



A Stochastic Primal-Dual Method for Optimization with Conditional Value at Risk Constraints

Avinash N. Madavan¹  · Subhonimesh Bose¹

Received: 23 September 2020 / Accepted: 9 June 2021 / Published online: 24 June 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

We study a first-order primal-dual subgradient method to optimize risk-constrained risk-penalized optimization problems, where risk is modeled via the popular conditional value at risk (CVaR) measure. The algorithm processes independent and identically distributed samples from the underlying uncertainty in an online fashion and produces an η/\sqrt{K} -approximately feasible and η/\sqrt{K} -approximately optimal point within K iterations with constant step-size, where η increases with tunable risk-parameters of CVaR. We find optimized step sizes using our bounds and precisely characterize the computational cost of risk aversion as revealed by the growth in η . Our proposed algorithm makes a simple modification to a typical primal-dual stochastic subgradient algorithm. With this mild change, our analysis surprisingly obviates the need to impose a priori bounds or complex adaptive bounding schemes for dual variables to execute the algorithm as assumed in many prior works. We also draw interesting parallels in sample complexity with that for chance-constrained programs derived in the literature with a very different solution architecture.

Keywords Primal-dual optimization · Stochastic optimization · Risk-sensitive optimization · Conditional value at risk

Mathematics Subject Classification 90C15 · 90C25 · 90C30

Communicated by Xiaolu Tan.

✉ Avinash N. Madavan
madavan2@illinois.edu

Subhonimesh Bose
booses@illinois.edu

¹ University of Illinois at Urbana-Champaign, Urbana, IL, USA

1 Introduction

We study iterative primal-dual stochastic subgradient algorithms to solve risk-sensitive optimization problems of the form

$$\begin{aligned} \mathcal{P}^{\text{CVaR}} : \quad & \underset{\mathbf{x} \in \mathbb{X}}{\text{minimize}} \quad F(\mathbf{x}) := \text{CVaR}_\alpha[f_\omega(\mathbf{x})], \\ & \text{subject to} \quad G^i(\mathbf{x}) := \text{CVaR}_{\beta^i}[g_\omega^i(\mathbf{x})] \leq 0, \quad i = 1, \dots, m, \end{aligned} \quad (1)$$

where $\omega \in \Omega$ is random and $\alpha, \beta := (\beta^1, \dots, \beta^m)$ in $[0, 1]$ define risk-aversion parameters. The collection of real-valued functions $f_\omega, g_\omega^1, \dots, g_\omega^m$ are assumed convex but not necessarily differentiable, over the closed convex set $\mathbb{X} \subseteq \mathbb{R}^n$, where \mathbb{R} and \mathbb{R}_+ stand for the set of real and nonnegative numbers, respectively. Denote by \mathbf{G} and \mathbf{g}_ω , the collection of G^i 's and g_ω^i 's, respectively, for $i = 1, \dots, m$. CVaR stands for conditional value at risk. For any $\delta \in [0, 1]$, $\text{CVaR}_\delta[y_\omega]$ of a scalar random variable y_ω with continuous distribution equals its expectation computed over the $1 - \delta$ tail of the distribution of y_ω . For y_ω with general distributions, CVaR is defined via the following variational characterization

$$\text{CVaR}_\delta[y_\omega] = \min_{u \in \mathbb{R}} \left\{ u + \frac{1}{1 - \delta} \mathbb{E}[y_\omega - u]^+ \right\}, \quad (2)$$

following [36]. For each $\mathbf{x} \in \mathbb{X}$, assume that $\mathbb{E}[|f_\omega(\mathbf{x})|]$ and $\mathbb{E}[|g_\omega^i(\mathbf{x})|]$ are finite, implying that F and \mathbf{G} are well defined everywhere in \mathbb{X} .

$\mathcal{P}^{\text{CVaR}}$ offers a modeler the flexibility to indicate her risk preference in α, β . With α close to zero, she indicates risk-neutrality toward the uncertain cost associated with the decision. With α closer to one, she expresses her risk aversion toward the same and seeks a decision that limits the possibility of large random costs associated with the decision. Similarly, β 's express the risk tolerance in constraint violation. Choosing β 's close to zero indicates that constraints should be satisfied on average over Ω rather than on each sample. Driving β 's to unity amounts to requiring the constraints to be met almost surely. Said succinctly, $\mathcal{P}^{\text{CVaR}}$ permits the modeler to customize risk preference between the risk-neutral choice of expected evaluations of functions to the conservative choice of robust evaluations.

There is a growing interest in solving risk-sensitive optimization problems with data. See [3, 20] for recent examples that tackle problems with generalized mean semi-deviation risk that equals $\mathbb{E}[y_\omega] + c\mathbb{E}[|y_\omega - \mathbb{E}[y_\omega]|^p]^{1/p}$ for $p > 1$ for a random variable y_ω . There is a long literature on risk measures, e.g., see [1, 12, 25, 33, 36, 37]. We choose CVaR for three particular reasons. First, it is a coherent risk measure, meaning that it is normalized, sub-additive, positively homogeneous and translation invariant, i.e.,

$$\begin{aligned} \text{CVaR}_\delta[0] &= 0, \quad \text{CVaR}_\delta[y_\omega^1 + y_\omega^2] \leq \text{CVaR}_\delta[y_\omega^1] + \text{CVaR}_\delta[y_\omega^2], \\ \text{CVaR}_\delta[ty_\omega] &= t\text{CVaR}_\delta[y_\omega], \quad \text{CVaR}_\delta[y_\omega + t'] = \text{CVaR}_\delta[y_\omega] + t' \end{aligned}$$

for random variables $y_\omega, y_\omega^1, y_\omega^2, t > 0$ and $t' \in \mathbb{R}$. An important consequence of coherence is that F and \mathbf{G} in $\mathcal{P}^{\text{CVaR}}$ inherit the convexity of f_ω and \mathbf{g}_ω . Convexity together with the variational characterization in (2) allow us to design sampling based primal-dual methods for $\mathcal{P}^{\text{CVaR}}$ for which we are able to provide finite sample analysis of approximate optimality and feasibility. The popularity of the CVaR measure is our second reason to study $\mathcal{P}^{\text{CVaR}}$. Following Rockafellar and Uryasev's seminal work in [36], CVaR has found applications in various engineering domains, e.g., see [22, 27], and therefore we anticipate wide applications of our result. Our third and final reason to study $\mathcal{P}^{\text{CVaR}}$ is its close relation to other optimization paradigms in the literature as we describe next.

$\mathcal{P}^{\text{CVaR}}$ without constraints and $\alpha = 0$ reduces to the minimization of $\mathbb{E}[f_\omega(\mathbf{x})]$, the canonical stochastic optimization problem. With $\alpha \uparrow 1$, the problem description of $\mathcal{P}^{\text{CVaR}}$ approaches that of a robust optimization problem (see [4]) of the form $\min_{\mathbf{x} \in \mathbb{X}} \text{ess sup}_{\omega \in \Omega} f_\omega(\mathbf{x})$, where ess sup denotes the essential supremum. Driving β 's to unity, $\mathcal{P}^{\text{CVaR}}$ demands the constraints to be enforced almost surely. Such robust constraint enforcement is common in multi-stage stochastic optimization problems with recourse and discrete-time optimal control problems, e.g., in [16, 39, 40]. CVaR-based constraints are closely related to chance constraints introduced by Charnes and Cooper in [12] that enforce $\Pr\{g_\omega(\mathbf{x}) \leq 0\} > 1 - \varepsilon$ where \Pr refers to the probability measure on Ω . Even if g_ω is convex, chance-constraints typically describe a nonconvex feasible set. It is well known that CVaR-based constraints provide a convex inner approximation of chance-constraints. Restricting the probability of constraint violation does not limit the extent of any possible violation, while CVaR-based enforcement does so in expectation. CVaR is also intimately related to the buffered probability of exceedence (bPOE) introduced and studied more recently in [25, 45]. In fact, bPOE is the inverse function of CVaR, and hence, problems with bPOE-constraints can often be reformulated as instances of $\mathcal{P}^{\text{CVaR}}$.

It can be challenging to compute CVaR of $f_\omega(\mathbf{x})$ or $\mathbf{g}_\omega(\mathbf{x})$ for a given decision variable \mathbf{x} with respect to a general distribution on Ω for two reasons. First, if samples from Ω are obtained from a simulation tool, an explicit representation of the probability distribution on Ω may not be available. Second, even if such a distribution is available, computation of CVaR (or even the expectation) can be difficult. For example, with f_ω as the positive part of an affine function and ω being uniformly distributed over a unit hypercube, computation of $\mathbb{E}[f_\omega]$ via a multivariate integral is #P-hard according to [17, Corollary 1]. Therefore, we do not assume knowledge of F and \mathbf{G} but rather study a sampling-based algorithm to solve $\mathcal{P}^{\text{CVaR}}$.

Solution architectures for $\mathcal{P}^{\text{CVaR}}$ via sampling come in two flavors. The first approach is sample average approximation (SAA) that replaces the expectation in (2) by an empirical average over N samples. One can then solve the sampled problem as a deterministic convex program.¹ We take the second and alternate approach of stochastic approximation and process independent and identically distributed (i.i.d.) samples from Ω in an online fashion. Iterative stochastic approximation algorithms for the unconstrained problem have been studied since the early works by Robbins

¹ For the unconstrained problem, variance-reduced stochastic gradient descent methods can efficiently minimize the resulting finite sum as in [19, 38].

and Monro in [34] and by Kiefer and Wolfowitz in [21]. See [24] for a more recent survey. Zinkevich in [46] proposed a projected stochastic subgradient method that can be applied to tackle constraints in such problems. Without directly knowing \mathbf{G} , we cannot easily project the iterates on the feasible set $\{\mathbf{x} \in \mathbb{X} \mid \mathbf{G}(\mathbf{x}) \leq 0\}$. We circumvent the challenge by associating Lagrange multipliers $\mathbf{z} \in \mathbb{R}_+^m$ to the constraints and iteratively updating \mathbf{x}, \mathbf{z} by using $f_\omega, \mathbf{g}_\omega$ and their subgradients via a first-order stochastic primal-dual algorithm for $\mathcal{P}^{\text{CVaR}}$ along the lines of [30, 42, 44].

In Sect. 2, we first design and analyze Algorithm 1 for $\mathcal{P}^{\text{CVaR}}$ with $\alpha = 0, \boldsymbol{\beta} = 0$, i.e., the optimization problem

$$\begin{aligned} \mathcal{P}^E : \quad & \underset{\mathbf{x} \in \mathbb{X}}{\text{minimize}} \quad F(\mathbf{x}) := \mathbb{E}[f_\omega(\mathbf{x})], \\ & \text{subject to} \quad G^i(\mathbf{x}) := \mathbb{E}[g_\omega^i(\mathbf{x})] \leq 0, \quad i = 1, \dots, m. \end{aligned} \quad (3)$$

First-order stochastic primal-dual algorithms have a long history, dating back almost forty years, including that in [15, 24, 30–32]. The analyses of these algorithms often require a bound on the possible growth of the dual variables. Borkar and Meyn in [8] stress the importance of compactness assumptions in their analysis of stochastic approximation algorithms. A priori bounds used in [31] are difficult to know in practice and techniques for iterative construction of such bounds as in [30] require extra computational effort. A regularization term in the dual update has been proposed in [23, 26] to circumvent this limitation. Instead, we propose a different modification to the classical primal-dual stochastic subgradient algorithm. With this simple modification, we are able to circumvent the need to bound the dual variables in executing the algorithm. As will become clear in Sect. 2, we rely on the existence of a saddle point of the Lagrangian function for \mathcal{P}^E , which is typically guaranteed under Slater-type constraint qualification. However, knowledge of that saddle point or a strictly feasible “Slater” point is not required to execute the algorithm nor derive its convergence rate. While the classical primal-dual approach samples once for a single update of the primal and the dual variables, we sample twice—once to update the primal variable and then again to update the dual variable with the most recent primal iterate—thus, adopting a Gauss–Seidel approach in place of a Jacobi framework. For Algorithm 1, we bound the expected optimality gap and constraint violations at a suitably weighted average of the iterates by η/\sqrt{K} for a constant η with a constant step-size algorithm. Using these bounds, we then carefully optimize the step-size that allows us to reach within a given threshold of suboptimality and constraint violation with the minimum number of iterations. While we do not bound the dual variables to execute the algorithm or to characterize the $1/\sqrt{K}$ convergence rate, we do require an overestimate of the distance of the dual initialization from an optimal point to calculate the constant η that in turn is required to optimize the constant step-size. The additional sample required in our update aids in the analysis; however, it comes at the price of making the sample complexity double of the iteration complexity. Given the popularity of decaying step-sizes in first-order algorithms, we also provide stability analysis of our algorithm with such step-sizes. This analysis exploits a dissipation inequality that we derive for our Gauss–Seidel approach. Such a stability analysis is crucial for our

primal-dual algorithm, given that we do not explicitly restrict the growth of the dual variables.

In Sect. 3, we solve $\mathcal{P}^{\text{CVaR}}$ with general risk aversion parameters α, β using Algorithm 1 on an instance of \mathcal{P}^E obtained through a standard reformulation via the variational formula for CVaR in (2) from [36]. We then bound the expected suboptimality and constraint violation at a weighted average of the iterates for $\mathcal{P}^{\text{CVaR}}$ by $\eta(\alpha, \beta)/\sqrt{K}$. Upon utilizing the optimized step-sizes from the analysis of \mathcal{P}^E , we are then able to study the precise growth in the required iteration (and sample) complexity of $\mathcal{P}^{\text{CVaR}}$ as a function of α, β . Not surprisingly, the more risk-averse a problem one aims to solve, the greater this complexity increases. A modeler chooses risk aversion parameters primarily driven by attitudes toward risk in specific applications. Our precise characterization of the growth in sample complexity with risk aversion will permit the modeler to balance between desired risk levels and computational challenges in handling that risk. We remark that the algorithmic architecture for the risk neutral problem may not directly apply to the risk-sensitive variant for general risk measures. For example, the algorithm described in [20] for general mean-semideviation-type risk measures is considerably more complex than that required for the risk-neutral problem. We are able to extend our algorithm and its analysis for \mathcal{P}^E to $\mathcal{P}^{\text{CVaR}}$, thanks to the variational form in (2) that CVaR admits. See the discussion after the proof of Theorem 3.1 for a precise list of properties a risk measure must exhibit for us to apply the same trick. Using concentration inequalities, we also report an interesting connection of our results to that in [10,11] on scenario approximations to chance-constrained programs. The resemblance in sample complexity is surprising, given that the approach in [10,11] solves a deterministic convex program with sampled constraints, while we process samples in an online fashion.

We illustrate properties of our algorithm through a stylized example. Our experiments reveal that the optimized iteration count (and sample complexity) for even a simple example is quite high. This limitation is unfortunately common for subgradient algorithms and likely cannot be overcome in optimizing general nonsmooth functions that we study. While the bounds are order-optimal, our numerical experiments reveal that a solution with desired risk tolerance can be found in less iterations than obtained from the upper bound. This is an artifact of optimizing step-sizes based on upper bounds on suboptimality and constraint violation. We end the paper in Sect. 4 with discussions on possible extensions of our analysis.

