuss some future directions.

7.1 Summary of contributions

The main contributions can be quantified as follows.
(i) Stochastic and risk-averse nonsmooth convex optimization. In the context of nonsmooth two-stage stochastic/risk-averse optimization, we develop a smoothed accelerated variance-reduced scheme that is characterized by an optimal rate of a convergence of $\mathcal{O}(1/k)$ and a near optimal sample complexity of $\mathcal{O}(1/\epsilon^{2+\delta})$ to compute an ϵ-optimal solution where $\delta > 0$. Notably, this framework can contend with a broad range of two-stage stochastic nonlinear optimization problems where the second-stage problems may be nonlinear convex programs (with a suitable structure).

(ii) Stochastic mixed-integer nonlinear optimization. When one overlays discreteness in the first and second-stage problems, we apply a stochastic branch-and-bound scheme in which the relaxations are continuous and convex two-stage optimization problems. We leverage the techniques from (i) and show that this framework can contend with a broad range of two-stage mixed-integer nonlinear programs (albeit with convex relaxations). In particular, the scheme provides sampling-based upper and lower bounds. In addition, we demonsrate that the introduction of cuts provides further improvements on the quality of the solution developed.

(iii) Stochastic mathematical programs with equilibrium constraints. Complementarity constraints emerge in some models presented in the GO competition. Such problems lead to nonconvexity and ill-posedness and to date, there have been no efficient schemes for their resolution even in deterministic regimes. We present an implicit framework that can contend with a subclass of

regimes where the lower-level problems are characterized by strong monotonicity of the lower-level map. Notably, we present efficient schemes for computing an ϵ-solution for either a first or a two-stage stochastic MPEC when the implicit problem is either convex or nonconvex. Notably, when one overlays discreteness, a branching scheme can again be employed where the relaxations can be resolved as per the schemes developed in this section.

7.2 Future work on methodology and computation.

(a) *Nonconvexity.* As part of our efforts on addressing complementarity constraints, we have been able to deal with one form of nonconvexity. But contending with nonconvexities arising from AC power flow equations remain. We will consider two approaches to address such a question. The first lies in assessing whether computing stationary points via stochastic variants of sequential quadratic programming schemes (as developed in our prior work [208]) may be employed. The second avenue may consider utilizing a combination of branching schemes for computing under and over estimators, similar to those employed in the commercial solver `baron`.

(b) *Parallel implementations.* A key concern in developing branching schemes lies in scalability. There has been an effort to develop parallel implementations of branching schemes. Our intention is to consider this avenue in developing scalable tools for contending with mixed-integer variants of stochastic optimization problems, complicated by risk, nonlinearity, and the presence of complementarity constraints.

References

[1] T. Yong and R. Lasseter, "Stochastic optimal power flow: formulation and solution," in *Power Engineering Society Summer Meeting, 2000. IEEE*, vol. 1, pp. 237–242, IEEE, 2000.

[2] H. Gangammanavar, S. Sen, and V. M. Zavala, "Stochastic optimization of sub-hourly economic dispatch with wind energy," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 949–959, 2016.

[3] R. Doherty and M. O'malley, "A new approach to quantify reserve demand in systems with significant installed wind capacity," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 587–595, 2005.

[4] J. M. Morales, A. J. Conejo, and J. Pérez-Ruiz, "Economic valuation of reserves in power systems with high penetration of wind power," *IEEE Transactions on Power Systems*, vol. 24, no. 2, pp. 900–910, 2009.

[5] T. Summers, J. Warrington, M. Morari, and J. Lygeros, "Stochastic optimal power flow based on convex approximations of chance constraints," in *Power Systems Computation Conference (PSCC), 2014*, pp. 1–7, IEEE, 2014.

[6] J. Warrington, P. Goulart, S. Mariéthoz, and M. Morari, "Policy-based reserves for power systems," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4427–4437, 2013.

[7] R. A. Jabr, S. Karaki, and J. A. Korbane, "Robust multi-period opf with storage and renewables," *IEEE Transactions on Power Systems*, vol. 30, no. 5, pp. 2790–2799, 2015.

[8] R. A. Jabr, "Adjustable robust opf with renewable energy sources," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4742–4751, 2013.

[9] D. Munoz-Alvarez, E. Bitar, L. Tong, and J. Wang, "Piecewise affine dispatch policies for economic dispatch under uncertainty," in *PES General Meeting— Conference & Exposition, 2014 IEEE*, pp. 1–5, IEEE, 2014.

[10] X. Bai, L. Qu, and W. Qiao, "Robust ac optimal power flow for power networks with wind power generation," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 4163–4164, 2016.

[11] R. Louca and E. Bitar, "Stochastic ac optimal power flow with affine recourse," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pp. 2431–2436, IEEE, 2016.

[12] D. Bienstock, M. Chertkov, and S. Harnett, "Chance-constrained optimal power flow: Risk-aware network control under uncertainty," *SIAM Review*, vol. 56, no. 3, pp. 461–495, 2014.

