Noname manuscript No.
(will be inserted by the editor)

Primal-Dual Stochastic Hybrid Approximation Algorithm (Draft)

Tomás Tinoco De Rubira · Gabriela Hug

Received: date / Accepted: date

Abstract A new algorithm for solving convex stochastic optimization problems with expectation functions in both the objective and constraints is presented. The algorithm combines a stochastic hybrid procedure, which was originally designed to solve problems with expectation only in the objective, with dual stochastic gradient ascent. More specifically, the algorithm generates primal iterates by minimizing deterministic approximations of the Lagrangian that are updated using noisy subgradients, and dual iterates by applying stochastic gradient ascent to the true Lagrangian. The sequence of primal iterates produced by the algorithm is shown to have a subsequence that converges almost surely to an optimal point under certain conditions. Numerical experience with the new and benchmark algorithms that include a primal-dual stochastic approximation algorithm and algorithms based on sample-average approximations is reported. The test problem used originates in power systems operations planning under high penetration of renewable energy, where optimal risk-averse power generation policies are sought. In particular, the problem consists of a two-stage stochastic optimization problem with quadratic objectives and a Conditional Value-At-Risk constraint.

Keywords Stochastic Approximation · Stochastic Hybrid Approximation · Expected-Value Constraints · Conditional Value-At-Risk

1 Introduction

In many applications, optimization problems arise whose objective function or constraints have expectation functions. For example, in portfolio optimization, one is in-

T. Tinoco De Rubira
Power Systems Laboratory, ETH Zurich, Switzerland
Tel.: +41-44-632-43-84 E-mail: tomast@eeh.ee.ethz.ch

G. Hug
Power Systems Laboratory, ETH Zurich, Switzerland
Tel.: +41-44-632-81-91 E-mail: ghug@eeh.ee.ethz.ch

terested in selecting the portfolio that results in the lowest expected loss, and a bound on the risk of high losses may be enforced using Conditional Value-At-Risk (CVaR) [1]. In power systems operations planning, an important task is planning generator output powers in such a way that the expected system operation cost is minimized in the presence of uncertain inputs from renewable energy sources [2]. In communication networks, one goal is to maximize throughput while bounding the expected packet loss or time delay in transmission [3]. In call centers, one is interested in minimizing cost while ensuring that the expected number of answered calls are above a certain level [4]. Solving optimization problems such as these that have expectation functions is in general a challenging task. The reason is that the multi-dimensional integrals that define the expectation functions cannot be computed with high accuracy in practice when the uncertainty has more than a few dimensions [5]. Hence, existing approaches used for solving these problems work with approximations instead of the exact expectation functions.

Two families of approaches have been studied for trying to solve optimization problems with expectation functions in a computationally tractable way. The first approach is based on *external sampling*, which consists of sampling realizations of the uncertainty and replacing the stochastic problem with a deterministic approximation of it. This approach is also referred to as the Sample-Average Approximation (SAA) approach since the deterministic approximation is obtained by replacing expected values with sample averages [6]. The resulting deterministic problem, which is typically very large, is solved using techniques that exploit the problem structure, such as the Benders or L-shaped decomposition method [7] [8]. Several authors have analyzed the performance of SAA-based algorithms in the literature. The reader is referred to [1], [6], [9] and [10] for discussions on asymptotic convergence of solutions and optimal values, and on the accuracy obtained using finite sample sizes.

The second approach for solving optimization problems with expectation functions is based on *internal sampling*, which consists of periodically sampling realizations of the uncertainty during the execution of the algorithm. Some of the most studied algorithms of this family are Stochastic Approximation (SA) algorithms. These algorithms update a solution estimate by taking steps along random directions. These random directions typically have the property that on expectation are equal to negative gradients or subgradients of the function to be minimized [6]. The advantage of these algorithms is that they typically have low demand for computer memory. Their convergence, however, is very sensitive to the choice of step lengths, *i.e.*, the lengths of the steps taken along the random directions. The fundamental issue is that the step lengths need to be small enough to adequately suppress noise, while at the same time large enough to avoid slow convergence. Since first proposed by Robbins and Monro [11], SA algorithms have been an active area of research. Most of the work has been done in the context of stochastic optimization problems with deterministic constraints. The reader is referred to [5] and [12] for important advancements and insights on these algorithms. To handle expectation functions in the constraints, to the best of our knowledge, SA approaches explored in the literature have been based on the Lagrangian method [3] [13] [14] [15]. More specifically, stochastic gradient or similar steps are used for minimizing the Lagrangian with respect to the primal variables and for maximizing the Lagrangian with respect to the dual variables.