Very recently, it was brought to our attention that the work in [7] done concurrently presents a related approach to tackle optimization of composite nonconvex functions under related but different assumptions. In fact, their work appeared at the same time as our early version and claims a similar result that does not require bounds on the dual variables. Our analysis does not require or analyze the case with strongly convex functions within our setup and therefore Nesterov-style acceleration remains untenable. As a result, our algorithm is different. Our focus on CVaR permits us to further analyze the growth in optimized sample complexity with risk aversion and its connection to chance-constrained optimization that is quite different.

2 Algorithm for \mathcal{P}^E and Its Analysis

We present the primal-dual stochastic subgradient method to solve \mathcal{P}^E in Algorithm 1.

Algorithm 1: Primal-dual stochastic subgradient method for \mathcal{P}^E .

Initialization: Choose $\mathbf{x}_1 \in \mathbb{X}$, $\mathbf{z}_1 = 0$, and a positive sequence γ .

1 **for** $k \geq 1$ **do**

2 Sample $\omega_k \in \Omega$. Update \mathbf{x} as

$$\mathbf{x}_{k+1} \leftarrow \operatorname{argmin}_{\mathbf{x} \in \mathbb{X}} \left\langle \nabla f_{\omega_k}(\mathbf{x}_k) + \sum_{i=1}^m z_k^i \nabla g_{\omega_k}^i(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \right\rangle + \frac{1}{2\gamma_k} \|\mathbf{x} - \mathbf{x}_k\|^2. \quad (4)$$

3 Sample $\omega_{k+1/2} \in \Omega$. Update \mathbf{z} as

$$z_{k+1} \leftarrow \operatorname{argmax}_{z \in \mathbb{R}_+^m} \left\langle \mathbf{g}_{\omega_{k+1/2}}(\mathbf{x}_{k+1}), z - z_k \right\rangle - \frac{1}{2\gamma_k} \|z - z_k\|^2. \quad (5)$$

The notation $\langle \cdot, \cdot \rangle$ stands for the usual inner product in Euclidean space and $\|\cdot\|$ denotes the induced ℓ_2 -norm. Here, $\nabla h(\mathbf{x})$ stands for a subgradient of an arbitrary convex function h at \mathbf{x} . For our analysis, the subgradient in Algorithm 1 for functions $f_\omega(\mathbf{x})$ and $g_\omega(\mathbf{x})$ can be arbitrary elements of the closed convex subdifferential sets $\partial f_\omega(\mathbf{x})$ and $\partial g_\omega(\mathbf{x})$, respectively. We assume that these subdifferential sets are nonempty everywhere in \mathbb{X} .

The primal-dual method in Algorithm 1 leverages Lagrangian duality theory. Specifically, define the Lagrangian function for \mathcal{P}^E as

$$\mathcal{L}(\mathbf{x}, \mathbf{z}) := F(\mathbf{x}) + \mathbf{z}^\top \mathbf{G}(\mathbf{x}) = \mathbb{E}[\mathcal{L}_\omega(\mathbf{x}, \mathbf{z})], \quad (6)$$

for $\mathbf{x} \in \mathbb{X}$, $\mathbf{z} \in \mathbb{R}_+^m$, where $\mathcal{L}_\omega(\mathbf{x}, \mathbf{z}) := f_\omega(\mathbf{x}) + \mathbf{z}^\top \mathbf{g}_\omega(\mathbf{x})$. Then, \mathcal{P}^E admits the standard reformulation as a min-max problem of the form

$$p_\star^E := \min_{\mathbf{x} \in \mathbb{X}} \max_{\mathbf{z} \in \mathbb{R}_+^m} \mathcal{L}(\mathbf{x}, \mathbf{z}). \quad (7)$$

Denote its optimal set by $\mathbb{X}_\star \subseteq \mathbb{X}$. Define the dual problem of \mathcal{P}^E as

$$d_\star^E := \max_{\mathbf{z} \in \mathbb{R}_+^m} \min_{\mathbf{x} \in \mathbb{X}} \mathcal{L}(\mathbf{x}, \mathbf{z}). \quad (8)$$

Denote its optimal set by $\mathbb{Z}_\star \subseteq \mathbb{R}_+^m$. Weak duality then guarantees $p_\star^E \geq d_\star^E$. When the inequality is met with an equality, the problem is said to satisfy strong duality. A point $(\mathbf{x}_\star, \mathbf{z}_\star) \in \mathbb{X} \times \mathbb{R}_+^m$ is a saddle point of \mathcal{L} if

$$\mathcal{L}(\mathbf{x}_\star, \mathbf{z}) \leq \mathcal{L}(\mathbf{x}_\star, \mathbf{z}_\star) \leq \mathcal{L}(\mathbf{x}, \mathbf{z}_\star) \quad (9)$$

for all $(\mathbf{x}, z) \in \mathbb{X} \times \mathbb{R}_+^m$. The following well-known saddle point theorem (see [6, Theorem 2.156]) relates saddle points with primal-dual optimal solutions.

Theorem (Saddle point theorem) *A saddle point of \mathcal{L} exists if and only if \mathcal{P}^E satisfies strong duality, i.e., $p_\star^E = d_\star^E$. Moreover, the set of saddle points of \mathcal{L} is given by $\mathbb{X}_\star \times \mathbb{Z}_\star$.*

Our convergence analysis of Algorithm 1 requires the following assumptions.

Assumption 1 \mathcal{P}^E must satisfy the following properties:

- (a) Subgradients of F and \mathbf{G} are bounded, i.e., $\|\nabla F(\mathbf{x})\| \leq C_F$, $\|\nabla G^i(\mathbf{x})\| \leq C_G^i$ for each $i = 1, \dots, m$ and all $\mathbf{x} \in \mathbb{X}$.
- (b) ∇f_ω and ∇g_ω^i for $i = 1, \dots, m$ have bounded variance, i.e., $\mathbb{E}\|\nabla f_\omega(\mathbf{x}) - \mathbb{E}[\nabla f_\omega(\mathbf{x})]\|^2 \leq \sigma_F^2$ and $\mathbb{E}\|\nabla g_\omega^i(\mathbf{x}) - \mathbb{E}[\nabla g_\omega^i(\mathbf{x})]\|^2 \leq [\sigma_G^i]^2$ for all $\mathbf{x} \in \mathbb{X}$.
- (c) $\mathbf{g}_\omega(\mathbf{x})$ has a bounded second moment, i.e., $\mathbb{E}\|\mathbf{g}_\omega(\mathbf{x})\|^2 \leq [D_G^i]^2$ for all $\mathbf{x} \in \mathbb{X}$.
- (d) The Lagrangian function \mathcal{L} admits a saddle point $(\mathbf{x}_\star, z_\star) \in \mathbb{X} \times \mathbb{R}_+^m$.

The subgradient of F and the variance of its noisy estimate are assumed bounded. Such an assumption is standard in the convergence analysis of unconstrained stochastic subgradient methods. The assumptions regarding \mathbf{G} are similar, but we additionally require the second moment of the noisy estimate of \mathbf{G} to be bounded over \mathbb{X} . Boundedness of \mathbf{G} in primal-dual subgradient methods has appeared in prior literature, e.g., in [42,44]. The second moment remains bounded if g_ω^i is uniformly bounded over \mathbb{X} and Ω for each i . It is also satisfied if \mathbf{G} remains bounded over \mathbb{X} and its noisy estimate has a bounded variance. Convergence analysis of unconstrained optimization problems typically assumes the existence of a finite optimal solution. We extend that requirement to the existence of a saddle point in the primal-dual setting, which by the saddle point theorem is equivalent to the existence of finite primal and dual optimal solutions. A variety of conditions imply the existence of such a point; the next result delineates two such sufficient conditions in (a) and (b), where (a) implies (b).

Lemma 2.1 (Sufficient conditions for existence of a saddle point) *For \mathcal{P}^E , the Lagrangian function \mathcal{L} admits a saddle point, if either of the following conditions hold:*

- (a) \mathbb{X}_\star is nonempty, p_\star^E is finite and Slater's constraint qualification holds, i.e., there exists \mathbf{x} in the relative interior of \mathbb{X} for which $\mathbf{G}(\mathbf{x}) < 0$.
- (b) \mathcal{P}^E admits a finite $(\mathbf{x}_\star, z_\star) \in \mathbb{X} \times \mathbb{R}_+^m$ that satisfies the generalized Karush–Kuhn–Tucker (KKT) conditions given by

$$0 \in \partial_{\mathbf{x}} \mathcal{L}(\mathbf{x}_\star, z_\star) + \mathcal{N}_{\mathbb{X}}(\mathbf{x}_\star), \quad G^i(\mathbf{x}_\star) \leq 0, \quad z_\star^i G^i(\mathbf{x}_\star) = 0 \quad (10)$$

for $i = 1, \dots, m$, where $\mathcal{N}_{\mathbb{X}}(\mathbf{x}_\star)$ denotes the normal cone of \mathbb{X} at \mathbf{x}_\star .

Proof Part (a) is a direct consequence of [6, Theorem 1.265]. To prove part (b), notice that (10) ensures the existence of subgradients $\nabla F(\mathbf{x}_\star) \in \partial F(\mathbf{x}_\star)$, $\nabla G^i(\mathbf{x}_\star) \in \partial G^i(\mathbf{x}_\star)$, $i = 1, \dots, m$ and $\mathbf{n} \in \mathcal{N}_{\mathbb{X}}(\mathbf{x}_\star)$ for which

$$\nabla F(\mathbf{x}_\star) + \sum_{i=1}^m z_\star^i \nabla G^i(\mathbf{x}_\star) + \mathbf{n} = 0. \quad (11)$$

Then, for any $x \in \mathbb{X}$, we have

$$\underbrace{\langle \nabla F(x_*) - F(x_*) \rangle}_{\leq F(x) - F(x_*)} + \sum_{i=1}^m \underbrace{z_*^i \langle \nabla G^i(x_*), x - x_* \rangle}_{\leq z_*^i [G^i(x) - G^i(x_*)]} + \underbrace{\langle n, x - x_* \rangle}_{\leq 0} = 0. \quad (12)$$

The inequalities in the above relation follow from the convexity of F and G^i 's, non-negativity of z_* , and the definition of the normal cone. From the above inequalities, we conclude $\mathcal{L}(x, z_*) \geq \mathcal{L}(x_*, z_*)$ for all $x \in \mathbb{X}$. Furthermore, for any $z \geq 0$, we have

$$\mathcal{L}(x_*, z_*) - \mathcal{L}(x_*, z) = z_*^\top G(x_*) - z^\top G(x_*) \geq 0, \quad (13)$$

where the last step follows from the nonnegativity of z and (10), completing the proof. \square

We now present our first main result that provides a bound on the expected distance to optimality and constraint violation at a weighted average of the iterates generated by the algorithm on \mathcal{P}^E under Assumption 1. Denote by \mathbf{C}_G , \mathbf{D}_G , and σ_G the collections of C_G^i , D_G^i , and σ_G^i , respectively. We make use of the following notation.

$$\begin{aligned} P_1 &:= 2\|x_1 - x_*\|^2 + 4\|1 + z_*\|^2, \\ P_2 &:= 8(4C_F^2 + \sigma_F^2) + 2\|\mathbf{D}_G\|^2, \\ P_3 &:= 8m(4\|\mathbf{C}_G\|^2 + \|\sigma_G\|^2). \end{aligned} \quad (14)$$

Theorem 2.1 (Convergence result for \mathcal{P}^E) *Suppose Assumption 1 holds. For a positive sequence $\{\gamma_k\}_{k=1}^K$, if $P_3 \sum_{k=1}^K \gamma_k^2 < 1$, then the iterates generated by Algorithm 1 satisfy*

$$\mathbb{E}[F(\bar{x}_{K+1})] - p_*^E \leq \frac{1}{4 \sum_{k=1}^K \gamma_k} \left(\frac{P_1 + P_2 \sum_{k=1}^K \gamma_k^2}{1 - P_3 \sum_{k=1}^K \gamma_k^2} \right), \quad (15)$$

$$\mathbb{E}[G^i(\bar{x}_{K+1})] \leq \frac{1}{4 \sum_{k=1}^K \gamma_k} \left(\frac{P_1 + P_2 \sum_{k=1}^K \gamma_k^2}{1 - P_3 \sum_{k=1}^K \gamma_k^2} \right) \quad (16)$$

for each $i = 1, \dots, m$, where $\bar{x}_{K+1} := \frac{\sum_{k=1}^K \gamma_k x_{k+1}}{\sum_{k=1}^K \gamma_k}$. Moreover, if $\gamma_k = \gamma / \sqrt{K}$ for $k = 1, \dots, K$ with $0 < \gamma < P_3^{-1/2}$, then

$$\mathbb{E}[F(\bar{x}_{K+1})] - p_*^E \leq \frac{\eta}{\sqrt{K}}, \quad \mathbb{E}[G^i(\bar{x}_{K+1})] \leq \frac{\eta}{\sqrt{K}} \quad (17)$$

for $i = 1, \dots, m$, where $\eta := \frac{P_1 + P_2 \gamma^2}{4\gamma(1 - P_3 \gamma^2)}$.

A constant step-size of η/\sqrt{K} over a fixed number of K iterations yields the $\mathcal{O}(1/\sqrt{K})$ decay rate in the expected distance to optimality and constraint violation of Algorithm 1. This is indeed order optimal, as implied by Nesterov's celebrated result in [32, Theorem 3.2.1].

Remark 2.1 While we present the proof for an i.i.d. sequence of samples, we believe that the result can be extended to the case where ω 's follow a Markov chain with geometric mixing rate following the technique in [41]. For such settings, the expectations in the definition of F , G should be computed with respect to the stationary distribution of the chain. The results will then possibly apply to Markov decision processes with applications in stochastic control.

Given that the literature on primal-dual subgradient methods is extensive, it is important for us to relate and distinguish Algorithm 1 and Theorem 2.1 with prior work. Using the Lagrangian in (6), Algorithm 1 can be written as

$$\begin{aligned} \mathbf{x}_{k+1} &:= \text{proj}_{\mathbb{X}}[\mathbf{x}_k - \gamma_k \nabla_{\mathbf{x}} \mathcal{L}_{\omega}(\mathbf{x}_k, \mathbf{z}_k)], \\ \mathbf{z}_{k+1} &:= \text{proj}_{\mathbb{R}_{+}^m}[\mathbf{z}_k + \gamma_k \nabla_{\mathbf{z}} \mathcal{L}_{\omega}(\mathbf{x}_{k+1}, \mathbf{z}_k)], \end{aligned} \quad (18)$$

where $\text{proj}_{\mathbb{A}}$ projects its argument on set \mathbb{A} . The vectors $\nabla_{\mathbf{x}} \mathcal{L}_{\omega}$ and $\nabla_{\mathbf{z}} \mathcal{L}_{\omega}$ are stochastic subgradients of the Lagrangian function with respect to \mathbf{x} and \mathbf{z} , respectively. Therefore, Algorithm 1 is a projected stochastic subgradient algorithm that seeks to solve the saddle-point reformulation of \mathcal{P}^E in (7). Implicit in our algorithm is the assumption that projection on \mathbb{X} is computationally easy. Any functional constraints describing \mathbb{X} that makes such projection challenging should be included in G .