[13] E. Ela and M. O'Malley, "Studying the variability and uncertainty impacts of variable generation at multiple timescales," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1324–1333, 2012.

[14] P. P. Varaiya, F. F. Wu, and J. W. Bialek, "Smart operation of smart grid: Risk-limiting dispatch," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 40–57, 2011.

[15] R. Rajagopal, E. Bitar, P. Varaiya, and F. Wu, "Risk-limiting dispatch for integrating renewable power," *International Journal of Electrical Power & Energy Systems*, vol. 44, no. 1, pp. 615–628, 2013.

[16] C. Peng and Y. Hou, "Risk-limiting dispatch with operation constraints," in *PES General Meeting— Conference & Exposition, 2014 IEEE*, pp. 1–5, IEEE, 2014.

[17] B. Zhang, R. Rajagopal, and D. Tse, "Network risk limiting dispatch: Optimal control and price of uncertainty," *IEEE Transactions on Automatic Control*, vol. 59, no. 9, pp. 2442–2456, 2014.

[18] G. B. Sheble and G. N. Fahd, "Unit commitment literature synopsis," *Power Systems, IEEE Transactions on*, vol. 9, no. 1, pp. 128–135, 1994.

[19] G. Lauer, N. Sandell Jr, D. Bertsekas, and T. Posbergh, "Solution of large-scale optimal unit commitment problems," *Power Apparatus and Systems, IEEE Transactions on*, no. 1, pp. 79–86, 1982.

[20] A. I. Cohen and M. Yoshimura, "A branch-and-bound algorithm for unit commitment," *IEEE Trans. Power Appar. Syst.;(United States)*, vol. 102, no. 2, 1983.

[21] A. Merlin and P. Sandrin, "A new method for unit commitment at electricité de france," *IEEE Trans. Power Appar. Syst.;(United States)*, vol. 102, no. 5, 1983.

[22] S. Tong and S. Shahidehpour, "Combination of lagrangian-relaxation and linear-programming approaches for fuel-constrained unit-commitment problems," in *IEE Proceedings C (Generation, Transmission and Distribution)*, vol. 136, pp. 162–174, IET, 1989.

[23] K. Aoki, M. Itoh, T. Satoh, K. Nara, and M. Kanezashi, "Optimal long-term unit commitment in large scale systems including fuel constrained thermal and pumped-storage hydro," *Power Systems, IEEE Transactions on*, vol. 4, no. 3, pp. 1065–1073, 1989.

[24] S. Feltenmark, K. C. Kiwiel, and P.-O. Lindberg, "Solving unit commitment problems in power production planning," in *Operations Research Proceedings 1996*, pp. 236–241, Springer, 1997.

[25] Q. P. Zheng, J. Wang, P. M. Pardalos, and Y. Guan, "A decomposition approach to the two-stage stochastic unit commitment problem," *Annals of Operations Research*, vol. 210, no. 1, pp. 387–410, 2013.

[26] A. Jalilzadeh, U. V. Shanbhag, J. H. Blanchet, and P. W. Glynn, "Smoothed variable sample-size accelerated proximal methods for nonsmooth stochastic convex programs," *arXiv preprint arXiv:1803.00718*, 2018.

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

[28] H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, vol. 35 of *Applications of Mathematics (New York)*. second ed., 2003.

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

[30] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming, SIAM, Philadelphia*. 2009.

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

[32] S. Ghadimi and G. Lan, "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2061–2089, 2013.

[33] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Mathematical Programming*, vol. 156, no. 1-2, pp. 59–99, 2016.

[34] A. Jofré and P. Thompson, "On variance reduction for stochastic smooth convex optimization with multiplicative noise," *arXiv preprint arXiv:1705.02969*, 2017.

[35] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

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

[37] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 ed., 2014.

[38] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *International Conference on Machine Learning*, pp. 71–79, 2013.

[39] G. Lan, "An optimal method for stochastic composite optimization," *Mathematical Programming*, vol. 133, no. 1, pp. 365–397, 2012.

[40] S. Ghadimi and G. Lan, "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1469–1492, 2012.

[41] C. D. Dang and G. Lan, "Stochastic block mirror descent methods for nonsmooth and stochastic optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 856–881, 2015.

[42] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Mathematical Programming*, vol. 146, no. 1-2, pp. 37–75, 2014.

[43] P. Dvurechensky and A. Gasnikov, "Stochastic intermediate gradient method for convex problems with stochastic inexact oracle," *Journal of Optimization Theory and Applications*, vol. 171, no. 1, pp. 121–145, 2016.

[44] F. Yousefian, A. Nedić, and U. V. Shanbhag, "On stochastic gradient and subgradient methods with adaptive steplength sequences," *Automatica*, vol. 48, no. 1, pp. 56–67, 2012.

[45] D. Newton, R. Pasupathy, and F. Yousefian, "Recent trends in stochastic gradient descent for machine learning and big data," in *Proceedings of the 2018 Winter Simulation Conference*, pp. 366–380, IEEE Press, 2018.