Another important internal-sampling algorithm, which is referred to as stochastic hybrid approximation algorithm, is proposed in [16] for solving stochastic problems with deterministic constraints. It consists of solving a sequence of deterministic approximations of the original problem whose objective functions are periodically updated using noisy subgradients. Its development was motivated by a dynamic resource allocation problem, and it belongs to an important line of research that uses techniques from approximate dynamic programming for solving problems of such type [17] [18] [19]. The authors in [16] show that the iterates produced by the algorithm convergence almost surely to an optimal point under certain conditions. In [2], the authors apply this algorithm to a two-stage stochastic optimization problem with quadratic objectives and linear constraints for determining optimal generation policies in power systems with high penetration of renewable energy. The initial deterministic approximation used consists of the certainty-equivalent problem, *i.e.*, the problem obtained by replacing random variables with their expected values. The results obtained show the superiority of the hybrid algorithm over benchmark algorithms that include stochastic gradient descent and (sequential) algorithms based on SAA with Benders decomposition on the problem considered. In particular, the hybrid algorithm is shown to produce better iterates than the other algorithms, and to be more robust against noise during the initial iterations compared to stochastic gradient descent. The authors attribute these properties to the quality of the initial deterministic approximation used, and to an interesting connection between the hybrid algorithm and a stochastic proximal gradient algorithm when applied to the problem considered.

In this work, the stochastic hybrid approximation algorithm proposed in [16] and further investigated in [2] is extended to also handle constraints with expectation functions. This is done by applying the hybrid procedure to the Lagrangian and combining it with dual stochastic gradient ascent. An analysis of the convergence of this new algorithm is presented. Furthermore, numerical experience obtained from applying this and benchmark algorithms to a risk-averse version of the problem considered in [2] is reported. The benchmark algorithms include a primal-dual SA algorithm and SAA-based algorithms with and without decomposition techniques. The test problem considered consists of a two-stage stochastic problem with quadratic objectives and a CVaR constraint that bounds the risk of high second-stage costs. The instances of the test problem used for the experiments come from real electric power networks having around 2500 and 3000 nodes and hence several thousand combined first- and second-stage variables. The empirical results obtained suggest that the hybrid algorithm can exploit accurate initial deterministic approximations of the expectation functions and converge to a solution faster than the benchmark algorithms.

This paper is organized as follows: In Section 2, the general stochastic optimization problem considered as well as important notation are introduced. In Sections 3 and 4, the SAA-based and primal-dual SA benchmark algorithms are described. In Section 5, the new primal-dual stochastic hybrid approximation algorithm is presented along with a convergence analysis and simple examples that illustrate its properties. In Section 6, an application from power systems operations planning of stochastic optimization involving expectation functions in both the objective and constraints is described, and the numerical experience obtained from applying the

algorithms considered in this work to this application is reported. Lastly, in Section 7, the key contributions and findings of this work are summarized, and next research directions are discussed.

2 General Problem

The general stochastic optimization problem considered is of the form

$$\begin{aligned} \underset{x\in\mathscr{X}}{\text{minimize}} \quad & \mathscr{F}(x) \\ \text{subject to} \quad & \mathscr{G}(x) \leq 0, \end{aligned} \tag{1}$$

where $\mathscr{X}$ is a compact convex set, $\mathscr{F} := \mathbb{E}[F(\cdot, w)]$, $F(\cdot, w)$ is convex for each random vector $w \in \Omega$, $\mathscr{G} := \mathbb{E}[G(\cdot, w)]$, $G(\cdot, w)$ is a vector-valued function composed of convex functions for each $w \in \Omega$, Ω is a compact set, and $\mathbb{E}[\cdot]$ denotes expectation with respect to w. It is assumed that all functions are continuous in $\mathscr{X}$, and that Slater's condition and hence strong duality holds [20].

The Lagrangian of problem (1) is given by

$$\mathscr{L}(x, \lambda) := \mathscr{F}(x) + \lambda^T \mathscr{G}(x),$$

and the "noisy" Lagrangian is defined as

$$L(x, \lambda, w) := F(x, w) + \lambda^T G(x, w).$$

The vector x^* denotes a primal optimal point of problem (1). Similarly, the vector λ^* denotes a dual optimal point of problem (1). It is assumed that $\lambda^* \in \Lambda$, where

$$\Lambda := \{\, \lambda \mid 0 \leq \lambda < 1\lambda^{\max} \,\},$$

1 is the vector of ones, and $\lambda^{\max}$ is some known positive scalar.