Closest in spirit to our work on \mathcal{P}^E are the papers by Baes et al. in [2], Yu et al. in [44], Xu in [42], and Nedic and Ozdaglar in [30]. Stochastic mirror-prox algorithm in [2] and projected subgradient method in [30] are similar in their updates to ours except in two ways. First, these algorithms in the context of \mathcal{P}^E update the dual variable \mathbf{z}_k based on G or its noisy estimate evaluated at \mathbf{x}_k , while we update it based on the estimate at \mathbf{x}_{k+1} . Second, both project the dual variable on a compact subset of \mathbb{R}_{+}^m that contains the optimal set of dual multipliers. While authors in [2] assume an a priori set to project on, authors in [30] compute such a set from a “Slater point” that satisfies $G(\mathbf{x}) < 0$. Specifically, Slater's condition guarantees that the set of optimal dual solutions \mathbb{Z}_{*} is bounded (see [6, Theorem 1.265], [18]). Moreover, a Slater point can be used to construct a compact set that contains \mathbb{Z}_{*} , e.g., using [30, Lemma 4.1]. While one can project dual variables on such a set in each iteration, execution of the algorithm then requires a priori knowledge of such a point. We do not assume knowledge of such a point (or any explicit bound on \mathbb{Z}_{*}) to execute Algorithm 1. Rather, our proof provides an explicit bound on the growth of the dual variable sequence for Algorithm 1, much in line with Xu's analysis in [42]. Much to our surprise, a minor modification of using a Gauss–Seidel style dual update as opposed to the popular Jacobi style dual update obviates the need for this assumption in the literature for the proofs to work. Unfortunately, our Gauss–Seidel style dual update comes at an additional cost of an extra sample required per iteration of the primal-dual algorithm, making the sample complexity double of the iteration complexity. The constant factor of two, however,

does not impact the order-wise complexity. We surmise that the additional sample and the Gauss–Seidel update of the dual variable helps to decouple the analysis of the primal and dual updates and points to a possible extension of our result to an asynchronous setting, often useful in engineering applications. We remark that while we do not utilize a priori knowledge of a dual optimal solution to explicitly restrict the dual variables within a set containing it, an overestimate of $\|z_1 - z_*\|$ is required to compute η to calculate the precise bound in (17). In other words, gauging the quality of the ergodic mean after K iterations still requires that knowledge. We suspect that the distance of the ergodic mean \bar{z}_{K+1} to the dual optimal set is crucial to bound the extent of expected suboptimality and constraint violation. While analysis such as that in [2] achieves it by explicitly imposing a bound on the entire trajectory of z_k 's, we do so by assuming a bound on the distance of the initial point to the optimal set and then characterizing the growth over that trajectory.

Our work shares some parallels with that in [42], but has an important difference. Xu considers a collection of deterministic constraint functions, i.e., g^ω is identical for all $\omega \in \Omega$, and considers a modified augmented Lagrangian function of the form $\tilde{\mathcal{L}}(\mathbf{x}, \mathbf{z}) := F(\mathbf{x}) + \frac{1}{m} \sum_{i=1}^m \varphi_\delta(\mathbf{x}, z^i)$, where

$$\varphi_\delta(\mathbf{x}, z^i) := \begin{cases} z^i g^i(\mathbf{x}) + \frac{\delta}{2} [g^i(\mathbf{x})]^2, & \text{if } \delta g^i(\mathbf{x}) + z^i \geq 0, \\ -\frac{(z^i)^2}{2\delta}, & \text{otherwise} \end{cases} \quad (19)$$

for $i = 1, \dots, m$ with a suitable time-varying sequence of δ 's. His algorithm is similar to Algorithm 1 but performs a randomized coordinate update for the dual variable instead of (5). To the best of our knowledge, Xu's analysis in [42] with such a Lagrangian function does not directly apply to our setting with stochastic constraints that is crucial for the subsequent analysis of the risk-sensitive problem $\mathcal{P}^{\text{CVaR}}$.

Finally, Yu et al.'s work in [44] provides an analysis of the algorithm that updates its dual variables using

$$z_{k+1} := \underset{\mathbf{z} \in \mathbb{R}_+^m}{\operatorname{argmax}} \langle \mathbf{v}_k, \mathbf{z} - \mathbf{z}_k \rangle - \frac{1}{2\gamma_k} \|\mathbf{z} - \mathbf{z}_k\|^2, \quad (20)$$

where $v^i := g_{\omega_k}^i(\mathbf{x}_k) + \langle \nabla g_{\omega_k}^i(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle$ for $i = 1, \dots, m$. In contrast, our \mathbf{z} -update in (5) samples $\omega_{k+1/2}$ and sets $\mathbf{v}_k := \mathbf{g}_{\omega_{k+1/2}}(\mathbf{x}_{k+1})$ at the already computed point \mathbf{x}_{k+1} . We are able to recover the $\mathcal{O}(1/\sqrt{K})$ decay rate of suboptimality and constraint violation with a proof technique much closer to the classical analysis of subgradient methods in [9, 30]. Unlike [44], we provide a clean characterization of the constant η in (17) that is crucial to study the growth in sample (and iteration) complexity of Algorithm 1 applied to a reformulation of $\mathcal{P}^{\text{CVaR}}$.

2.1 Proof of Theorem 2.1

The proof proceeds in three steps.

(a) We establish the following dissipation inequality that consecutive iterates of the algorithm satisfy.

$$\begin{aligned} \gamma_k \mathbb{E}[\mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}) - \mathcal{L}(\mathbf{x}, \mathbf{z}_k)] + \frac{1}{2} \mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}\|^2 + \frac{1}{2} \mathbb{E}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 \\ \leq \frac{1}{2} \mathbb{E}\|\mathbf{x}_k - \mathbf{x}\|^2 + \frac{1}{2} \mathbb{E}\|\mathbf{z}_k - \mathbf{z}\|^2 + \frac{1}{4} P_2 \gamma_k^2 + \frac{1}{4} P_3 \gamma_k^2 \mathbb{E}\|\mathbf{z}_k\|^2 \end{aligned} \quad (21)$$

for any $\mathbf{x} \in \mathbb{X}$ and $\mathbf{z} \in \mathbb{R}_+^m$.

(b) Next, we bound $\mathbb{E}\|\mathbf{z}_k\|^2$ generated by our algorithm from above using step (a) as

$$\mathbb{E}\|\mathbf{z}_k\|^2 \leq \frac{P_1 + P_2 A_K}{1 - P_3 A_K} \quad (22)$$

for $k = 1, \dots, K$, where $A_K := \sum_{k=1}^K \gamma_k^2$.

(c) We combine the results in steps (a) and (b) to complete the proof.

Define the filtration $\mathcal{W}_1 \subset \mathcal{W}_{1+1/2} \subset \mathcal{W}_2 \subset \dots$, where \mathcal{W}_k is the σ -algebra generated by the samples $\omega_1, \dots, \omega_{k-1/2}$ for k being multiples of $1/2$, starting from unity. Then, $\{\mathbf{x}_1, \mathbf{z}_1, \dots, \mathbf{x}_k, \mathbf{z}_k\}$ becomes \mathcal{W}_k -measurable, while $\{\mathbf{x}_1, \mathbf{z}_1, \dots, \mathbf{x}_k, \mathbf{z}_k, \mathbf{x}_{k+1}\}$ is $\mathcal{W}_{k+1/2}$ -measurable.

- *Step (a)—Proof of (21):* We first utilize the \mathbf{x} -update in (4) to prove

$$\begin{aligned} \mathbb{E}[F(\mathbf{x}_{k+1}) - F(\mathbf{x}) + \mathbf{z}_k^\top \mathbf{G}(\mathbf{x}_{k+1}) - \mathbf{z}_k^\top \mathbf{G}(\mathbf{x}) | \mathcal{W}_k] + \frac{1}{2\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}\|^2 | \mathcal{W}_k] \\ \leq \frac{1}{2\gamma_k} \|\mathbf{x}_k - \mathbf{x}\|^2 + 2\gamma_k(4C_F^2 + \sigma_F^2) + 2\gamma_k m(4\|\mathbf{C}_G\|^2 + \|\boldsymbol{\sigma}_G\|^2) \|\mathbf{z}_k\|^2 \end{aligned} \quad (23)$$

for all $\mathbf{x} \in \mathbb{X}$. Then, we utilize the \mathbf{z} -update in (5) to prove

$$\begin{aligned} \mathbb{E}[\mathbf{z}^\top \mathbf{G}(\mathbf{x}_{k+1}) - \mathbf{z}_k^\top \mathbf{G}(\mathbf{x}_{k+1}) | \mathcal{W}_k] + \frac{1}{2\gamma_k} \mathbb{E}[\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 | \mathcal{W}_k] \\ \leq \frac{1}{2\gamma_k} \|\mathbf{z}_k - \mathbf{z}\|^2 + \frac{\gamma_k}{2} \|\mathbf{D}_G\|^2 \end{aligned} \quad (24)$$

for all $\mathbf{z} \in \mathbb{R}_+^m$. The law of total probability is then applied to the sum of (23) and (24) followed by a multiplication by γ_k yielding the desired result in (21).

Proof of (23): The \mathbf{x} -update in (4) yields

$$\left\langle \mathbf{x}_{k+1} - \mathbf{x}, \nabla f_\omega(\mathbf{x}_k) + \sum_{i=1}^m z_k^i \nabla \mathbf{g}_\omega^i(\mathbf{x}_k) + \frac{1}{\gamma_k} (\mathbf{x}_{k+1} - \mathbf{x}_k) \right\rangle \leq 0. \quad (25)$$

We now simplify the inner product. The product with $\nabla f_\omega(\mathbf{x}_k)$ can be expressed as

$$\begin{aligned} \langle \mathbf{x}_{k+1} - \mathbf{x}, \nabla f_\omega(\mathbf{x}_k) \rangle &= \underbrace{\langle \mathbf{x}_{k+1} - \mathbf{x}_k, \nabla F(\mathbf{x}_{k+1}) \rangle}_{\geq F(\mathbf{x}_{k+1}) - F(\mathbf{x}_k)} + \langle \mathbf{x}_k - \mathbf{x}, \nabla f_\omega(\mathbf{x}_k) \rangle \\ &\quad - \langle \mathbf{x}_{k+1} - \mathbf{x}_k, \nabla F(\mathbf{x}_{k+1}) - \nabla f_\omega(\mathbf{x}_k) \rangle, \end{aligned} \quad (26)$$

where $\nabla F(\mathbf{x}_{k+1})$ denotes a subgradient of F at \mathbf{x}_{k+1} . The inequality for the first term follows from the convexity of F . Since $\mathbb{E}[\nabla f_\omega(\mathbf{x}_k) | \mathcal{W}_k] \in \partial F(\mathbf{x}_k)$ from [5], the expectation of the second summand on the right-hand side (RHS) of (26) satisfies

$$\mathbb{E}[\langle \mathbf{x}_k - \mathbf{x}, \nabla f_\omega(\mathbf{x}_k) \rangle | \mathcal{W}_k] = \langle \mathbf{x}_k - \mathbf{x}, \nabla F(\mathbf{x}_k) \rangle \geq F(\mathbf{x}_k) - F(\mathbf{x}). \quad (27)$$

Taking expectations in (26), the above relation implies

$$\begin{aligned} \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}, \nabla f_\omega(\mathbf{x}_k) \rangle | \mathcal{W}_k] \\ \geq \mathbb{E}[F(\mathbf{x}_{k+1}) - F(\mathbf{x}) - \langle \mathbf{x}_{k+1} - \mathbf{x}_k, \nabla F(\mathbf{x}_{k+1}) - \nabla f_\omega(\mathbf{x}_k) \rangle | \mathcal{W}_k]. \end{aligned} \quad (28)$$

Next, we bound the inner product with the second term on the RHS of (26). To that end, utilize the convexity of member functions in \mathbf{g}_ω and \mathbf{G} along the above lines to infer

$$\begin{aligned} &\sum_{i=1}^m \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}, z_k^i \nabla g_\omega^i(\mathbf{x}_k) \rangle | \mathcal{W}_k] \\ &\geq \sum_{i=1}^m \mathbb{E}[z_k^i G^i(\mathbf{x}_{k+1}) - z_k^i G^i(\mathbf{x}) | \mathcal{W}_k] \\ &\quad - \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}, z_k^i \nabla G^i(\mathbf{x}_{k+1}) - z_k^i \nabla g_\omega^i(\mathbf{x}_k) \rangle | \mathcal{W}_k] \\ &= \mathbb{E}[z_k^\top \mathbf{G}(\mathbf{x}_{k+1}) - z_k^\top \mathbf{G}(\mathbf{x}) | \mathcal{W}_k] \\ &\quad - \sum_{i=1}^m \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}, z_k^i \nabla G^i(\mathbf{x}_{k+1}) - z_k^i \nabla g_\omega^i(\mathbf{x}_k) \rangle | \mathcal{W}_k]. \end{aligned} \quad (29)$$

To tackle the inner product with the third term in the RHS of (25), we use the identity

$$\begin{aligned} &\left\langle \mathbf{x}_{k+1} - \mathbf{x}, \frac{1}{\gamma_k} (\mathbf{x}_{k+1} - \mathbf{x}_k) \right\rangle \\ &= \frac{1}{2\gamma_k} [\|\mathbf{x}_{k+1} - \mathbf{x}\|^2 - \|\mathbf{x}_k - \mathbf{x}\|^2 + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2]. \end{aligned} \quad (30)$$

The inequalities in (28), (29), and the equality in (30) together give

$$\begin{aligned}
 & \mathbb{E}[F(\mathbf{x}_{k+1}) - F(\mathbf{x}) + \mathbf{z}_k^\top \mathbf{G}(\mathbf{x}_{k+1}) - \mathbf{z}_k^\top \mathbf{G}(\mathbf{x}) | \mathcal{W}_k] \\
 & - \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}_k, \nabla F(\mathbf{x}_{k+1}) - \nabla f_\omega(\mathbf{x}_k) \rangle | \mathcal{W}_k] \\
 & - \sum_{i=1}^m \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}_k, z_k^i \nabla G^i(\mathbf{x}_{k+1}) - z_k^i \nabla g_\omega^i(\mathbf{x}_k) \rangle | \mathcal{W}_k] \\
 & + \frac{1}{2\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}\|^2 + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] \\
 & \leq \frac{1}{2\gamma_k} \|\mathbf{x}_k - \mathbf{x}\|^2.
 \end{aligned} \tag{31}$$

To simplify the above relation, apply Young's inequality to obtain

$$\begin{aligned}
 & \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}_k, \nabla F(\mathbf{x}_{k+1}) - \nabla f_\omega(\mathbf{x}_k) \rangle | \mathcal{W}_k] \\
 & \leq \frac{1}{4\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] + \gamma_k \mathbb{E}[\|\nabla F(\mathbf{x}_{k+1}) - \nabla f_\omega(\mathbf{x}_k)\|^2 | \mathcal{W}_k] \\
 & \leq \frac{1}{4\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] + 2\gamma_k \mathbb{E}[\|\nabla F(\mathbf{x}_{k+1}) \\
 & \quad - \mathbb{E}\nabla f_\omega(\mathbf{x}_k)\|^2 + \|\mathbb{E}\nabla f_\omega(\mathbf{x}_k) - \nabla f_\omega(\mathbf{x}_k)\|^2 | \mathcal{W}_k].
 \end{aligned} \tag{32}$$

Recall that $\mathbb{E}[\nabla f_\omega(\mathbf{x}_k) | \mathcal{W}_k] \in \partial F(\mathbf{x}_k)$, subgradients of F are bounded and ∇f_ω has bounded variance. Therefore, we infer from the above inequality that

$$\begin{aligned}
 & \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}_k, \nabla F(\mathbf{x}_{k+1}) - \nabla f_\omega(\mathbf{x}_k) \rangle | \mathcal{W}_k] \\
 & \leq \frac{1}{4\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] + 2\gamma_k(4C_F^2 + \sigma_F^2).
 \end{aligned} \tag{33}$$

Appealing to Young's inequality m times and a similar line of argument as above gives

$$\begin{aligned}
 & \sum_{i=1}^m \mathbb{E}[\langle \mathbf{x}_{k+1} - \mathbf{x}_k, z_k^i \nabla G^i(\mathbf{x}_{k+1}) - z_k^i \nabla g_\omega^i(\mathbf{x}_k) \rangle | \mathcal{W}_k] \\
 & \leq \frac{1}{4\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] + 2\gamma_k m \sum_{i=1}^m (4[C_G^i]^2 + [\sigma_G^i]^2) \|z_k\|^2.
 \end{aligned} \tag{34}$$

Leveraging the relations in (33) and (34) in (31), we get

$$\begin{aligned}
 & \mathbb{E}[F(\mathbf{x}_{k+1}) - F(\mathbf{x}) + \mathbf{z}_k^\top \mathbf{G}(\mathbf{x}_{k+1}) - \mathbf{z}_k^\top \mathbf{G}(\mathbf{x}) | \mathcal{W}_k] \\
 & + \frac{1}{2\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}\|^2 + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] \\
 & \leq \frac{1}{2\gamma_k} \|\mathbf{x}_k - \mathbf{x}\|^2 + \frac{1}{4\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] + 2\gamma_k(4C_F^2 + \sigma_F^2) \\
 & + \frac{1}{4\gamma_k} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] + 2\gamma_k m \sum_{i=1}^m (4[C_G^i]^2 + [\sigma_G^i]^2) \|\mathbf{z}_k\|^2,
 \end{aligned} \tag{35}$$

that upon simplification gives (23).