[46] U. V. Shanbhag and J. H. Blanchet, "Budget-constrained stochastic approximation," in *Proceedings of the 2015 Winter Simulation Conference, Huntington Beach, CA, USA, December 6-9, 2015*, pp. 368–379, 2015.

[47] A. Jalilzadeh and U. V. Shanbhag, "eg-VSSA: An extragradient variable sample-size stochastic approximation scheme: Error analysis and complexity trade-offs," in *Winter Simulation Conference, WSC 2016.*, pp. 690–701, 2016.

[48] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical programming*, vol. 103, no. 1, pp. 127–152, 2005.

[49] Y. Nesterov, "Excessive gap technique in nonsmooth convex minimization," *SIAM Journal on Optimization*, vol. 16, no. 1, pp. 235–249, 2005.

[50] Q. Tran-Dinh, O. Fercoq, and V. Cevher, "A smooth primal-dual optimization framework for nonsmooth composite convex minimization," *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 96–134, 2018.

[51] R. I. Boţ and C. Hendrich, "A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems," *Computational Optimization and Applications*, vol. 54, no. 2, pp. 239–262, 2013.

[52] O. Devolder, F. Glineur, and Y. Nesterov, "Double smoothing technique for large-scale linearly constrained convex optimization," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 702–727, 2012.

[53] Q. Tran-Dinh, "Adaptive smoothing algorithms for nonsmooth composite convex minimization," *Computational Optimization and Applications*, vol. 66, no. 3, pp. 425–451, 2017.

[54] H. Ouyang and A. Gray, "Stochastic smoothing for nonsmooth minimizations: Accelerating SGD by exploiting structure," *arXiv preprint arXiv:1205.4481*, 2012.

[55] W. Zhong and J. Kwok, "Accelerated stochastic gradient method for composite regularization," in *Artificial Intelligence and Statistics*, pp. 1086–1094, 2014.

[56] Q. Van Nguyen, O. Fercoq, and V. Cevher, "Smoothing technique for nonsmooth composite minimization with linear operator," *arXiv preprint arXiv:1706.05837*, 2017.

[57] F. Orabona, A. Argyriou, and N. Srebro, "Prisma:proximal iterative smoothing algorithm," *arXiv:1206.2372*, 2012.

[58] R. I. Boţ and C. Hendrich, "A variable smoothing algorithm for solving convex optimization problems," *Top*, vol. 23, no. 1, pp. 124–150, 2015.

[59] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA: SIAM, 2017.

[60] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances in neural information processing systems*, pp. 1458–1466, 2011.

[61] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bull. Soc. Math. France*, vol. 93, no. 2, pp. 273–299, 1965.

[62] C. Planiden and X. Wang, "Strongly convex functions, Moreau envelopes, and the generic nature of convex functions with strong minimizers," *SIAM Journal on Optimization*, vol. 26, no. 2, pp. 1341–1364, 2016.

[63] F. Facchinei and J.-S. Pang, *Finite-dimensional Variational Inequalities and Complementarity Problems. Vols. I,II.* Springer Series in Operations Research, New York: Springer-Verlag, 2003.

[64] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.*, vol. 40, p. 120145, May 2011.

[65] A. Beck and M. Teboulle, "Smoothing and first order methods: A unified framework," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 557–580, 2012.

[66] D. Veberic, "Having fun with Lambert $W(x)$ function," *arXiv:1003.1628*, 2010.

[67] I. Chatzigeorgiou, "Bounds on the lambert function and their application to the outage analysis of user cooperation," *IEEE Communications Letters*, vol. 17, no. 8, pp. 1505–1508, 2013.

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

[69] F. Capitanescu, J. M. Ramos, P. Panciatici, D. Kirschen, A. M. Marcolini, L. Platbrood, and L. Wehenkel, "State-of-the-art, challenges, and future trends in security constrained optimal power flow," *Electric Power Systems Research*, vol. 81, no. 8, pp. 1731–1741, 2011.

[70] B. Stott, J. Jardim, and O. Alsaç, "Dc power flow revisited," *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1290–1300, 2009.

[71] P. Bie, H.-D. Chiang, B. Zhang, and N. Zhou, "Online multiperiod power dispatch with renewable uncertainty and storage: A two-parameter homotopy-enhanced methodology," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6321–6331, 2018.

[72] S. Wong and J. D. Fuller, "Pricing energy and reserves using stochastic optimization in an alternative electricity market," *IEEE Transactions on Power Systems*, vol. 22, no. 2, pp. 631–638, 2007.

[73] R. A. Jabr, S. Karaki, and J. A. Korbane, "Robust multi-period opf with storage and renewables," *IEEE Transactions on Power Systems*, vol. 30, no. 5, pp. 2790–2799, 2014.

[74] A. A. Thatte and L. Xie, "A metric and market construct of inter-temporal flexibility in time-coupled economic dispatch," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3437–3446, 2015.