In the following sections, unless otherwise stated, all derivatives and subdifferentials are assumed to be with respect to the primal variables x.

3 Algorithms Based on Sample-Average Approximations

A widely-used *external-sampling* approach for solving stochastic optimization problems of the form of (1) is to sample independent realizations of the random vector w, say $\{w_l\}_{l=1}^N$, where $N \in \mathbb{Z}_{++}$, and replace expected values with samples averages. The resulting deterministic problem approximation is

$$\begin{aligned} \underset{x\in\mathscr{X}}{\text{minimize}} \quad & N^{-1} \sum_{l=1}^{N} F(x, w_l) \\ \text{subject to} \quad & N^{-1} \sum_{l=1}^{N} G(x, w_l) \leq 0. \end{aligned} \tag{2}$$

It can be shown that under certain conditions and sufficiently large N the solutions of (2) are arbitrarily close to those of (1) with high probability [1] [6].

Due to the requirement of choosing N sufficiently large in order to get an accurate estimate of a solution of (1), problem (2) can be difficult to solve with standard (deterministic) optimization algorithms. This is particularly true when $F(\cdot,w)$ or $G(\cdot,w)$ are computationally expensive to evaluate, *e.g.*, when they are defined in terms of optimal values of other optimization problems that depend on both x and w. In this case, a common strategy for solving (2) is to use decomposition. One widely-used decomposition approach is based on cutting planes and consists of first expressing (2) as

$$\begin{aligned} \underset{x\in\mathscr{X}}{\text{minimize}} \quad & \bar{F}(x)+N^{-1}\sum_{l=1}^{N}\hat{F}(x,w_l) \\ \text{subject to} \quad & \bar{G}(x)+N^{-1}\sum_{l=1}^{N}\hat{G}(x,w_l)\leq 0, \end{aligned}$$

and then performing the following step at each iteration $k\in\mathbb{Z}_+$:

$$x_k=\arg\min\ \Big\{\bar{F}(x)+\max_{0\leq j<k} a_j+b_j^T(x-x_j)\ \Big|\ x\in\mathscr{X},\ \bar{G}_i(x)+\max_{0\leq j<k} c_{ij}+d_{ij}^T(x-x_j)\leq 0,\ \forall i\Big\}, \quad (3)$$

where

$$\begin{aligned} a_j &= N^{-1}\sum_{l=1}^{N}\hat{F}(x_j,w_l), \qquad & b_j &\in N^{-1}\sum_{l=1}^{N}\partial\hat{F}(x_j,w_l), \\ c_{ij} &= N^{-1}\sum_{l=1}^{N}\hat{G}_i(x_j,w_l), \qquad & d_{ij} &\in N^{-1}\sum_{l=1}^{N}\partial\hat{G}_i(x_j,w_l), \end{aligned}$$

and $\bar{G}_i(x,w)$ and $\hat{G}_i(x,w)$ denote the i-th component of $\bar{G}(x,w)$ and $\hat{G}(x,w)$, respectively. Some key properties of this algorithm are that the functions

$$\max_{0\leq j<k} a_j+b_j^T(x-x_j) \quad \text{and} \quad \max_{0\leq j<k} c_{ij}+d_{ij}^T(x-x_j)$$

are gradually-improving piecewise linear under-estimators of $N^{-1}\sum_{l=1}^{N}\hat{F}(\cdot,w_l)$ and $N^{-1}\sum_{l=1}^{N}\hat{G}_i(\cdot,w_l)$, respectively, and that parallelization can be easily exploited for computing a_j, b_j, c_{ij} and d_{ij} efficiently. The interested reader is referred to [7] and [8] for more details about decomposition techniques based on cutting planes.