Proof of (24): From the z -update in (5), we obtain

$$\left\langle \mathbf{z}_{k+1} - \mathbf{z}, -\mathbf{g}_\omega(\mathbf{x}_{k+1}) + \frac{1}{\gamma_k}(\mathbf{z}_{k+1} - \mathbf{z}_k) \right\rangle \leq 0 \tag{36}$$

for all $\mathbf{z} \geq 0$. Again, we deal with the two summands in the second factor of the inner product of (36) separately. The expectation of the inner product with the first term yields²

$$\begin{aligned}
 & \mathbb{E}[\langle \mathbf{z}_{k+1} - \mathbf{z}, -\mathbf{g}_\omega(\mathbf{x}_{k+1}) \rangle | \mathcal{W}_{k+1/2}] \\
 & = \mathbb{E}[\langle \mathbf{z}_{k+1} - \mathbf{z}_k, -\mathbf{g}_\omega(\mathbf{x}_{k+1}) \rangle | \mathcal{W}_{k+1/2}] + \mathbb{E}[\langle \mathbf{z}_k - \mathbf{z}, -\mathbf{g}_\omega(\mathbf{x}_{k+1}) \rangle | \mathcal{W}_{k+1/2}] \\
 & \geq -\frac{1}{2\gamma_k} \mathbb{E}[\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 | \mathcal{W}_{k+1/2}] - \frac{\gamma_k}{2} \mathbb{E}[\|\mathbf{g}_\omega(\mathbf{x}_{k+1})\|^2 | \mathcal{W}_{k+1/2}] \\
 & + \mathbb{E}[\langle \mathbf{z}_k - \mathbf{z}, -\mathbf{g}_\omega(\mathbf{x}_{k+1}) \rangle | \mathcal{W}_{k+1/2}] \\
 & \geq -\frac{1}{2\gamma_k} \mathbb{E}[\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 | \mathcal{W}_{k+1/2}] - \frac{\gamma_k}{2} \|\mathbf{D}_G\|^2 + \langle \mathbf{z}_k - \mathbf{z}, -\mathbf{G}(\mathbf{x}_{k+1}) \rangle \\
 & = -\frac{1}{2\gamma_k} \mathbb{E}[\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 | \mathcal{W}_{k+1/2}] - \frac{\gamma_k}{2} \|\mathbf{D}_G\|^2 + \mathbf{z}^\top \mathbf{G}(\mathbf{x}_{k+1}) - \mathbf{z}_k^\top \mathbf{G}(\mathbf{x}_{k+1}).
 \end{aligned} \tag{37}$$

In the above derivation, we have utilized Young's inequality and the boundedness of the second moment of \mathbf{g}_ω . Since $\mathcal{W}_k \subset \mathcal{W}_{k+1/2}$, the law of total probability can be used to condition (37) on \mathcal{W}_k rather than on $\mathcal{W}_{k+1/2}$. To simplify the inner product with the second term in (36), we use the identity

$$\left\langle \mathbf{z}_{k+1} - \mathbf{z}, \frac{1}{\gamma_k}(\mathbf{z}_{k+1} - \mathbf{z}_k) \right\rangle = \frac{1}{2\gamma_k} [\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \|\mathbf{z}_k - \mathbf{z}\|^2 + \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2]. \tag{38}$$

² $\mathbb{E}[\mathbf{g}_\omega(\mathbf{x}_{k+1}) | \mathcal{W}_{k+1/2}] = \mathbf{G}(\mathbf{x}_{k+1})$ requires that we sample ω once more for the z -update.

Utilizing (37) and (38) in (36) gives (24). Adding (23) and (24) followed by a multiplication by γ_k yields

$$\begin{aligned} & \gamma_k \mathbb{E}[\mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}) - \mathcal{L}(\mathbf{x}, \mathbf{z}_k) | \mathcal{W}_k] + \frac{1}{2} \mathbb{E} \left[\|\mathbf{x}_{k+1} - \mathbf{x}\|^2 + \|\mathbf{z}_{k+1} - \mathbf{z}\|^2 | \mathcal{W}_k \right] \\ & \leq \frac{1}{2} \|\mathbf{x}_k - \mathbf{x}\|^2 + \frac{1}{2} \|\mathbf{z}_k - \mathbf{z}\|^2 + \frac{1}{4} P_2 \gamma_k^2 + \frac{1}{4} P_3 \gamma_k^2 \|\mathbf{z}_k\|^2. \end{aligned} \quad (39)$$

Taking the expectation and applying the law of total probability completes the proof of (21).

• *Step (b)—Proof of (22):* Plugging $(\mathbf{x}, \mathbf{z}) = (\mathbf{x}_\star, \mathbf{z}_\star)$ in the inequality for the one-step update in (21) and summing it over $k = 1, \dots, \kappa$ for $\kappa \leq K$ gives

$$\begin{aligned} & \sum_{k=1}^{\kappa} \gamma_k \underbrace{\mathbb{E}[\mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}_\star) - \mathcal{L}(\mathbf{x}_\star, \mathbf{z}_k)]}_{\geq 0 \text{ from (9)}} + \frac{1}{2} \sum_{k=1}^{\kappa} \left[\mathbb{E} \|\mathbf{x}_{k+1} - \mathbf{x}_\star\|^2 + \mathbb{E} \|\mathbf{z}_{k+1} - \mathbf{z}_\star\|^2 \right] \\ & \leq \frac{1}{2} \sum_{k=1}^{\kappa} \left[\mathbb{E} \|\mathbf{x}_k - \mathbf{x}_\star\|^2 + \mathbb{E} \|\mathbf{z}_k - \mathbf{z}_\star\|^2 \right] + \frac{1}{4} P_2 \sum_{k=1}^{\kappa} \gamma_k^2 + \frac{1}{4} P_3 \sum_{k=1}^{\kappa} \gamma_k^2 \mathbb{E} \|\mathbf{z}_k\|^2 \end{aligned} \quad (40)$$

for $\kappa = 1, \dots, K$. The above then yields

$$\begin{aligned} & \underbrace{\mathbb{E} \|\mathbf{x}_{\kappa+1} - \mathbf{x}_\star\|^2 + \mathbb{E} \|\mathbf{z}_{\kappa+1} - \mathbf{z}_\star\|^2}_{\geq 0} \\ & \leq \|\mathbf{x}_1 - \mathbf{x}_\star\|^2 + \|\mathbf{z}_1 - \mathbf{z}_\star\|^2 + \frac{1}{2} P_2 \sum_{k=1}^{\kappa} \gamma_k^2 + \frac{1}{2} P_3 \sum_{k=1}^{\kappa} \gamma_k^2 \mathbb{E} \|\mathbf{z}_k\|^2. \end{aligned} \quad (41)$$

Notice that $2\mathbb{E} \|\mathbf{z}_{\kappa+1} - \mathbf{z}_\star\|^2 + 2\|\mathbf{z}_\star\|^2 \geq \mathbb{E} \|\mathbf{z}_{\kappa+1}\|^2$. This inequality and $\mathbf{z}_1 = 0$ in (41) gives

$$\begin{aligned} \mathbb{E} \|\mathbf{z}_{\kappa+1}\|^2 & \leq 2\|\mathbf{x}_1 - \mathbf{x}_\star\|^2 + 4\|\mathbf{z}_\star\|^2 + P_2 \sum_{k=1}^{\kappa} \gamma_k^2 + P_3 \sum_{k=1}^{\kappa} \gamma_k^2 \mathbb{E} \|\mathbf{z}_k\|^2 \\ & \leq P_1 + P_2 \sum_{k=1}^{\kappa} \gamma_k^2 + P_3 \sum_{k=1}^{\kappa} \gamma_k^2 \mathbb{E} \|\mathbf{z}_k\|^2. \end{aligned} \quad (42)$$

We argue the bound on $\mathbb{E} \|\mathbf{z}_k\|^2$ for $k = 1, \dots, K$ inductively. Since $\mathbf{z}_1 = 0$, the base case trivially holds. Assume that the bound holds for $k = 1, \dots, \kappa$ for $\kappa < K$. With the notation $A_K = \sum_{k=1}^K \gamma_k^2$, the relation in (42) implies

$$\mathbb{E} \|\mathbf{z}_{\kappa+1}\|^2 \leq P_1 + P_2 \sum_{k=1}^{\kappa} \gamma_k^2 + P_3 \sum_{k=1}^{\kappa} \gamma_k^2 \frac{P_1 + P_2 A_K}{1 - P_3 A_K}$$

$$\begin{aligned}
&\leq P_1 + P_2 A_K + P_3 \frac{P_1 + P_2 A_K}{1 - P_3 A_K} A_K \\
&= \frac{P_1 + P_2 A_K}{1 - P_3 A_K},
\end{aligned} \tag{43}$$

completing the proof of step (b).

• *Step (c)—Combining steps (a) and (b) to prove Theorem 2.1:* For any $z \geq 0$, the inequality in (21) with $x = x_*$ from step (a) summed over $k = 1, \dots, K$ gives

$$\begin{aligned}
&\sum_{k=1}^K \gamma_k \mathbb{E}[\mathcal{L}(x_{k+1}, z) - \mathcal{L}(x_*, z_k)] + \frac{1}{2} \sum_{k=1}^K \left[\mathbb{E}\|x_{k+1} - x_*\|^2 + \mathbb{E}\|z_{k+1} - z\|^2 \right] \\
&\leq \frac{1}{2} \sum_{k=1}^K \left[\mathbb{E}\|x_k - x_*\|^2 + \mathbb{E}\|z_k - z\|^2 \right] + \frac{1}{4} P_2 \sum_{k=1}^K \gamma_k^2 + \frac{1}{4} P_3 \sum_{k=1}^K \gamma_k^2 \mathbb{E}\|z_k\|^2.
\end{aligned} \tag{44}$$

Using $z_1 = 0$ and an appeal to the saddle point property of (x_*, z_*) yields

$$\begin{aligned}
&\sum_{k=1}^K \gamma_k \mathbb{E}[\mathcal{L}(x_{k+1}, z) - \mathcal{L}(x_*, z_*)] + \frac{1}{2} \mathbb{E}\|x_{K+1} - x_*\|^2 + \frac{1}{2} \mathbb{E}\|z_{K+1} - z\|^2 \\
&\leq \frac{1}{2} \mathbb{E}\|x_1 - x_*\|^2 + \frac{1}{4} P_2 \sum_{k=1}^K \gamma_k^2 + \frac{1}{4} P_3 \sum_{k=1}^K \gamma_k^2 \mathbb{E}\|z_k\|^2 + \frac{1}{2} \|z\|^2 \\
&\leq \frac{1}{4} P_1 - \|\mathbf{1} + z_*\|^2 + \frac{1}{4} P_2 A_K + \frac{1}{4} P_3 \sum_{k=1}^K \gamma_k^2 \frac{P_1 + P_2 A_K}{1 - P_3 A_K} + \frac{1}{2} \|z\|^2 \\
&= \frac{1}{4} P_1 + \frac{1}{4} P_2 A_K + \frac{1}{4} P_3 A_K \frac{P_1 + P_2 A_K}{1 - P_3 A_K} + \frac{1}{2} \|z\|^2 - \|\mathbf{1} + z_*\|^2 \\
&= \frac{1}{4} \left(\frac{P_1 + P_2 A_K}{1 - P_3 A_K} \right) + \frac{1}{2} \|z\|^2 - \|\mathbf{1} + z_*\|^2.
\end{aligned} \tag{45}$$

In deriving the above inequality, we have utilized the bound on $\mathbb{E}\|z_k\|^2$ from step (b) and the definition of P_1 and A_K . To further simplify the above inequality, notice that the saddle point property of (x_*, z_*) in (9) yields

$$F(x_*) = \mathcal{L}(x_*, 0) \leq \mathcal{L}(x_*, z_*) = F(x_*) + z_*^\top \mathbf{G}(x_*), \tag{46}$$

which implies $z_*^\top \mathbf{G}(x_*) \geq 0$. However, the saddle point theorem guarantees that x_* is an optimizer of \mathcal{P}^E , meaning that x_* is feasible and $\mathbf{G}(x_*) \leq 0$, implying $z_*^\top \mathbf{G}(x_*) \leq 0$ as $z_* \in \mathbb{R}_+^m$. Taken together, we infer

$$z_*^\top \mathbf{G}(x_*) = 0 \implies \mathcal{L}(x_*, z_*) = F(x_*) = p_*^E. \tag{47}$$

Since $\mathcal{L}(\mathbf{x}, z)$ is convex in \mathbf{x} , Jensen's inequality and (47) implies

$$\sum_{k=1}^K \gamma_k \mathbb{E}[\mathcal{L}(\mathbf{x}_{k+1}, z) - \mathcal{L}(\mathbf{x}_*, z_*)] \geq \left(\sum_{k=1}^K \gamma_k \right) \mathbb{E}[\mathcal{L}(\bar{\mathbf{x}}_{K+1}, z) - p_*^E], \quad (48)$$

where recall that $\bar{\mathbf{x}}_{K+1}$ is the γ -weighted average of the iterates. Utilizing (48) in (45), we get

$$\left(\sum_{k=1}^K \gamma_k \right) \mathbb{E}[\mathcal{L}(\bar{\mathbf{x}}_{K+1}, z) - p_*^E] \leq \frac{1}{4} \left(\frac{P_1 + P_2 A_K}{1 - P_3 A_K} \right) + \frac{1}{2} \|z\|^2 - \|\mathbf{1} + z_*\|^2. \quad (49)$$