[75] A. Lorca and X. A. Sun, "Adaptive robust optimization with dynamic uncertainty sets for multi-period economic dispatch under significant wind," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1702–1713, 2014.

[76] H. Pandžić, Y. Dvorkin, T. Qiu, Y. Wang, and D. S. Kirschen, "Toward cost-efficient and reliable unit commitment under uncertainty," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 970–982, 2015.

[77] Y. Cho, T. Ishizaki, N. Ramdani, and J.-i. Imura, "Box-based temporal decomposition of multi-period economic dispatch for two-stage robust unit commitment," *IEEE Transactions on Power Systems*, 2019.

[78] Q. P. Zheng, J. Wang, and A. L. Liu, "Stochastic optimization for unit commitmenta review," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1913–1924, 2014.

[79] S. Cerisola, Á. Baíllo, J. M. Fernández-López, A. Ramos, and R. Gollmer, "Stochastic power generation unit commitment in electricity markets: A novel formulation and a comparison of solution methods," *Operations Research*, vol. 57, no. 1, pp. 32–46, 2009.

[80] L. Wu, M. Shahidehpour, and T. Li, "Stochastic security-constrained unit commitment," *IEEE Transactions on Power Systems*, vol. 22, no. 2, pp. 800–811, 2007.

[81] J. Wang, M. Shahidehpour, and Z. Li, "Security-constrained unit commitment with volatile wind power generation," *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 1319–1327, 2008.

[82] J. Zou, S. Ahmed, and X. A. Sun, "Multistage stochastic unit commitment using stochastic dual dynamic integer programming," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 1814–1823, 2018.

[83] A. Kalantari, J. F. Restrepo, and F. D. Galiana, "Security-constrained unit commitment with uncertain wind generation: The loadability set approach," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1787–1796, 2012.

[84] A. Papavasiliou, S. S. Oren, and R. P. O'Neill, "Reserve requirements for wind power integration: A scenario-based stochastic programming framework," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2197–2206, 2011.

[85] A. Papavasiliou and S. S. Oren, "Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network," *Operations Research*, vol. 61, no. 3, pp. 578–592, 2013.

[86] H. Gangammanavar, S. Sen, and V. M. Zavala, "Stochastic optimization of sub-hourly economic dispatch with wind energy," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 949–959, 2015.

[87] O. Mégel, G. Andersson, and J. L. Mathieu, "Reducing the computational effort of stochastic multi-period dc optimal power flow with storage," in *2016 Power Systems Computation Conference (PSCC)*, pp. 1–7, IEEE, 2016.

[88] B. Hua, D. Schiro, T. Zheng, R. Baldick, and E. Litvinov, "Pricing in multi-interval real-time markets," *IEEE Transactions on Power Systems*, 2019.

[89] J. Zhao, T. Zheng, and E. Litvinov, "A multi-period market design for markets with intertemporal constraints," *arXiv preprint arXiv:1812.07034*, 2018.

[90] E. Ela and M. O'Malley, "Scheduling and pricing for expected ramp capability in real-time power markets," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1681–1691, 2015.

[91] H. Zhang and P. Li, "Chance constrained programming for optimal power flow under uncertainty," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2417–2424, 2011.

[92] F. Bouffard, F. D. Galiana, and A. J. Conejo, "Market-clearing with stochastic security-part i: formulation," *IEEE Transactions on Power Systems*, vol. 20, no. 4, pp. 1818–1826, 2005.

[93] G. Pritchard, G. Zakeri, and A. Philpott, "A single-settlement, energy-only electric power market for unpredictable and intermittent participants," *Operations research*, vol. 58, no. 4-part-2, pp. 1210–1219, 2010.

[94] J. M. Morales, M. Zugno, S. Pineda, and P. Pinson, "Electricity market clearing with improved scheduling of stochastic production," *European Journal of Operational Research*, vol. 235, no. 3, pp. 765–774, 2014.

[95] D. S. Kirschen and G. Strbac, *Fundamentals of power system economics*. John Wiley & Sons, 2004.

[96] G. C. Pflug, "Some remarks on the value-at-risk and the conditional value-at-risk," in *Probabilistic constrained optimization*, pp. 272–281, Springer, 2000.

[97] A. Beck, *First-Order Methods in Optimization*, vol. 25. SIAM, 2017.

[98] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[99] R. M. Van Slyke and R. Wets, "L-shaped linear programs with applications to optimal control and stochastic programming," *SIAM Journal on Applied Mathematics*, vol. 17, no. 4, pp. 638–663, 1969.

[100] J. M. Mulvey and A. Ruszczyński, "A new scenario decomposition method for large-scale stochastic optimization," *Operations research*, vol. 43, no. 3, pp. 477–490, 1995.

[101] J. Eckstein and D. P. Bertsekas, "On the douglasrachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, 1992.

[102] A. Ruszczyński, "Decomposition methods in stochastic programming," *Mathematical programming*, vol. 79, no. 1, pp. 333–353, 1997.