4 Primal-Dual Stochastic Approximation Algorithm

As already mentioned, the authors in [13] and [14] describe an *internal-sampling* primal-dual stochastic approximation algorithm for solving problems of the form of

(1). It consists of performing the following steps at each iteration $k \in \mathbb{Z}_+$:

$$x_{k+1} = \Pi_x(x_k - \alpha_k \xi_k) \tag{4a}$$

$$\lambda_{k+1} = \Pi_\lambda\left(\lambda_k + \alpha_k G(x_k, w_k)\right), \tag{4b}$$

where $\xi_k \in \partial L(x_k, \lambda_k, w_k)$, α_k are scalar step lengths, Π_x is projection on $\mathscr{X}$, Π_λ is projection on Λ, and w_k are independent identically distributed (i.i.d.) samples of the random vector w. Since ξ_k is a noisy subgradient of $\mathscr{L}(\cdot, \lambda_k)$ at x_k, and $G(x_k, w_k)$ is a noisy gradient of $\mathscr{L}(x_k, \cdot)$ at λ_k, *i.e.*,

$$\mathbb{E}[\xi_k | \mathscr{W}_{k-1}] \in \partial \mathscr{L}(x_k, \lambda_k) \quad \text{and} \quad \mathbb{E}[G(x_k, w_k) | \mathscr{W}_{k-1}] = \nabla_\lambda \mathscr{L}(x_k, \lambda_k),$$

where $\mathscr{W}_k$ denotes the observation history $(w_0, \ldots, w_k)$, it is clear that this algorithm performs stochastic subgradient descent to minimize the Lagrangian with respect to x, and stochastic gradient ascent to maximize the Lagrangian with respect to λ.

In [13], the authors provide a convergence proof of this algorithm under certain conditions based on the "ODE method". These conditions include, among others, $\mathscr{F}$ being strictly convex and continuously differentiable, the components of $\mathscr{G}$ being convex and continuously differentiable, and the existence of a point $\tilde{x}$ such that $\mathscr{G}(\tilde{x}) < 0$.

5 Primal-Dual Stochastic Hybrid Approximation Algorithm

In [16], the authors propose an *internal-sampling* stochastic hybrid approximation procedure for solving stochastic optimization problems of the form

$$\underset{x \in \mathscr{X}}{\text{minimize}} \quad \mathbb{E}[F(x, w)], \tag{5}$$

where $\mathscr{X}$ and $F(\cdot, w)$ have the same properties as those of problem (1). The procedure consists of generating iterates

$$x_k = \arg\min \{ \mathscr{F}_k(x) \mid x \in \mathscr{X} \}$$

at each iteration $k \in \mathbb{Z}_+$, where $\mathscr{F}_k$ are strongly convex deterministic differentiable approximations of $\mathbb{E}[F(\cdot, w)]$. The first approximation $\mathscr{F}_0$ is provided by the user, and the rest are obtained by performing "noisy" slope corrections at each iteration using

$$\mathscr{F}_{k+1}(x) = \mathscr{F}_k(x) + \alpha_k(\hat{g}_k - \nabla \mathscr{F}_k(x_k))^T x,$$

where $\hat{g}_k \in \partial F(x_k, w_k)$, w_k are i.i.d samples of the random vector w, and α_k are step lengths that satisfy conditions that are common in stochastic approximation algorithms [14] [21]. The slope corrections performed at each iteration are "noisy" because they are based on subgradients associated with specific realizations of the random vector. Roughly speaking, the conditions satisfied by the step lengths ensure that this noise is eventually averaged out. The authors show that the iterates x_k convergence almost surely to an optimal point of (5) under the stated and other minor assumptions. They claim that the strengths of this approach are its ability to exploit

an accurate initial approximation $\mathscr{F}_0$, and the fact that the slope corrections do not change the structure of the approximations.

In [2], the authors apply this stochastic hybrid approximation procedure to a two-stage stochastic generator dispatch problem that arises in power systems with high penetration of renewable energy. Motivated by the fact that in power systems operations planning certainty-equivalent models are common, the authors use the function $\mathscr{F}_0 = F(\cdot, \mathbb{E}[w])$ as the initial approximation. The results obtained show how the algorithm outperforms common SAA-based algorithms on the problem considered due to its ability to exploit the accuracy of the given initial function approximation. The algorithm is also shown to be less susceptible to noise compared to stochastic gradient descent, and this is attributed to an interesting connection between the algorithm and a stochastic proximal gradient algorithm when applied to the problem considered.