The above relation defines a bound on $\mathbb{E}[\mathcal{L}(\bar{\mathbf{x}}_{K+1}, z)]$ for every $z \geq 0$. Choosing $z = 0$ and noting $\|\mathbf{1} + z_*\|^2 \geq 0$, we get the bound on expected suboptimality in (15). To derive the bound on expected constraint violation in (16), notice that the saddle point property in (9) and (47) implies

$$\begin{aligned} & \mathbb{E}[\mathcal{L}(\bar{\mathbf{x}}_{K+1}, \mathbf{1}^i + z_*) - p_*^E] \\ &= \mathbb{E}[\mathcal{L}(\bar{\mathbf{x}}_{K+1}, z_*) - \mathcal{L}(\mathbf{x}_*, z_*)] + \mathbb{E}\left[[\mathbf{1}^i]^\top \mathbf{G}(\bar{\mathbf{x}}_{K+1}) \right] \\ &\geq \mathbb{E}[G^i(\bar{\mathbf{x}}_{K+1})], \end{aligned} \quad (50)$$

where $\mathbf{1}^i \in \mathbb{R}^m$ is a vector of all zeros except the i -th entry that is unity. Choosing $z = \mathbf{1}^i + z_*$ in (49) and the observation in (50) then gives

$$\begin{aligned} \mathbb{E}[G^i(\bar{\mathbf{x}}_{K+1})] &\leq \frac{1}{4 \sum_{k=1}^K \gamma_k} \left(\frac{P_1 + P_2 A_K}{1 - P_3 A_K} + 2\|\mathbf{1}^i + z_*\|^2 - 4\|\mathbf{1} + z_*\|^2 \right) \\ &\leq \frac{1}{4 \sum_{k=1}^K \gamma_k} \left(\frac{P_1 + P_2 A_K}{1 - P_3 A_K} \right) \end{aligned} \quad (51)$$

for each $i = 1, \dots, m$. This completes the proof of (16). The bounds in (17) are immediate from that in (15)–(16). This completes the proof of Theorem 2.1. \square

Remark 2.2 The bound in (16) can be sharpened to

$$\sum_{i=1}^m \mathbb{E}\left[G^i(\bar{\mathbf{x}}_{K+1}) \right]^+ \leq \frac{1}{4 \sum_{k=1}^K \gamma_k} \left(\frac{P_1 + P_2 \sum_{k=1}^K \gamma_k^2}{1 - P_3 \sum_{k=1}^K \gamma_k^2} \right) \quad (52)$$

using z defined by $z^i := z_*^i + \mathbb{I}_{\{G^i(\bar{\mathbf{x}}_{K+1}) > 0\}}$ for $i = 1, \dots, m$ in (49). Here, $\mathbb{I}_{\{A\}}$ is the indicator function, evaluating to 1 if A holds and 0 otherwise. This improved bound was suggested to us by an anonymous reviewer. Notice that (52) is a much tighter bound on the expected constraint violation per constraint than (16) when m is large.

In what follows, we offer insights into two specific aspects of our proof. First, we present our conjecture on where the Gauss–Seidel nature of our dual update obtained with an extra sample helps us circumvent the need for an a priori bound on the dual variable. Notice that our dual update allows us to derive the third line of (37) that ultimately yields the term $-\mathbf{z}_k^\top \mathbf{G}(\mathbf{x}_{k+1})$ in (24). This term conveniently disappears when (24) is added to the inequality in (23) obtained from the primal update. We conjecture that this cancellation made possible by our dual update makes the theoretical analysis particularly easy. We anticipate that the classical Jacobi-style dual iteration derived with one sample shared within the primal and the dual steps will not lead to said cancellation and yield a term of the form $\mathbf{z}_k^\top [\mathbf{G}(\mathbf{x}_{k+1}) - \mathbf{G}(\mathbf{x}_k)]$. Bounding the growth of such a term might prove challenging without an available bound on $\|\mathbf{z}_k\|$ and will likely require a different argument. A detailed comparison between the proof techniques of the Jacobi and the Gauss–Seidel updates is left for future endeavors.

Second, we comment on the presence of a dimensionless constant $\mathbf{1}$ in P_1 together with \mathbf{z}_* . We use the inequality in (21) to establish (49) that is valid at all $\mathbf{z} \geq 0$. Inspired by arguments in [42], we then utilize (49) not only at the dual iterate \mathbf{z}_k —that is often the case with many prior analyses—but also at $\mathbf{z} = 0$ and $\mathbf{z} = \mathbf{1}^i + \mathbf{z}_*$. Specifically, the nature of the Lagrangian function $\mathcal{L}(\mathbf{x}, \mathbf{z})$ in \mathbf{z} permits us to relate these evaluations at $\mathbf{z} = 0$ and $\mathbf{z} = \mathbf{1}^i + \mathbf{z}_*$ to the extents of suboptimality and constraint violation, respectively, using

$$\mathcal{L}(\mathbf{x}, 0) = F(\mathbf{x}), \quad \mathcal{L}(\mathbf{x}, \mathbf{1}^i + \mathbf{z}) = \mathcal{L}(\mathbf{x}, \mathbf{z}) + G^i(\mathbf{x}). \quad (53)$$

The deliberate inclusion of $\|\mathbf{1} + \mathbf{z}_*\|^2$ in constant P_1 aids in drowning the effect of the term $\frac{1}{2}\|\mathbf{z}\|^2$ in (49) evaluated at $\mathbf{z} = \mathbf{1}^i + \mathbf{z}_*$ when deriving the bound on the extent of constraint violation, without impacting the same when evaluated at $\mathbf{z} = 0$, used in deriving the bound on the extent of suboptimality.

2.2 Optimal Step Size Selection

We exploit the bounds in Theorem 2.1 to select a step size that minimizes the iteration count to reach an ε -approximately feasible and optimal solution to \mathcal{P}^E and solve³

$$\begin{aligned} & \underset{K, \gamma > 0}{\text{minimize}} && K, \\ & \text{subject to} && \frac{\eta}{\sqrt{K}} = \frac{P_1 + P_2\gamma^2}{4\gamma\sqrt{K}(1 - P_3\gamma^2)} \leq \varepsilon, \quad P_3\gamma^2 < 1. \end{aligned} \quad (54)$$

The following characterization of optimal step sizes and the resulting iteration count from Proposition 2.1 will prove useful in studying the growth in iteration complexity in solving \mathcal{P}^{CVaR} with the risk-aversion parameters α, β in the following section.

³ The integrality of K is ignored for notational convenience.

Proposition 2.1 For any $\varepsilon > 0$, the optimal solution of (54) satisfies

$$\gamma_*^2 = \frac{2P_3^{-1}}{2 + y + \sqrt{y^2 + 8y}}, \quad K_* = \frac{(P_1 + P_2\gamma_*^2)^2}{16\gamma_*^2(1 - P_3\gamma_*^2)^2\varepsilon^2}, \quad (55)$$

where $y = 1 + \frac{P_2}{P_1P_3}$.

Proof It is evident from (55) that $\gamma_*^2 < P_3^{-1}$. Then, it suffices to show that γ_* from (55) minimizes

$$\sqrt{K} = \frac{P_1 + P_2\gamma^2}{4\gamma(1 - P_3\gamma^2)\varepsilon} \quad (56)$$

over $\gamma > 0$. To that end, notice that

$$\frac{d}{d\gamma} \left(\frac{P_1 + P_2\gamma^2}{\gamma(1 - P_3\gamma^2)} \right) = \frac{P_2P_3\gamma^4 + (P_2 + 3P_1P_3)\gamma^2 - P_1}{\gamma^2(1 - P_3\gamma^2)^2}. \quad (57)$$

The above derivative is negative at $\gamma = 0^+$ and vanishes only at γ_* over positive values of γ , certifying it as the global minimizer. \square

Parameter P_1 is generally not known a priori. However, it is often possible to bound it from above. One can calculate γ_* and K_* using (55), replacing P_1 with its overestimate. Notice that

$$\frac{dK_*}{dP_1} := \frac{\partial K_*}{\partial P_1} + \frac{\partial K_*}{\partial \gamma_*} \frac{d\gamma_*}{dy} \frac{dy}{dP_1}. \quad (58)$$

It is straightforward to verify that $\frac{\partial K_*}{\partial P_1} > 0$, $\frac{dy}{dP_1} \leq 0$, and $\frac{\partial \gamma_*}{\partial y} \leq 0$, and hence, overestimating P_1 results in a smaller γ_* . Finally, $\frac{\partial K_*}{\partial \gamma} > 0$ for $\gamma > \gamma_*$, implying that K_* calculated with an overestimate of P_1 is larger than the optimal iteration count—the computational burden we must bear for not knowing P_1 . Our algorithm does require knowledge of P_3 to implement the algorithm that in turn depends only on the nature of the functions defining the constraints and not a primal-dual optimizer.

2.3 Asymptotic Almost Sure Convergence with Decaying Step-Sizes

Subgradient methods are often studied with decaying nonsummable square-summable step sizes, for which they converge to an optimizer in the unconstrained setting. The result holds even for distributed variants and for mirror descent methods (see [13]). Establishing convergence of Algorithm 1 to a primal-dual optimizer of \mathcal{P}^E is much more challenging without assumptions of strong convexity in the objective. With such step-sizes, we provide the following result to guarantee the stability of our algorithm, which is reminiscent of [28, Theorem 4].

Proposition 2.2 Suppose Assumption 1 holds and $\{\gamma_k\}_{k=1}^{\infty}$ is a nonsummable square-summable nonnegative sequence, i.e., $\sum_{k=1}^{\infty} \gamma_k = \infty$, $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$. Then, $(\mathbf{x}_k, \mathbf{z}_k)$ generated by Algorithm 1 remains bounded and $\lim_{k \rightarrow \infty} \mathcal{L}(\mathbf{x}_k, \mathbf{z}_*) - \mathcal{L}(\mathbf{x}_*, \mathbf{z}_k) = 0$ almost surely.

This ‘gap’ function $\mathcal{L}(\mathbf{x}, \mathbf{z}_*) - \mathcal{L}(\mathbf{x}_*, \mathbf{z})$ looks notoriously similar to the duality gap at (\mathbf{x}, \mathbf{z}) , but is not the same. We are unaware of any results on asymptotic almost sure convergence of primal-dual first-order algorithms to an optimizer for constrained convex programs with convex, but not necessarily strongly convex, objectives. A recent result in [43] establishes such a convergence in primal-dual dynamics in continuous time; our attempts at leveraging discretizations of the same have yet proven unsuccessful.

The proof of Proposition 2.2 takes advantage of the one-step update in (21) that makes it amenable to the well-studied almost supermartingale convergence result by Robbins and Siegmund in [35, Theorem 1].

Theorem (Convergence of almost supermartingales) *Let m_k, n_k, r_k, s_k be \mathcal{F}_k -measurable finite nonnegative random variables, where $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$ describes a filtration. If $\sum_{k=1}^{\infty} s_k < \infty$, $\sum_{k=1}^{\infty} r_k < \infty$, and*

$$\mathbb{E}[m_{k+1} | \mathcal{F}_k] \leq m_k(1 + s_k) + r_k - n_k, \quad (59)$$

then $\lim_{k \rightarrow \infty} m_k$ exists and is finite and $\sum_{k=1}^{\infty} n_k < \infty$ almost surely.

Proof of Proposition 2.2 Using notation from the proof of Theorem 2.1, (23) and (24) together yields

$$\begin{aligned} & \gamma_k \mathbb{E}[\mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}_*) - \mathcal{L}(\mathbf{x}, \mathbf{z}_k) | \mathcal{W}_k] + \frac{1}{2} \mathbb{E} \left[\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 + \|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2 | \mathcal{W}_k \right] \\ & \leq \frac{1}{2} \left[\|\mathbf{x}_k - \mathbf{x}_*\|^2 + \|\mathbf{z}_k - \mathbf{z}_*\|^2 \right] + \frac{1}{4} P_2 \gamma_k^2 + \frac{1}{4} P_3 \gamma_k^2 \|\mathbf{z}_k\|^2. \end{aligned} \quad (60)$$

We utilize the above to derive a similar inequality replacing $\mathbb{E}[\mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}_*) | \mathcal{W}_k]$ with $\mathcal{L}(\mathbf{x}_k, \mathbf{z}_*)$ by bounding the difference between them. Then, we apply the almost supermartingale convergence theorem to the result to conclude the proof. To bound said difference, the convexity of \mathcal{L} in \mathbf{x} and Young’s inequality together implies

$$\begin{aligned} & \mathcal{L}(\mathbf{x}_k, \mathbf{z}_*) - \mathbb{E}[\mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}_*) | \mathcal{W}_k] \\ & \leq \langle \nabla \mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}_*), \mathbf{x}_k - \mathbf{x}_{k+1} \rangle \\ & \leq \frac{\gamma_k}{2} \mathbb{E}[\|\nabla_x \mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}_*)\|^2 | \mathcal{W}_k] + \frac{1}{2\gamma_k} \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 | \mathcal{W}_k], \end{aligned} \quad (61)$$

where $\nabla_x \mathcal{L}$ denotes a subgradient of \mathcal{L} w.r.t. \mathbf{x} . To further bound the RHS of (61), Assumption 1 allows us to deduce

$$\begin{aligned} \|\nabla \mathcal{L}(\mathbf{x}, \mathbf{z}_\star)\|^2 &\leq 2\|\nabla F(\mathbf{x})\|^2 + 2m \sum_{i=1}^m \|\mathbf{z}_\star^i \nabla G^i(\mathbf{x})\|^2 \\ &\leq 2C_F^2 + 2m \|\mathbf{z}_\star\|^2 \|\mathbf{C}_G\|^2 \\ &:= 2Q_1. \end{aligned} \quad (62)$$

for any $\mathbf{x} \in \mathbb{X}$. Furthermore, the \mathbf{x} -update in (18) and the nonexpansive nature of the projection operator yield

$$\begin{aligned} &\frac{1}{\gamma_k^2} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] \\ &\leq \mathbb{E} \left[\left\| \nabla f_\omega(\mathbf{x}_k) + \sum_{i=1}^m \mathbf{z}_k^i \nabla g_\omega^i(\mathbf{x}_k) \right\|^2 | \mathcal{W}_k \right] \\ &\leq 2\mathbb{E}[\|\nabla f_\omega(\mathbf{x}_k)\|^2 | \mathcal{W}_k] + 2m \sum_{i=1}^m \mathbb{E} \left[(\mathbf{z}_k^i)^2 \|\nabla g_\omega^i(\mathbf{x}_k)\|^2 | \mathcal{W}_k \right]. \end{aligned} \quad (63)$$