[103] J. E. Kelley, Jr, "The cutting-plane method for solving convex programs," *Journal of the society for Industrial and Applied Mathematics*, vol. 8, no. 4, pp. 703–712, 1960.

[104] J. L. Higle and S. Sen, "Stochastic decomposition: An algorithm for two-stage linear programs with recourse," *Mathematics of operations research*, vol. 16, no. 3, pp. 650–669, 1991.

[105] A. J. King and R. J. Wets*, "Epi-consistency of convex stochastic programs," *Stochastics and Stochastic Reports*, vol. 34, no. 1-2, pp. 83–92, 1991.

[106] S. M. Robinson, "Analysis of sample-path optimization," *Mathematics of Operations Research*, vol. 21, no. 3, pp. 513–528, 1996.

[107] A. J. King and R. T. Rockafellar, "Asymptotic theory for solutions in statistical estimation and stochastic programming," *Mathematics of Operations Research*, vol. 18, no. 1, pp. 148–162, 1993.

[108] A. Shapiro, "Asymptotic properties of statistical estimators in stochastic programming," *The Annals of Statistics*, pp. 841–858, 1989.

[109] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[110] H. J. Kushner and D. S. Clark, "Stochastic approximation methods for constrained and unconstrained systems," 1978.

[111] M. B. Nevel'son and R. Z. Khas' minskii, *Stochastic approximation and recursive estimation*, vol. 47. American Mathematical Society Providence, 1976.

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

[113] A. Nemirovski and D. Yudin, "On cezaris convergence of the steepest descent method for approximating saddle point of convex-concave functions," in *Soviet Math. Dokl*, vol. 19, 1978.

[114] Y. M. Ermoliev, "On the stochastic quasi-gradient method and stochastic quasi-feyer sequences," *Kibernetika*, vol. 2, pp. 72–83, 1969.

[115] A. Gavironski, "Implementation of stochastic quasigradient methods," *Numerical Techniques for Stochastic Optimization. Springer-Verlag*, 1988.

[116] R. Wets, "Stochastic programming: Solution techniques and approximation schemes," in *Mathematical programming The state of the art*, pp. 566–603, Springer, 1983.

[117] U. V. Shanbhag and J. H. Blanchet, "Budget-constrained stochastic approximation," in *Winter Simulation Conference (WSC), 2015*, pp. 368–379, IEEE, 2015.

[118] A. Jalilzadeh and U. V. Shanbhag, "eg-vssa: an extragradient variable sample-size stochastic approximation scheme: error analysis and complexity trade-offs," in *Proceedings of the 2016 Winter Simulation Conference*, pp. 690–701, IEEE Press, 2016.

[119] A. Jalilzadeh, U. V. Shanbhag, J. H. Blanchet, and P. W. Glynn, "On the analysis of unaccelerated and accelerated variable sample-size stochastic approximation (vssa) schemes." Manuscript submitted for publication., 2017.

[120] A. Jalilzadeh, U. V. Shanbhag, J. H. Blanchet, and P. W. Glynn, "Optimal smoothed variable sample-size accelerated proximal methods for structured nonsmooth stochastic convex programs," *arXiv preprint arXiv:1803.00718*, 2018.

[121] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, 2014.

[122] H. Gangammanavar, *Multiple timescale stochastic optimization with application to integrating renewable resources in power systems*. PhD thesis, The Ohio State University, 2013.

[123] B. G. Brown, R. W. Katz, and A. H. Murphy, "Time series models to simulate and forecast wind speed and wind power," *Journal of climate and applied meteorology*, vol. 23, no. 8, pp. 1184–1195, 1984.

[124] A. Garcia, J. Torres, E. Prieto, and A. De Francisco, "Fitting wind speed distributions: a case study," *Solar energy*, vol. 62, no. 2, pp. 139–144, 1998.

[125] W. K. K. Haneveld and M. H. van der Vlerk, "Optimizing electricity distribution using two-stage integer recourse models," in *Stochastic optimization: algorithms and applications*, pp. 137–154, Springer, 2001.

[126] M. A. Dempster, M. Fisher, L. Jansen, B. Lageweg, J. K. Lenstra, and A. Rinnooy Kan, "Analytical evaluation of hierarchical planning systems," *Operations Research*, vol. 29, no. 4, pp. 707–716, 1981.

[127] G. Laporte, F. Louveaux, and H. Mercure, "The vehicle routing problem with stochastic travel times," *Transportation science*, vol. 26, no. 3, pp. 161–170, 1992.

[128] R. Schultz, "Continuity properties of expectation functions in stochastic integer programming," *Mathematics of Operations Research*, vol. 18, no. 3, pp. 578–589, 1993.

[129] R. Schultz, "On structure and stability in stochastic programs with random technology matrix and complete integer recourse," *Mathematical Programming*, vol. 70, no. 1-3, pp. 73–89, 1995.

[130] V. I. Norkin, G. C. Pflug, and A. Ruszczyński, "A branch and bound method for stochastic global optimization," *Mathematical programming*, vol. 83, no. 1-3, pp. 425–450, 1998.