In order to apply the stochastic hybrid approximation procedure to solve problem (1), which also has expectation functions in the constraints as opposed to only in the objective, a primal-dual extension is proposed here. The proposed algorithm works with deterministic approximations of the Lagrangian, and combines the stochastic hybrid approximation procedure with dual stochastic gradient ascent, the latter which is motivated from algorithm (4). More specifically, the proposed algorithm consists of performing the following steps at each iteration $k \in \mathbb{Z}_+$:

$$x_k = \arg\min \{ \mathscr{F}_k(x) + \lambda_k^T \mathscr{G}_k(x) \mid x \in \mathscr{X} \} \tag{6a}$$

$$\lambda_{k+1} = \Pi_\lambda \left(\lambda_k + \alpha_k \hat{G}_k \right) \tag{6b}$$

$$\mathscr{F}_{k+1}(x) = \mathscr{F}_k(x) + \alpha_k (\hat{g}_k - g_k(x_k))^T x \tag{6c}$$

$$\mathscr{G}_{k+1}(x) = \mathscr{G}_k(x) + \alpha_k \left(\hat{J}_k - J_k(x_k) \right) x, \tag{6d}$$

where $\mathscr{F}_k$ and $\mathscr{G}_k$ are deterministic differentiable approximations of the stochastic functions $\mathscr{F}$ and $\mathscr{G}$, respectively, $\hat{G}_k := G(x_k, w_k)$, $\hat{g}_k \in \partial F(x_k, w_k)$, $g_k := \nabla \mathscr{F}_k$, $\hat{J}_k$ is a matrix whose rows are subgradients (transposed) of the components of $G(\cdot, w_k)$ at x_k, $J_k := \frac{\partial}{\partial x} \mathscr{G}_k$, α_k are scalar step lengths, and w_k are i.i.d. samples of the random vector w. Step (6a) generates primal iterates by minimizing deterministic approximations of the Lagrangian. Step (6b) updates dual iterates using noisy measurements of feasibility, namely, $\hat{G}_k = G(x_k, w_k)$. Lastly, steps (6c) and (6d) perform slope corrections on the deterministic approximations $\mathscr{F}_k$ and $\mathscr{G}_k$, respectively, using noisy subgradients of the stochastic function $\mathscr{F}$ and of the components of $\mathscr{G}$.

5.1 Convergence

Almost-sure convergence of a subsequence of the primal iterates produced by algorithm (6) to an optimal point of problem (1) is shown under certain assumptions. The assumptions include the problem-specific ones stated in Section 2 as well as the additional ones stated below. The proof uses approximate primal and dual solutions x_k^σ

and λ_k^σ, respectively, of problem (1) that solve the regularized problem

$$\begin{array}{ll} \underset{x\in\mathscr{X}}{\text{minimize}} & \mathscr{F}(x)+\dfrac{\sigma}{2}\|x-x_k\|_2^2 \\ \text{subject to} & \mathscr{G}(x)\leq 0 \end{array} \tag{7}$$

for $\sigma>0$ and $k\in\mathbb{Z}_+$. The proof also uses the approximate Lagrangian defined by

$$\mathscr{L}_k(x,\lambda):=\mathscr{F}_k(x)+\lambda^T\mathscr{G}_k(x).$$

Assumption 1 (Initial Function Approximations) *The initial objective function approximation $\mathscr{F}_0$ is strongly convex and continuously differentiable in $\mathscr{X}$. The components of the initial constraint function approximation $\mathscr{G}_0$ are convex and continuously differentiable in $\mathscr{X}$.*

Assumption 2 (Sampled Subgradients) *The subgradients $\hat{g}_k$ and matrices of subgradients $\hat{J}_k$ defined above are uniformly bounded for all $k\in\mathbb{Z}_+$.*

Assumption 3 (Step Lengths) *The step lengths α_k lie inside the open interval $(0,1)$, and satisfy $\sum_{k=0}^{\infty}\alpha_k=\infty$ almost surely and $\sum_{k=0}^{\infty}\mathbb{E}\left[\alpha_k^2\right]<\infty$.*

Assumption 4 (Approximate Dual Solutions) *For each $\sigma>0$ there exists a constant $M>0$ such that the approximate dual solutions λ_k^σ satisfy*

$$\|\lambda_{k+1}^\sigma-\lambda_k^\sigma\|_2\leq M\big(\|x_{k+1}-x_k\|_2+\|x_{k+1}^\sigma-x_k^\sigma\|_2\big)$$

for all $k\in\mathbb{Z}_+$. Furthermore, for all $\sigma>0$ small enough, $\lambda_k^\sigma\in\Lambda$ for all $k\in\mathbb{Z}_+$.