From Assumption 1, we get

$$\begin{aligned} \mathbb{E}[\|\nabla f_\omega(\mathbf{x}_k)\|^2 | \mathcal{W}_k] &\leq 2\mathbb{E}[\|\nabla f_\omega(\mathbf{x}_k) - \mathbb{E}\nabla f_\omega(\mathbf{x}_k)\|^2 + \|\mathbb{E}\nabla f_\omega(\mathbf{x}_k)\|^2 | \mathcal{W}_k] \\ &\leq 2\sigma_F^2 + 2C_F^2, \end{aligned} \quad (64)$$

and along similar lines

$$\sum_{i=1}^m \mathbb{E} \left[(\mathbf{z}_k^i)^2 \|\nabla g_\omega^i(\mathbf{x}_k)\|^2 | \mathcal{W}_k \right] \leq 2(\|\boldsymbol{\sigma}_G\|^2 + \|\mathbf{C}_G\|^2) \|\mathbf{z}_k\|^2, \quad (65)$$

that together in (63) yield

$$\frac{1}{\gamma_k^2} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{W}_k] \leq \underbrace{4(\sigma_F^2 + C_F^2)}_{:=2Q_2} + \underbrace{4m(\|\boldsymbol{\sigma}_G\|^2 + \|\mathbf{C}_G\|^2)}_{:=2Q_3} \|\mathbf{z}_k\|^2. \quad (66)$$

Combining the above with (62) in (61) gives

$$\gamma_k (\mathcal{L}(\mathbf{x}_k, \mathbf{z}_\star) - \mathbb{E}[\mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}_\star) | \mathcal{W}_k]) \leq \gamma_k^2 (Q_1 + Q_2 + Q_3 \|\mathbf{z}_k\|^2). \quad (67)$$

Adding (67) to (60) and simplifying, we obtain

$$\begin{aligned} & \frac{1}{2} \mathbb{E} \left[\|\mathbf{x}_{k+1} - \mathbf{x}_\star\|^2 + \|\mathbf{z}_{k+1} - \mathbf{z}_\star\|^2 \mid \mathcal{W}_k \right] \\ & \leq \frac{1}{2} \left[\|\mathbf{x}_k - \mathbf{x}_\star\|^2 + \|\mathbf{z}_k - \mathbf{z}_\star\|^2 \right] - \gamma_k [\mathcal{L}(\mathbf{x}_k, \mathbf{z}_\star) - \mathcal{L}(\mathbf{x}_\star, \mathbf{z}_k)] \\ & \quad + \gamma_k^2 \left(\frac{1}{4} P_2 + Q_1 + Q_2 \right) + \gamma_k^2 \left(\frac{1}{4} P_3 + Q_3 \right) \|\mathbf{z}_k\|^2. \end{aligned} \quad (68)$$

The above inequality with

$$\|\mathbf{z}_k\|^2 \leq 2\|\mathbf{x}_k - \mathbf{x}_\star\|^2 + 2\|\mathbf{z}_k - \mathbf{z}_\star\|^2 + 2\|\mathbf{z}_\star\|^2 \quad (69)$$

becomes (59), where

$$\begin{aligned} m_k &= \frac{1}{2} \mathbb{E} \|\mathbf{x}_k - \mathbf{x}_\star\|^2 + \frac{1}{2} \mathbb{E} \|\mathbf{z}_k - \mathbf{z}_\star\|^2, \quad n_k = \gamma_k [\mathcal{L}(\mathbf{x}_k, \mathbf{z}_\star) - \mathcal{L}(\mathbf{x}_\star, \mathbf{z}_k)], \\ r_k &= \gamma_k^2 \left[\frac{1}{4} P_2 + Q_1 + Q_2 + \left(\frac{1}{2} P_3 + 2Q_3 \right) \|\mathbf{z}_\star\|^2 \right], \quad s_k = \gamma_k^2 \left(\frac{1}{2} P_3 + 2Q_3 \right). \end{aligned} \quad (70)$$

Each term is nonnegative, owing to (9), and γ defines a square-summable sequence. Applying [35, Theorem 1], m_k converges to a constant and $\sum_{k=1}^{\infty} n_k < \infty$. The latter combined with the nonsummability of γ implies the result. \square

3 Algorithm for $\mathcal{P}^{\text{CVaR}}$ and Its Analysis

We now devote our attention to solving $\mathcal{P}^{\text{CVaR}}$ via a primal-dual algorithm. To do so, we reformulate it as an instance of \mathcal{P}^E and utilize Algorithm 1 to solve that reformulation with constant step-sizes under a stronger set of assumptions given below. In the sequel, we use \mathcal{L} to denote the Lagrangian function defined in (6), but with F and G as defined in $\mathcal{P}^{\text{CVaR}}$.

Assumption 2 $\mathcal{P}^{\text{CVaR}}$ must satisfy the following properties:

- Subgradients of F and G are bounded, i.e., $\|\nabla f_\omega(\mathbf{x})\| \leq C_F$ and $\|\nabla g_\omega^i(\mathbf{x})\| \leq C_G^i$ almost surely for all $\mathbf{x} \in \mathbb{X}$.
- $\mathbf{g}_\omega(\mathbf{x})$ is bounded, i.e., $\|g_\omega^i(\mathbf{x})\| \leq D_G^i$ for all $\mathbf{x} \in \mathbb{X}$, almost surely.
- The Lagrangian function \mathcal{L} admits a saddle point $(\mathbf{x}_\star, \mathbf{z}_\star) \in \mathbb{X} \times \mathbb{R}_+^m$.

Using the variational characterization (2) of CVaR, rewrite $\mathcal{P}^{\text{CVaR}}$ as

$$\begin{aligned} & \underset{\mathbf{x} \in \mathbb{X}}{\text{minimize}} \quad \min_{u^0 \in \mathbb{R}} \mathbb{E}[\psi_\omega^f(\mathbf{x}, u^0; \alpha)], \\ & \text{subject to} \quad \min_{u^i \in \mathbb{R}} \mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)] \leq 0, \quad i = 1, \dots, m, \end{aligned} \quad (71)$$

⁴ Lemma 2.1 provides sufficient conditions for the existence of such a saddle point.

where $\psi_\omega^h(\mathbf{x}, u; \delta) := u + \frac{1}{1-\delta}[h_\omega(\mathbf{x}) - u]^+$ for any collection of convex functions $h_\omega : \mathbb{R}^n \rightarrow \mathbb{R}$, $\omega \in \Omega$. Coupled with Assumption 2, we will show that we can bound $|u^i| \leq D_G^i$ for each $i = 1, \dots, m$ ⁵ that allows us to rewrite $\mathcal{P}^{\text{CVaR}}$ as

$$\begin{aligned} \mathcal{P}^{\text{E}'} : \underset{\substack{\mathbf{x} \in \mathbb{X}, u^0 \in \mathbb{R}, \\ |\mathbf{u}| \leq \mathbf{D}_G}}{\text{minimize}} \quad & \mathbb{E}[\psi_\omega^f(\mathbf{x}, u^0; \alpha)], \\ \text{subject to} \quad & \mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)] \leq 0, \quad \text{for each } i = 1, \dots, m, \end{aligned} \quad (72)$$

where $|\cdot|$ denotes the element-wise absolute value. Call the optimal value of $\mathcal{P}^{\text{CVaR}}$ as p_\star^{CVaR} in the sequel.

Theorem 3.1 (Convergence result for $\mathcal{P}^{\text{CVaR}}$) *Suppose Assumption 2 holds. The iterates generated by Algorithm 1 on $\mathcal{P}^{\text{E}'}$ for $\mathcal{P}^{\text{CVaR}}$ with parameters α, β satisfy*

$$\mathbb{E}[\text{CVaR}_\alpha(f_\omega(\bar{\mathbf{x}}_{K+1}))] - p_\star^{\text{CVaR}} \leq \frac{\eta(\alpha, \beta)}{\sqrt{K}}, \quad (73)$$

$$\mathbb{E}[\text{CVaR}_{\beta^i}(g_\omega^i(\bar{\mathbf{x}}_{K+1}))] \leq \frac{\eta(\alpha, \beta)}{\sqrt{K}} \quad (74)$$

for $i = 1, \dots, m$ with step sizes $\gamma_k = \gamma/\sqrt{K}$ for $k = 1, \dots, K$ with $0 < \gamma < P_3^{-1/2}(\alpha, \beta)$, where $\eta(\alpha, \beta) := \frac{P_1 + \gamma^2 P_2(\alpha, \beta)}{4\gamma(1 - \gamma^2 P_3(\alpha, \beta))}$ and

$$\begin{aligned} P_2(\alpha, \beta) &:= \frac{16(C_F^2 + 1)}{(1 - \alpha)^2} + 2 \left\| \text{diag}(\mathbb{1} + \beta) \text{diag}(\mathbb{1} - \beta)^{-1} \mathbf{D}_G \right\|^2, \\ P_3(\alpha, \beta) &:= 16m \left\| \begin{pmatrix} \text{diag}(\mathbb{1} - \beta)^{-1} \mathbf{C}_G \\ \text{diag}(\mathbb{1} - \beta)^{-1} \mathbb{1} \end{pmatrix} \right\|^2. \end{aligned} \quad (75)$$

Proof We prove the result in the following steps.

- Under Assumption 2, we revise P_2 and P_3 in Theorem 2.1 for $\mathcal{P}^{\text{E}'}$.
- We show that if f_ω, g_ω satisfy Assumption 2, then ψ_ω^f and $\psi_\omega^{g^i}, i = 1, \dots, m$ satisfy Assumption 2, but with different bounds on the gradients and function values. Leveraging these bounds, we obtain $P_2(\alpha, \beta)$ and $P_3(\alpha, \beta)$ for $\mathcal{P}^{\text{E}'}$ using step (a).
- Using Assumption 2, we prove that the Lagrangian function $\mathcal{L}' : \mathbb{X} \times \mathbb{R} \times \mathbb{U} \times \mathbb{R}_+^m \rightarrow \mathbb{R}$ defined as

$$\mathcal{L}'(\mathbf{x}, u^0, \mathbf{u}, \mathbf{z}) := \mathbb{E}[\psi_\omega^f(\mathbf{x}, u^0; \alpha)] + \sum_{i=1}^m z^i \mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)] \quad (76)$$

admits a saddle point in $\mathbb{X} \times \mathbb{R} \times \mathbb{U} \times \mathbb{R}_+^m$, where $\mathbb{U} := \{\mathbf{u} \in \mathbb{R}^m \mid |\mathbf{u}| \leq \mathbf{D}_G\}$.

⁵ CVaR of any random variable can only vary between the mean and the maximum value that random variable can take.

(d) We then apply Theorem 2.1 with $P_2(\alpha, \beta)$ and $P_3(\alpha, \beta)$ from step (b) on $\mathcal{P}^{E'}$ to derive the bounds in (73) and (74).

• *Step (a)—Revising Theorem 2.1 with Assumption 2:* Recall that in the derivation of (33) in the proof of Theorem 2.1, Assumption 1 yields

$$\|\nabla F(\mathbf{x}_{k+1}) - \nabla f_\omega(\mathbf{x}_k)\|^2 \leq 2(4C_F^2 + \sigma_F^2). \quad (77)$$

Assumption 2 allows us to bound the same by $4C_F^2$, yielding $P_2 = 16C_F^2 + 2\|\mathbf{D}_G\|^2$. Along the same lines, we get $P_3 = 16m\|\mathbf{C}_G\|^2$.

• *Step (b)—Deriving properties of ψ_ω^f :* Consider the stochastic subgradient of $\psi_\omega^f(\mathbf{x}, t; \alpha)$ given by

$$\nabla \psi_\omega^f(\mathbf{x}, u; \alpha) = \begin{pmatrix} \frac{1}{1-\alpha} \nabla f_\omega(\mathbf{x}) \mathbb{I}_{\{f_\omega(\mathbf{x}) \geq u\}} \\ 1 - \frac{1}{1-\alpha} \mathbb{I}_{\{f_\omega(\mathbf{x}) \geq u\}} \end{pmatrix}, \quad (78)$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. Recall that $\|\nabla f_\omega(\mathbf{x})\| \leq C_F$ for all $\mathbf{x} \in \mathbb{X}$ almost surely. Therefore, we have

$$\begin{aligned} \|\nabla \psi_\omega^f(\mathbf{x}, u; \alpha)\|^2 &= \left\| \frac{1}{1-\alpha} \nabla f_\omega(\mathbf{x}) \mathbb{I}_{\{f_\omega(\mathbf{x}) \geq u\}} \right\|^2 + \left\| 1 - \frac{1}{1-\alpha} \mathbb{I}_{\{f_\omega(\mathbf{x}) \geq u\}} \right\|^2 \\ &\leq \frac{C_F^2 + 1}{(1-\alpha)^2}. \end{aligned} \quad (79)$$

Proceeding similarly, we obtain

$$\|\nabla \psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)\|^2 \leq \frac{[C_G^i]^2 + 1}{(1-\beta^i)^2}. \quad (80)$$

We also have

$$\|\psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)\| = \left\| \max \left\{ \frac{g_\omega^i(\mathbf{x}) - \beta^i u^i}{1 - \beta^i}, u^i \right\} \right\| \leq \frac{1 + \beta^i}{1 - \beta^i} D_G^i. \quad (81)$$

Then, (75) follows from step (a) using (79), (80), and (81).

• *Step (c)—Showing that \mathcal{L}' admits a saddle point:* According to [36, Theorem 10], the minimizers of $\mathbb{E}[\psi_\omega^f(\mathbf{x}, u^0; \alpha)]$ over u^0 define a nonempty closed bounded interval (possibly a singleton). Thus, we have

$$F(\mathbf{x}) = \mathbb{E}[\psi_\omega^f(\mathbf{x}, u^0(\mathbf{x}); \alpha)] \quad (82)$$

for some $u^0(\mathbf{x}) \in \mathbb{R}$ for each $\mathbf{x} \in \mathbb{X}$. Similarly, we infer

$$G^i(\mathbf{x}) = \mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i(\mathbf{x}); \beta^i)] \quad (83)$$

for some $u^i(\mathbf{x}) \in \mathbb{R}$ for each $\mathbf{x} \in \mathbb{X}$. Moreover, for all $u^i > D_G^i$, we have

$$\mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)] = u^i, \quad (84)$$

and for $u^i < -D_G^i$, we have

$$\mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)] = \frac{1}{1 - \beta^i} \left(\mathbb{E}[g_\omega^i(\mathbf{x})] - \beta^i u^i \right). \quad (85)$$

Thus, $\mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)]$ is nonincreasing in u^i below $-D_G^i$ and increasing in it beyond D_G^i . Hence, at least one among the minimizers of $\mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i; \beta^i)]$ must lie in $[-D_G^i, D_G^i]$. In the sequel, let $u^i(\mathbf{x})$ refer to such a minimizer.