[131] S. Ahmed, A. Shapiro, and E. Shapiro, "The sample average approximation method for stochastic programs with integer recourse," *Submitted for publication*, pp. 1–24, 2002.

[132] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2002.

[133] R. Hemmecke and R. Schultz, "Decomposition of test sets in stochastic integer programming," *Mathematical Programming*, vol. 94, no. 2-3, pp. 323–341, 2003.

[134] R. Schultz, L. Stougie, and M. H. Van Der Vlerk, "Solving stochastic programs with integer recourse by enumeration: A framework using gröbner basis," *Mathematical Programming*, vol. 83, no. 1-3, pp. 229–252, 1998.

[135] S. R. Tayur, R. R. Thomas, and N. Natraj, "An algebraic geometry algorithm for scheduling in presence of setups and correlated demands," *Mathematical Programming*, vol. 69, no. 1-3, pp. 369–401, 1995.

[136] C. C. Carøe, *Decomposition in stochastic integer programming*. Institute of Mathematical Sciences, Department of Operations Research, 1999.

[137] S. Sen and J. L. Higle, "The c 3 theorem and a d 2 algorithm for large scale stochastic mixed-integer programming: Set convexification," *Mathematical Programming*, vol. 104, no. 1, pp. 1–20, 2005.

[138] M. H. Van Der Vlerk, "Convex approximations for complete integer recourse models," *Mathematical Programming*, vol. 99, no. 2, pp. 297–310, 2004.

[139] G. Laporte and F. V. Louveaux, "The integer l-shaped method for stochastic integer programs with complete recourse," *Operations research letters*, vol. 13, no. 3, pp. 133–142, 1993.

[140] H. D. Sherali and B. M. Fraticelli, "A modification of benders' decomposition algorithm for discrete subproblems: An approach for stochastic programs with integer recourse," *Journal of Global Optimization*, vol. 22, no. 1-4, pp. 319–342, 2002.

[141] C. C. Carøe and J. Tind, "L-shaped decomposition of two-stage stochastic programs with integer recourse," *Mathematical Programming*, vol. 83, no. 1-3, pp. 451–464, 1998.

[142] S. Ahmed, M. Tawarmalani, and N. V. Sahinidis, "A finite branch-and-bound algorithm for two-stage stochastic integer programs," *Mathematical Programming*, vol. 100, no. 2, pp. 355–377, 2004.

[143] N. Kong, A. J. Schaefer, and B. Hunsaker, "Two-stage integer programs with stochastic right-hand sides: a superadditive dual approach," *Mathematical Programming*, vol. 108, no. 2-3, pp. 275–296, 2006.

[144] C. C. CarøE and R. Schultz, "Dual decomposition in stochastic integer programming," *Operations Research Letters*, vol. 24, no. 1-2, pp. 37–45, 1999.

[145] Y. Guan, S. Ahmed, and G. L. Nemhauser, "Cutting planes for multistage stochastic integer programs," *Operations research*, vol. 57, no. 2, pp. 287–298, 2009.

[146] L. Ntaimo and M. W. Tanner, "Computations with disjunctive cuts for two-stage stochastic mixed 0-1 integer programs," *Journal of Global Optimization*, vol. 41, no. 3, pp. 365–384, 2008.

[147] H. D. Sherali and X. Zhu, "On solving discrete two-stage stochastic programs having mixed-integer first-and second-stage variables," *Mathematical programming*, vol. 108, no. 2-3, pp. 597–616, 2006.

[148] S. Sen and H. D. Sherali, "Decomposition with branch-and-cut approaches for two-stage stochastic mixed-integer programming," *Mathematical Programming*, vol. 106, no. 2, pp. 203–223, 2006.

[149] A. Ruszczyński, "A regularized decomposition method for minimizing a sum of polyhedral functions," *Mathematical programming*, vol. 35, no. 3, pp. 309–333, 1986.

[150] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, "New variants of bundle methods," *Mathematical programming*, vol. 69, no. 1-3, pp. 111–147, 1995.

[151] J. Linderoth and S. Wright, "Decomposition algorithms for stochastic programming on a computational grid," *Computational Optimization and Applications*, vol. 24, no. 2-3, pp. 207–250, 2003.

[152] V. Zverovich, C. I. Fábián, E. F. Ellison, and G. Mitra, "A computational study of a solver system for processing two-stage stochastic lps with enhanced benders decomposition," *Mathematical Programming Computation*, vol. 4, no. 3, pp. 211–238, 2012.

[153] M. Lubin, K. Martin, C. G. Petra, and B. Sandıkçı, "On parallelizing dual decomposition in stochastic integer programming," *Operations Research Letters*, vol. 41, no. 3, pp. 252–258, 2013.

[154] R. T. Rockafellar and R. J.-B. Wets, "Scenarios and policy aggregation in optimization under uncertainty," *Mathematics of operations research*, vol. 16, no. 1, pp. 119–147, 1991.

[155] J.-P. Watson, R. J. Wets, and D. L. Woodruff, "Scalable heuristics for a class of chance-constrained stochastic programs," *INFORMS Journal on Computing*, vol. 22, no. 4, pp. 543–554, 2010.

[156] D. Gade, G. Hackebeil, S. M. Ryan, J.-P. Watson, R. J.-B. Wets, and D. L. Woodruff, "Obtaining lower bounds from the progressive hedging algorithm for stochastic mixed-integer programs," *Mathematical Programming*, vol. 157, no. 1, pp. 47–67, 2016.

[157] A. Märkert and R. Gollmer, "User?s guide to ddsip–ac package for the dual decomposition of two-stage stochastic programs with mixed-integer recourse," *Department of Mathematics, University of Duisburg-Essen, Duisburg*, 2008.

[158] R.-B. Wets, "Large scale linear programming techniques in stochastic programming," *IIASA Working Paper*, 1984.

[159] L. F. Escudero, M. A. Garín, G. Pérez, and A. Unzueta, "Scenario cluster decomposition of the lagrangian dual in two-stage stochastic mixed 0–1 optimization," *Computers & Operations Research*, vol. 40, no. 1, pp. 362–377, 2013.

[160] T. G. Crainic, M. Hewitt, and W. Rei, "Scenario grouping in a progressive hedging-based meta-heuristic for stochastic network design," *Computers & Operations Research*, vol. 43, pp. 90–99, 2014.

[161] C. Li and I. E. Grossmann, "An improved l-shaped method for two-stage convex 0–1 mixed integer nonlinear stochastic programs," *Computers & Chemical Engineering*, vol. 112, pp. 165–179, 2018.

[162] C. Li and I. E. Grossmann, "A finite ϵ-convergence algorithm for two-stage stochastic convex nonlinear programs with mixed-binary first and second-stage variables," *Journal of Global Optimization*, vol. 75, no. 4, pp. 921–947, 2019.

[163] V. I. Norkin, Y. M. Ermoliev, and A. Ruszczyński, "On optimal allocation of indivisibles under uncertainty," *Operations Research*, vol. 46, no. 3, pp. 381–395, 1998.

[164] A. Shapiro, D. Dentcheva, and A. Ruszcsynski, *Lectures on stochastic programming: modeling and theory*, vol. 16. SIAM, 2014.

[165] S. Cui, U. V. Shanbhag, and F. Yousefian, "Complexity guarantees for an implicit smoothing-enabled method for stochastic mpecs," 2021.

[166] B. F. Hobbs, C. B. Metzler, and J.-S. Pang, "Strategic gaming analysis for electric power systems: An MPEC approach," *IEEE Transactions on Power Systems*, vol. 15, pp. 638–645, 2000.

[167] S. Lawphongpanich and D. W. Hearn, "An MPEC approach to second-best toll pricing," *Math. Program.*, vol. 101, no. 1, Ser. B, pp. 33–55, 2004.

[168] H. D. Sherali, A. L. Soyster, and F. H. Murphy, "Stackelberg-Nash-Cournot equilibria: characterizations and computations," *Oper. Res.*, vol. 31, no. 2, pp. 253–276, 1983.

[169] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*. Cambridge: Cambridge University Press, 1996.

[170] J. Outrata, M. Kočvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, vol. 28 of *Nonconvex Optimization and its Applications*. Dordrecht: Kluwer Academic Publishers, 1998. Theory, applications and numerical results.

[171] H. Scheel and S. Scholtes, "Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity," *Math. Oper. Res.*, vol. 25, no. 1, pp. 1–22, 2000.

[172] A. U. Raghunathan and L. T. Biegler, "An interior point method for mathematical programs with complementarity constraints (MPCCs)," *SIAM J. Optim.*, vol. 15, no. 3, pp. 720–750 (electronic), 2005.

[173] R. Fletcher, S. Leyffer, D. Ralph, and S. Scholtes, "Local convergence of sqp methods for mathematical programs with equilibrium constraints," *SIAM Journal on Optimization*, vol. 17, no. 1, pp. 259–286, 2006.

[174] M. Anitescu, "On solving mathematical programs with complementarity constraints as nonlinear programs," *SIAM J. Optim.*, vol. 15(4), pp. 1203–1236, 2005.

[175] H. Jiang and D. Ralph, "Smooth SQP methods for mathematical programs with nonlinear complementarity constraints," *SIAM Journal on Optimization*, vol. 10(3), pp. 779–808, 2000.

[176] S. Leyffer, G. López-Calva, and J. Nocedal, "Interior methods for mathematical programs with complementarity constraints," *SIAM Journal on Optimization*, vol. 17, no. 1, pp. 52–77, 2006.

[177] X. Hu and D. Ralph, "Convergence of a penalty method for mathematical programming with complementarity constraints," *Journal of Optimization Theory and Applications*, vol. 123, no. 2, pp. 365–398, 2004.