Consider a saddle point $(\mathbf{x}_\star, z_\star) \in \mathbb{X} \times \mathbb{R}_+^m$ of $\mathcal{P}^{\text{CVaR}}$. We argue that $(\mathbf{x}_\star, u^0(\mathbf{x}_\star), \mathbf{u}(\mathbf{x}_\star), z_\star)$ is a saddle point of \mathcal{L}' . From the definitions of \mathcal{L} , \mathcal{L}' , (82), (83), and the saddle point property of $(\mathbf{x}_\star, z_\star)$, we obtain

$$\begin{aligned} \mathcal{L}'(\mathbf{x}_\star, u^0(\mathbf{x}_\star), \mathbf{u}(\mathbf{x}_\star), z_\star) &= \mathcal{L}(\mathbf{x}_\star, z_\star) \\ &\leq \mathcal{L}(\mathbf{x}, z_\star) \\ &= \mathbb{E}[\psi_\omega^f(\mathbf{x}, u^0(\mathbf{x}); \alpha)] + \sum_{i=1}^m z_\star^i \mathbb{E}[\psi_\omega^{g^i}(\mathbf{x}, u^i(\mathbf{x}); \beta^i)] \\ &\leq \mathcal{L}'(\mathbf{x}, u^0, \mathbf{u}, z_\star) \end{aligned} \quad (86)$$

for all $(\mathbf{x}, u^0, \mathbf{u}) \in \mathbb{X} \times \mathbb{R} \times \mathbb{U}$. Also, for all $z \in \mathbb{R}_+^m$, we have

$$\mathcal{L}'(\mathbf{x}_\star, u^0(\mathbf{x}_\star), \mathbf{u}(\mathbf{x}_\star), z) = \mathcal{L}(\mathbf{x}_\star, z) \leq \mathcal{L}(\mathbf{x}_\star, z_\star) = \mathcal{L}'(\mathbf{x}_\star, u^0(\mathbf{x}_\star), \mathbf{u}(\mathbf{x}_\star), z_\star). \quad (87)$$

• *Step (d)—Proof of (73) and (74):* By the saddle point theorem and (86), we have $\mathcal{L}(\mathbf{x}_\star, z_\star) = p_\star^{\text{CVaR}}$ that also equals the optimal value of $\mathcal{P}^{\mathcal{E}'}$. Applying Theorem 2.1 with revised P_2 and P_3 from step (b) to $\mathcal{P}^{\mathcal{E}'}$ for which $\mathbf{x}_0, \dots, \mathbf{x}_{K+1}$ and u_0^0, \dots, u_{K+1}^0 are $\mathcal{W}_{K+1/2}$ -measurable, we obtain

$$\begin{aligned} \mathbb{E}[\text{CVaR}_\alpha(f_\omega(\bar{\mathbf{x}}_{K+1}))] &= \mathbb{E} \left[\min_{u^0 \in \mathbb{R}} \mathbb{E}[\psi_\omega^f(\bar{\mathbf{x}}_{K+1}, u^0; \alpha) | \mathcal{W}_{K+1/2}] \right] \\ &\leq \mathbb{E} \left[\mathbb{E}[\psi_\omega^f(\bar{\mathbf{x}}_{K+1}, \bar{u}_{K+1}^0; \alpha) | \mathcal{W}_{K+1/2}] \right] \\ &= \mathbb{E} \left[\psi_\omega^f(\bar{\mathbf{x}}_{K+1}, \bar{u}_{K+1}^0; \alpha) \right] \\ &\leq p_\star^{\text{CVaR}} + \frac{\eta(\alpha, \beta)}{\sqrt{K}}. \end{aligned} \quad (88)$$

Following a similar argument for $i = 1, \dots, m$, we get

$$\mathbb{E} \left[\text{CVaR}_{\beta^i} (g_\omega^i(\bar{x}_{K+1})) \right] = \mathbb{E} \left[\min_{u^i \in \mathbb{R}} \mathbb{E}[\psi_\omega^{g^i}(\bar{x}_{K+1}, u^i; \beta^i) | \mathcal{W}_{K+1/2}] \right] \leq \frac{\eta(\alpha, \beta)}{\sqrt{K}}, \quad (89)$$

completing the proof. \square

Our proof architecture generalizes to problems with other risk measures as long as that measure preserves convexity of f_ω , g_ω , admits a variational characterization as in (2), and a subgradient for this modified objective can be easily computed and remains bounded. We restrict our attention to CVaR to keep the exposition concrete.

Opposed to sample average approximation (SAA) algorithms, we neither compute nor estimate $F(\mathbf{x}) = \text{CVaR}[f_\omega(\mathbf{x})]$, $G(\mathbf{x}) = \text{CVaR}[g_\omega(\mathbf{x})]$ for any given decision \mathbf{x} to run the algorithm. Yet, our analysis provides guarantees on the same at \bar{x}_{K+1} in expectation. If one needs to compute F at any decision variable, e.g., at \bar{x}_{K+1} , one can employ the variational characterization in (2). Such evaluation requires additional computational effort. Notice that Theorem 3.1 does not relate $F(\bar{x}_{K+1})$ to p_\star^{CVaR} in an almost sure sense; it only relates the two in expectation according to (73), where the expectation is evaluated with respect to the stochastic sample path.

CVaR of a random variable depends on the tail of its distribution. The higher the risk aversion, the further into the tail one needs to look, generally requiring more samples. Even if we do not explicitly compute the tail-dependent CVaR relevant to the objective or the constraints, it is natural to expect our sample complexity to grow with risk aversion, which the following result confirms.

Proposition 3.1 *Suppose Assumption 2 holds. For an ε -approximately feasible and optimal solution of $\mathcal{P}^{\text{CVaR}}$ with risk aversion parameters α, β using Algorithm 1 on $\mathcal{P}^{\text{E}'}$, then $\gamma_\star(\alpha, \beta)$ and $K_\star(\alpha, \beta)$ from Proposition 2.1, respectively, decreases and increases with both α and β .*

Proof We borrow the notation from Proposition 2.1 and tackle the variation with α and β separately.

- *Variation with α :* P_2 increases with α , implying γ_\star decreases with α because $\frac{d\gamma_\star^2}{dy} \leq 0$ and $\frac{dy}{dP_2} \geq 0$. Furthermore, using $\frac{\partial K_\star}{\partial \gamma_\star} < 0$ for $\gamma < \gamma_\star$ and $\frac{\partial K_\star}{\partial P_2} \geq 0$ in

$$\frac{dK_\star}{dP_2} = \frac{\partial K_\star}{\partial P_2} + \frac{\partial K_\star}{\partial \gamma_\star} \frac{d\gamma_\star}{dP_2} \quad (90)$$

we infer that K_\star increases with α .

- *Variation with β^i :* Both P_2 and P_3 increase with β^i and

$$\frac{d\gamma_\star^2}{d\beta^i} = \frac{\partial \gamma_\star^2}{\partial P_2} \frac{dP_2}{d\beta^i} + \frac{\partial \gamma_\star^2}{\partial P_3} \frac{dP_3}{d\beta^i}. \quad (91)$$

Following an argument similar to that for the variation with α , the first term on the RHS of the above equation can be shown to be nonpositive. Next, we show that the second

term is nonpositive to conclude that γ_* decreases with β^i , where we use $\frac{dP_3}{d\beta^i} \geq 0$. Utilizing $\frac{P_2}{P_1 P_3} = y - 1$, we infer

$$\begin{aligned} \frac{\partial \gamma_*^2}{\partial P_3} &= -\frac{2}{P_3^2(2+y+\sqrt{y^2+8y})} + \frac{\partial \gamma_*^2}{\partial y} \frac{\partial y}{\partial P_3} \\ &= -\frac{2}{P_3^2(2+y+\sqrt{y^2+8y})} + 2 \frac{4+y+\sqrt{y^2+8y}}{P_3 \sqrt{y^2+8y}(2+y+\sqrt{y^2+8y})^2} \frac{P_2}{P_1 P_3^2} \\ &= -2 \frac{5y+4+3\sqrt{y^2+8y}}{P_3^2 \sqrt{y^2+8y}(2+y+\sqrt{y^2+8y})^2} \\ &\leq 0. \end{aligned} \tag{92}$$

To characterize the variation of K_* , notice that

$$\frac{dK_*}{d\beta^i} = \frac{\partial K_*}{\partial P_2} \frac{\partial P_2}{\partial \beta^i} + \frac{\partial K_*}{\partial P_3} \frac{\partial P_3}{\partial \beta^i}. \tag{93}$$

Again, the first term on the RHS of the above relation is nonnegative, owing to an argument similar to that used for the variation of K_* with α . We show $\frac{\partial K_*}{\partial P_3} \leq 0$ to conclude the proof. Treating K_* as a function of P_3 and γ_* , we obtain

$$\frac{dK_*}{dP_3} = \frac{\partial K_*}{\partial P_3} + \frac{\partial K_*}{\partial \gamma_*} \frac{\partial \gamma_*}{\partial P_3}. \tag{94}$$

It is straightforward to verify that the first summand is nonnegative. We have already argued that γ_* decreases with P_3 , and $\frac{\partial K_*}{\partial \gamma_*} < 0$ for $\gamma < \gamma_*$, implying that the second summand is nonnegative as well, completing the proof. \square

It is easy to compute the optimized iteration count $K_*(\alpha, \beta)$ and the optimized constant step-size $\gamma_*(\alpha, \beta)/\sqrt{K_*(\alpha, \beta)}$ from Proposition 2.1. The formula is omitted for brevity. Instead, we derive additional insight by fixing β and driving α towards unity. For such an α, β , we have

$$P_2(\alpha, \beta) \sim (1-\alpha)^{-2}, \quad \gamma_*(\alpha, \beta) \sim (1-\alpha), \quad K_*(\alpha, \beta) \sim \frac{1}{\varepsilon^2(1-\alpha)^2}. \tag{95}$$

With α approaching unity, notice that $\mathcal{P}^{\text{CVaR}}$ approaches a robust optimization problem. Thus, Algorithm 1 for $\mathcal{P}^{\text{E}'}$ is aiming to solve a robust optimization problem via sampling. Not surprisingly, the sample complexity exhibits unbounded growth with such robustness requirements, since we do not assume Ω to be finite. Also, this growth matches that of solving the SAA problem within ε -tolerance on the unconstrained problem to minimize $\widehat{F}(\mathbf{x}) := \frac{1}{K} \sum_{j=1}^K \psi_{\omega^j}^f(\mathbf{x}, u; \alpha)$. To see this, apply Theorem 2.1 on $\widehat{F}(\mathbf{x})$ with optimized step size from Proposition 2.1, where $P_2 \sim \|\nabla \widehat{F}(\mathbf{x})\|^2 \sim (1-\alpha)^{-2}$ and $P_3 = 1$.

Parallelization can lead to stronger bounds. More precisely, run stochastic approximation in parallel on N machines, each with K samples and compute $\langle \bar{\mathbf{x}} \rangle_{K+1} := \frac{1}{N} \sum_{j=1}^N \bar{\mathbf{x}}_{K+1}[j]$ using $\bar{\mathbf{x}}_{K+1}[1], \dots, \bar{\mathbf{x}}_{K+1}[N]$ obtained from the N separate runs. Then, we have

$$\begin{aligned} & \Pr \left\{ G^i (\langle \bar{\mathbf{x}} \rangle_{K+1}) \geq (1 + \tau) \eta(\alpha, \beta) / \sqrt{K} \right\} \\ & \leq \Pr \left\{ \frac{1}{N} \sum_{j=1}^N \text{CVaR}_{\beta^i} \left[g_\omega^i (\bar{\mathbf{x}}_{K+1}[j]) \right] \geq (1 + \tau) \frac{\eta(\alpha, \beta)}{\sqrt{K}} \right\} \quad (96) \\ & \leq \exp \left(-\frac{N\tau^2 \eta^2(\alpha, \beta)}{K [D_G^i]^2} \right) \end{aligned}$$

for $i = 1, \dots, m$ and $\tau > 0$. The steps combine coherence of CVaR, convexity and uniform boundedness of g_ω^i , Hoeffding's inequality and Theorem 3.1. A similar bound can be derived for suboptimality. Thus, parallelized stochastic approximation produces a result whose $\mathcal{O}(1/\sqrt{K})$ -violation occurs with a probability that decays exponentially with the degree of parallelization N .

The bound in (96) reveals an interesting connection with results for chance constrained programs. To describe the link, notice that $\text{CVaR}_\delta[y_\omega] \leq 0$ implies $\Pr\{y_\omega \leq 0\} \geq 1 - \delta$ for any random variable y_ω and $\delta \in [0, 1)$. Therefore, (96) implies

$$\Pr \left\{ \Pr \left\{ g_\omega^i (\bar{\mathbf{x}}_{K+1}) \leq C / \sqrt{K} \right\} \geq 1 - \beta^i \text{ is violated} \right\} \leq \exp(-C'/K) \leq \nu, \quad (97)$$

for constants C, C' . Said differently, our stochastic approximation algorithm requires $\mathcal{O}(\log(1/\nu))$ samples to produce a solution that satisfies an $\mathcal{O}(1/\sqrt{\log(1/\sqrt{\nu}))}$ -approximate chance-constraint with a violation probability bounded by ν . This result bears a striking similarity to that derived in [11], where the authors deterministically enforce $\mathcal{O}(\log(1/\nu))$ sampled constraints to produce a solution that satisfies the exact chance-constraint $\Pr \{ g_\omega^i(\mathbf{x}) \leq 0 \} \geq 1 - \beta^i$ with a violation probability bounded by ν . This resemblance in order-wise sample complexity is intriguing, given the significant differences between the algorithms.

3.1 An Illustrative Example

We explore the use of our algorithm on the following example problem

$$\underset{-\frac{1}{2} \leq x \leq \frac{1}{2}}{\text{minimize}} \text{CVaR}_\alpha \left[\frac{1}{2} \left(x - \omega - \frac{1}{2} \right)^2 \right], \text{ subject to } \text{CVaR}_\beta [x + \omega] \leq 0. \quad (98)$$

Let $\omega \sim \frac{1}{3} \text{beta}(2, 2)$ and consider the specific choice of risk parameters $\alpha = 0.3, \beta = 0.2$. To gain intuition into the optimal solution for this example, we numerically estimate $F(x)$ and $G^1(x)$ for each x and plot them in Fig. 1a. To that end, we first obtain

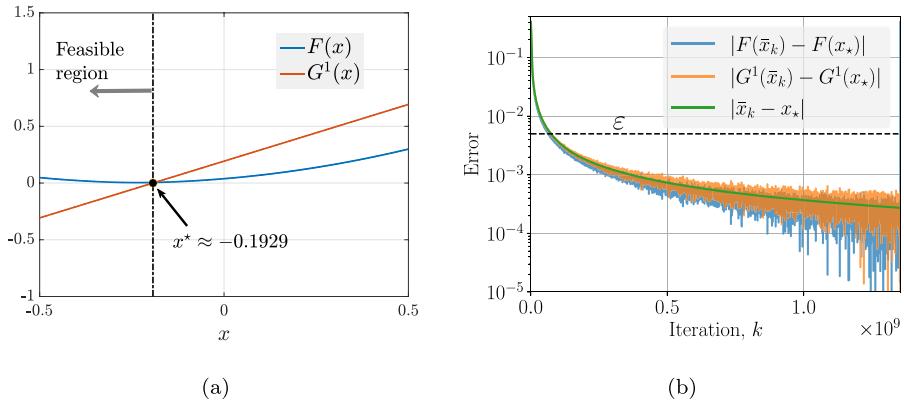


Fig. 1 Plots of **a** numerically estimated F and G^1 over $\mathbb{X} = [-\frac{1}{2}, \frac{1}{2}]$, and **b** convergence of the running ergodic mean and F , G evaluated at the mean for the example problem (98) with $\alpha = 0.3$, $\beta = 0.2$

a million samples of ω . Then, for each value of the decision variable x , we sort the objective function value $f_\omega(x)$ and the constraint function value $g_\omega(x)$ with these samples. We then estimate F and G^1 as the average of the highest $1 - \alpha = 70\%$ and $1 - \beta = 80\%$ among $f_\omega(x)$'s and $g_\omega(x)$'s, respectively, at each x with those samples. The unique optimum for (98) is numerically evaluated as $x_* \approx -0.1929$ for which $F(x_*) \approx 0.4042$ and $G^1(x_*) \approx 0$.

For this example, it is easy to show that $C_F = \frac{4}{3}$, $C_G = 1$ and $D_G = \frac{5}{6}$ that yields $P_2(0.3, 0.2) = \frac{8276}{93}$ and $P_3(0.3, 0.2) = 50$. To run Algorithm 1 on $\mathcal{P}^{E'}$ derived from (98), we can use constant step-size $\gamma_k = \gamma/\sqrt{K}$ with a pre-determined number of steps K for any $0 < \gamma < P_3^{-1/2}(0.3, 0.2) = \frac{1}{5\sqrt{2}}$. With any given K , Theorem 3.1 guarantees that the expected distance to $F(x_*)$ and the expected constraint violation evaluated at \bar{x}_{K+1} decays as $1/\sqrt{K}$. For a given K and $\gamma < \frac{1}{5\sqrt{2}}$, calculating the precise bound $\eta(0.3, 0.2)/\sqrt{K}$ requires the knowledge of P_1 or its overestimate. For this example, $|x_*| \leq \frac{1}{2}$ and $|u_*^1| \leq D_G = \frac{5}{6}$. Also, $|u_*^0|$ is bounded above by the maximum value that $|f_\omega(x)|$ can take, that is given by $\frac{8}{9}$. Since we cannot determine z_* a priori, we assume $|z_*| \leq 2$ (that will later be shown to be consistent with our result). Starting from $(x_0, u_0^0, u_0^1, z_0) = 0$, we then obtain $P_1 = \frac{3197}{81}$. To solve $\mathcal{P}^{\text{CVaR}}$ (or equivalently $\mathcal{P}^{E'}$) with a tolerance of $\varepsilon = 5 \times 10^{-3}$, we require $\eta(0.3, 0.2)/\sqrt{K} \leq 5 \times 10^{-3}$. With this tolerance and the values of P_1, P_2, P_3 , Proposition 2.1 yields an optimized $\gamma_* = 0.0808$ and $K_* \approx 1.35 \times 10^9$. We run Algorithm 1 on $\mathcal{P}^{E'}$ with constant step-size $\gamma_*/\sqrt{K_*}$ and plot F and G^1 at the running ergodic mean of the iterates, i.e., at $\bar{x}_k := \frac{1}{k} \sum_{j=1}^k x_j$ for each k . Again F and G^1 are evaluated numerically using the CVaR-estimation procedure we outlined above.

Notice that Theorem 3.1 only guarantees a bound on $F(\bar{x}_{K_*+1}) - F(\bar{x}_*)$ and $G^1(\bar{x}_{K_*+1})$ in expectation. Thus, one would expect that only the average of the CVaR of F and G^1 evaluated at \bar{x}_{K_*+1} over multiple sample paths to respect the ε -bound. However, our simulation yielded $\bar{x}_{K_*+1} = -0.1926$ and $\bar{z}_{K_*+1} = 0.8976$, for which

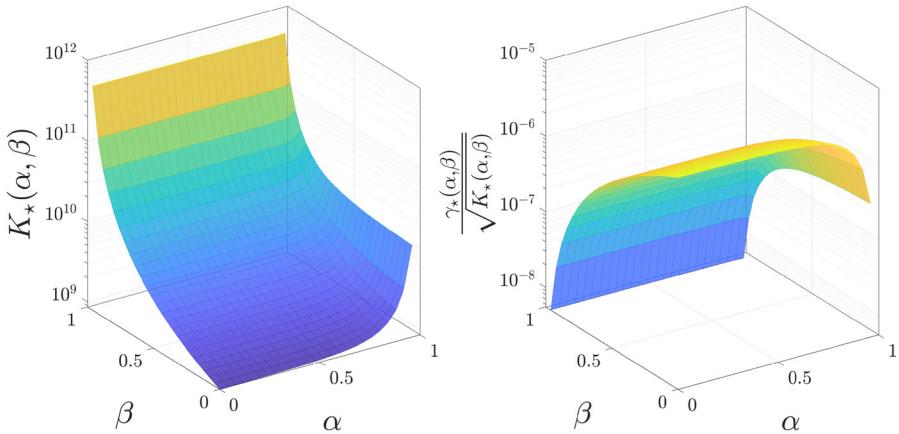


Fig. 2 Plot of the optimized number of iterations $K_*(\alpha, \beta)$ on the left and the optimized step size $\gamma_*(\alpha, \beta)/\sqrt{K_*(\alpha, \beta)}$ on the right to achieve a tolerance of $\varepsilon = 5 \times 10^{-3}$ for the example problem in (98)

$$\begin{aligned} F(\bar{x}_{K_*+1}) &\approx 0.4040 \leq F(x_*) + \varepsilon \approx 0.4042 + 0.0050 = 0.4092, \\ G^1(\bar{x}_{K_*+1}) &\approx 0.0002 \leq G^1(x_*) + \varepsilon \approx 0 + 0.0050 = 0.0050, \end{aligned} \quad (99)$$

i.e., the ergodic mean after K_* iterations respects the ε -bound over the plotted sample path. The same behavior was observed over multiple sample paths. The ergodic mean of the dual iterate is indeed consistent with our assumption $|z_*| \leq 2$ made in deriving $\eta(0.3, 0.2)$. We point out that the ergodic mean in Fig. 1b moves much more smoothly than our evaluation of F and G^1 at those means, especially for large k . The noise in F in G^1 emanates from the finitely many samples we use to evaluate F and G^1 . The errors appear much more pronounced at larger k , given the logarithmic scale of the plot.

The optimized iteration count $K_*(\alpha, \beta)$ from Proposition 2.1 with a modest $\alpha = 0.3$, $\beta = 0.2$ is quite high even for this simple example. This iteration count only grows with increased risk aversion as Fig. 2 reveals. Figure 1b suggests that the $\varepsilon = 5 \times 10^{-3}$ tolerance is met earlier than K_* iterations. This is the downside of optimizing upper bounds to decide step-sizes for subgradient methods. Carefully designed termination criteria may prove useful in practical implementations. In Fig. 2, we calculate K_* and γ_* with $P_1 = \frac{3197}{81}$ obtained using $|z_*| \leq 2$; extensive simulations with various $(\alpha, \beta) \in [0, 0.99]^2$ suggest that this over-estimate indeed holds.

We end the numerical example with a remark about the comparison of Algorithm 1 that uses Gauss–Seidel-type dual update in (5) and another that uses the popular Jacobi-type dual update on \mathcal{P}^E for (98) with $\alpha = 0.3$, $\beta = 0.2$. This alternate dual update replaces $\mathbf{g}_{\omega_{k+1/2}}(\mathbf{x}_{k+1})$ in (5) by $\mathbf{g}_{\omega_k}(\mathbf{x}_k)$. That is, the same sample ω_k is used for both the primal and the dual update. And, the primal iterate \mathbf{x}_k is used instead of \mathbf{x}_{k+1} to update the dual variable. We numerically compared this primal-dual algorithm with Algorithm 1 with various choices of step-sizes (consistent with the requirements of Theorem 3.1) and iteration count for our example and its variations. For each

run, we found that the iterates from both these algorithms moved very similarly. The differences are too small to report. The Jacobi-type update requires half the number of samples compared to Algorithm 1. While the extra sample helps us in the theoretical analysis, our experience with this stylized example does not suggest any empirical advantage. A more thorough comparison between these algorithms, both theoretically and empirically, is left to future work.

4 Conclusions and Future Work

In this paper, we study a stochastic approximation algorithm for CVaR-sensitive optimization problems. Such problems are remarkably rich in their modeling power and encompass a plethora of stochastic programming problems with broad applications. We study a primal-dual algorithm to solve that problem that processes samples in an online fashion, i.e., obtains samples and updates decision variables in each iteration. Such algorithms are useful when sampling is easy and intermediate approximate solutions, albeit inexact, are useful. The convergence analysis allows us to optimize the number of iterations required to reach a solution within a prescribed tolerance on expected suboptimality and constraint violation. The sample and iteration complexity predictably grows with risk-aversion. Our work affirms that a modeler must not only consider the attitude toward risk but also consider the computational burdens of risk in deciding the problem formulation.

Two possible extensions are of immediate interest to us. First, primal-dual algorithms find applications in multi-agent distributed optimization problems over a possibly time-varying communication network. We plan to extend our results to solve distributed risk-sensitive convex optimization problems over networks, borrowing techniques from [14,29]. Second, the relationship to sample complexity for chance-constrained programs in [11] encourages us to pursue a possible exploration of stochastic approximation for such optimization problems.

Acknowledgements We thank Eilyan Bitar, Rayadurgam Srikant, Tamer Başar, and Stan Uryasev for helpful discussions. This work was partially supported by the National Science Foundation under grant no. CAREER-2048065, the International Institute of Carbon-Neutral Energy Research (I²CNER), and the Power System Engineering Research Center (PSERC).

References

1. Ahmadi-Javid, A.: Entropic value-at-risk: a new coherent risk measure. *J. Optim. Theory Appl.* **155**(3), 1105–1123 (2012)
2. Baes, M., Bürgisser, M., Nemirovski, A.: A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems. *SIAM J. Optim.* **23**(2), 934–962 (2013)
3. Bedi, A.S., Koppel, A., Rajawat, K.: Nonparametric compositional stochastic optimization. arXiv preprint [arXiv:1902.06011](https://arxiv.org/abs/1902.06011) (2019)
4. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization, vol. 28. Princeton University Press, Princeton (2009)
5. Bertsekas, D.P.: Stochastic optimization problems with nondifferentiable cost functionals. *J. Optim. Theory Appl.* **12**(2), 218–231 (1973)
6. Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, Berlin (2013)

7. Boob, D., Deng, Q., Lan, G.: Stochastic first-order methods for convex and nonconvex functional constrained optimization. arXiv preprint [arXiv:1908.02734](https://arxiv.org/abs/1908.02734) (2019)
8. Borkar, V.S., Meyn, S.P.: The ode method for convergence of stochastic approximation and reinforcement learning. SIAM J. Control Optim. **38**(2), 447–469 (2000)
9. Boyd, S., Mutapcic, A.: Subgradient methods. Lecture notes of EE364b, Stanford University, Winter Quarter **2007** (2006)
10. Calafiori, G., Campi, M.C.: Uncertain convex programs: randomized solutions and confidence levels. Math. Program. **102**(1), 25–46 (2005)
11. Campi, M.C., Garatti, S.: The exact feasibility of randomized solutions of uncertain convex programs. SIAM J. Optim. **19**(3), 1211–1230 (2008)
12. Charnes, A., Cooper, W.W.: Chance-constrained programming. Manag. Sci. **6**(1), 73–79 (1959)
13. Doan, T.T., Bose, S., Nguyen, D.H., Beck, C.L.: Convergence of the iterates in mirror descent methods. IEEE Control Syst. Lett. **3**(1), 114–119 (2018)
14. Dominguez-Garcia, A.D., Hadjicostis, C.N.: Distributed matrix scaling and application to average consensus in directed graphs. IEEE Trans. Autom. Control **58**(3), 667–681 (2013)
15. Ermoliev, Y.M.: Methods of stochastic programming (1976)
16. Hadjjiyiannis, M.J., Goulart, P.J., Kuhn, D.: An efficient method to estimate the suboptimality of affine controllers. IEEE Trans. Autom. Control **56**(12), 2841–2853 (2011)
17. Hanasusanto, G.A., Kuhn, D., Wiesemann, W.: A comment on “computational complexity of stochastic programming problems”. Math. Program. **159**(1–2), 557–569 (2016)
18. Hirari-Urruty, J.B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms I: Fundamentals, vol. 305. Springer, Berlin (2013)
19. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems, pp. 315–323 (2013)
20. Kalogerias, D.S., Powell, W.B.: Recursive optimization of convex risk measures: mean-semideviation models. arXiv preprint [arXiv:1804.00636](https://arxiv.org/abs/1804.00636) (2018)
21. Kiefer, J., Wolfowitz, J., et al.: Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. **23**(3), 462–466 (1952)
22. Kisiala, J.: Conditional value-at-risk: theory and applications. arXiv preprint [arXiv:1511.00140](https://arxiv.org/abs/1511.00140) (2015)
23. Koppel, A., Sadler, B.M., Ribeiro, A.: Proximity without consensus in online multiagent optimization. IEEE Trans. Signal Process. **65**(12), 3062–3077 (2017)
24. Kushner, H., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications, vol. 35. Springer, Berlin (2003)
25. Mafusalov, A., Uryasev, S.: Buffered probability of exceedance: mathematical properties and optimization. SIAM J. Optim. **28**(2), 1077–1103 (2018)
26. Mahdavi, M., Jin, R., Yang, T.: Trading regret for efficiency: online convex optimization with long term constraints. J. Mach. Learn. Res. **13**(1), 2503–2528 (2012)
27. Miller, C.W., Yang, I.: Optimal control of conditional value-at-risk in continuous time. SIAM J. Control. Optim. **55**(2), 856–884 (2017)
28. Nedić, A., Lee, S.: On stochastic subgradient mirror-descent algorithm with weighted averaging. SIAM J. Optim. **24**(1), 84–107 (2014)
29. Nedić, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control **54**(1), 48–61 (2009)
30. Nedić, A., Ozdaglar, A.: Subgradient methods for saddle-point problems. J. Optim. Theory Appl. **142**(1), 205–228 (2009)
31. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
32. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Springer (2004)
33. Ogryczak, W., Ruszczyński, A.: From stochastic dominance to mean-risk models: semideviations as risk measures. Eur. J. Oper. Res. **116**(1), 33–50 (1999)
34. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
35. Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. In: Optimizing Methods in Statistics, pp. 233–257. Elsevier (1971)
36. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Bank. Finance **26**(7), 1443–1471 (2002)
37. Ruszczyński, A., Shapiro, A.: Optimization of convex risk functions. Math. Oper. Res. **31**(3), 433–452 (2006)

38. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. *Math. Program.* **162**(1–2), 83–112 (2017)
39. Shapiro, A., Philpott, A.: A tutorial on stochastic programming. Manuscript. Available at www2.isye.gatech.edu/ashapiro/publications.html 17 (2007)
40. Skaf, J., Boyd, S.P.: Design of affine controllers via convex optimization. *IEEE Trans. Autom. Control* **55**(11), 2476–2487 (2010)
41. Sun, T., Sun, Y., Yin, W.: On Markov chain gradient descent. In: *Advances in Neural Information Processing Systems*, pp. 9896–9905 (2018)
42. Xu, Y.: Primal-dual stochastic gradient method for convex programs with many functional constraints. arXiv preprint [arXiv:1802.02724v1](https://arxiv.org/abs/1802.02724v1) (2018)
43. Yamashita, S., Hatanaka, T., Yamauchi, J., Fujita, M.: Passivity-based generalization of primal-dual dynamics for non-strictly convex cost functions. *Automatica* **112**, 108712 (2020)
44. Yu, H., Neely, M., Wei, X.: Online convex optimization with stochastic constraints. In: *Advances in Neural Information Processing Systems*, pp. 1428–1438 (2017)
45. Zhang, T., Uryasev, S., Guan, Y.: Derivatives and subderivatives of buffered probability of exceedance. *Oper. Res. Lett.* **47**(2), 130–132 (2019)
46. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 928–936 (2003)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.