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

[179] A. Juditsky, A. Nemirovski, and C. Tauvel, "Solving variational inequalities with stochastic mirror-prox algorithm," *Stochastic Systems*, vol. 1, no. 1, pp. 17–58, 2011.

[180] F. Yousefian, A. Nedic, and U. V. Shanbhag, "On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems," *Math. Program.*, vol. 165, no. 1, pp. 391–431, 2017.

[181] V. DeMiguel and H. Xu, "A stochastic multiple-leader stackelberg model: Analysis, computation, and application," *Operations Research*, vol. 57, no. 5, pp. 1220–1235, 2009.

[182] M. Patriksson and L. Wynter, "Stochastic mathematical programs with equilibrium constraints," *Operations Research Letters*, vol. 25, pp. 159–167, 1999.

[183] H. Xu, "An implicit programming approach for a class of stochastic mathematical programs with complementarity constraints," *SIAM J. Optim.*, vol. 16, no. 3, pp. 670–696, 2006.

[184] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth analysis and control theory*, vol. 178 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.

[185] A. A. Goldstein, "Optimization of Lipschitz continuous functions," *Math. Programming*, vol. 13, no. 1, pp. 14–22, 1977.

[186] J. V. Burke, A. S. Lewis, and M. L. Overton, "A robust gradient sampling algorithm for nonsmooth, nonconvex optimization," *SIAM J. Optim.*, vol. 15, no. 3, pp. 751–779, 2005.

[187] X. Chen, "Smoothing methods for nonsmooth, nonconvex minimization," *Math. Program.*, vol. 134, no. 1, Ser. B, pp. 71–99, 2012.

[188] V. A. Steklov, "Sur les expressions asymptotiques decertaines fonctions dfinies par les quations diffrentielles du second ordre et leers applications au problme du dvelopement d'une fonction arbitraire en sries procdant suivant les diverses fonctions," *Comm. Charkov Math. Soc.*, vol. 2, no. 10, pp. 97–199, 1907.

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

[190] F. Yousefian, A. Nedić, and U. V. Shanbhag, "On stochastic gradient and subgradient methods with adaptive steplength sequences," *Automatica*, vol. 48, no. 1, pp. 56–67, 2012.

[191] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright, "Randomized smoothing for stochastic optimization," *SIAM Journal on Optimization (SIOPT)*, vol. 22, no. 2, pp. 674–701, 2012.

[192] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Found. Comput. Math.*, vol. 17, no. 2, pp. 527–566, 2017.

[193] A. S. Nemirovskij and D. B. Yudin, "Problem complexity and method efficiency in optimization," 1983.

[194] A. Flaxman, A. T. Kalai, and B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *SODA '05 Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394, January 2005.

[195] F. Yousefian, A. Nedi, and U. V. Shanbhag, "Convex nondifferentiable stochastic optimization: A local randomized smoothing technique," in *Proceedings of the 2010 American Control Conference*, pp. 4875–4880, 2010.

[196] D. Q. Mayne and E. Polak, "Nondifferential optimization via adaptive smoothing," *J. Optim. Theory Appl.*, vol. 43, no. 4, pp. 601–613, 1984.

[197] S. Ghadimi and G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2341–2368, 2013.

[198] H. D. Kaushik and F. Yousefian, "A method with convergence rates for optimization problems with variational inequality constraints," *arXiv:2007.15845v2*, 2021.

[199] K. Knopp, *Theory and applications of infinite series*. Bishopbriggs, Glasgow G64 2NZ, Scotland: Blackie & Son Ltd., 1951.

[200] A. Beck, *Introduction to nonlinear optimization*, vol. 19 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2014. Theory, algorithms, and applications with MATLAB.

[201] M. C. Ferris, S. P. Dirkse, and A. Meeraus, "Mathematical programs with equilibrium constraints: Automatic reformulation and solution via constrained optimization," 2002.

[202] J. F. Bard, "Convex two-level optimization," *Math. Programming*, vol. 40, no. 1, (Ser. A), pp. 15–27, 1988.

[203] M. Tawarmalani and N. V. Sahinidis, "A polyhedral branch-and-cut approach to global optimization," *Mathematical Programming*, vol. 103, pp. 225–249, 2005.

[204] N. V. Sahinidis, *BARON 21.1.13: Global Optimization of Mixed-Integer Nonlinear Programs,* User's Manual, 2017.

[205] J. Czyzyk, M. P. Mesnier, and J. J. Moré

[206] E. D. Dolan, "The neos server 4.0 administrative guide," Technical Memorandum ANL/MCS-TM-250, Mathematics and Computer Science Division, Argonne National Laboratory, 2001.

[207] W. Gropp and J. J. Moré, "Optimization environments and the neos server," in *Approximation Theory and Optimization* (M. D. Buhman and A. Iserles, eds.), p. 167, Cambridge University Press, 1997.

[208] A. A. Kulkarni and U. V. Shanbhag, "Recourse-based stochastic nonlinear programming: Properties and benders-sqp algorithms," *Comput. Optim. Appl.*, vol. 51, p. 77123, jan 2012.