Assumption 5 (Drift) *For all $\sigma>0$ small enough, the sequences of partial sums*

$$\sum_{k=0}^{K}(\lambda_{k+1}-\lambda_k)^T\Gamma_k,\quad \sum_{k=0}^{K}(\lambda_{k+1}^\sigma-\lambda_k^\sigma)^T\Psi_k,\quad \sum_{k=0}^{K}(x_{k+1}^\sigma-x_k^\sigma)^T\Upsilon_k, \tag{8}$$

where $\Gamma_k:=\mathscr{G}_k(x_k^\sigma)-\mathscr{G}_k(x_k)$, $\Psi_k:=\lambda_k^\sigma-\lambda_k$ and $\Upsilon_k:=\nabla\mathscr{L}_k(x_k^\sigma,\lambda_k)$, are almost-surely bounded above.

Assumptions 1, 2, and 3 made on the initial function approximations, subgradients, and step lengths, respectively, follow those made in [16]. Assumption 4 is reasonable from (7), the definition of Λ, and $\lambda^*\in\Lambda$. Assumption 5 is more obscure and hence Section 5.2 is devoted to providing an intuitive justification for why this may hold in practice. Furthermore, the examples in Section 5.3 show the behavior of the partial sums in (8) when using initial function approximations and dual bounds $\lambda^{\max}$ of various quality.

Lemma 1 *The functions $\mathscr{F}_k$ and $\mathscr{G}_k$ and their derivatives are uniformly Lipschitz for all $k\in\mathbb{Z}_+$ and hence also uniformly bounded in $\mathscr{X}$.*

Proof From (6d), the function $\mathscr{G}_k$ can be expressed as

$$\mathscr{G}_k(x) = \mathscr{G}_0(x) + R_k x, \ \forall x \in \mathscr{X}, \tag{9}$$

for $k \in \mathbb{Z}_+$, where R_k is a matrix and $R_0 = 0$. It follows that for all $k \in \mathbb{Z}_+$ and $x \in \mathscr{X}$,

$$\begin{aligned}\mathscr{G}_{k+1}(x) &= \mathscr{G}_k(x) + \alpha_k\big(\hat{J}_k - J_k(x_k)\big)x \\ &= \mathscr{G}_0(x) + R_k x + \alpha_k\big(\hat{J}_k - J_0(x_k) - R_k\big)x \\ &= \mathscr{G}_0(x) + \big((1-\alpha_k)R_k + \alpha_k\big(\hat{J}_k - J_0(x_k)\big)\big)x.\end{aligned}$$

Hence

$$R_{k+1} = (1-\alpha_k)R_k + \alpha_k\big(\hat{J}_k - J_0(x_k)\big).$$

From Assumptions 1 and 2, there is some constant M_1 such that

$$\|\hat{J}_k - J_0(x_k)\|_2 \le M_1$$

for all $k \in \mathbb{Z}_+$. Hence, Assumption 3 gives that $\|R_k\|_2 \le M_1$ implies

$$\begin{aligned}\|R_{k+1}\|_2 &= \big\|(1-\alpha_k)R_k + \alpha_k\big(\hat{J}_k - J_0(x_k)\big)\big\|_2 \\ &\le (1-\alpha_k)\|R_k\|_2 + \alpha_k\|\hat{J}_k - J_0(x_k)\|_2 \\ &\le (1-\alpha_k)M_1 + \alpha_k M_1 \\ &= M_1.\end{aligned}$$

Since $R_0 = 0$, it holds by induction that $\|R_k\|_2 \le M_1$ for all $k \in \mathbb{Z}_+$. It follows that for any x and $y \in \mathscr{X}$ and $k \in \mathbb{Z}_+$,

$$\begin{aligned}\|\mathscr{G}_k(x) - \mathscr{G}_k(y)\|_2 &= \|\mathscr{G}_0(x) + R_k x - \mathscr{G}_0(y) - R_k y\|_2 \\ &\le \|R_k\|_2\|x-y\|_2 + \|\mathscr{G}_0(x) - \mathscr{G}_0(y)\|_2 \\ &\le (M_1 + M_2)\,\|x-y\|_2,\end{aligned}$$

where $M_2 > 0$ exists because Assumption 1 implies that $\mathscr{G}_0$ is Lipschitz in the compact set $\mathscr{X}$.

From (9), the Jacobian of $\mathscr{G}_k$ is given by

$$J_k(x) = J_0(x) + R_k, \ \forall x \in \mathscr{X},$$

for each $k \in \mathbb{Z}_+$. Assumption 1 and $\|R_k\|_2 \le M_1$ imply that the functions J_k, $k \in \mathbb{Z}_+$, are uniformly Lipschitz in $\mathscr{X}$.

The derivation of these properties for the functions $\mathscr{F}_k$, $k \in \mathbb{Z}_+$, is similar and hence omitted. □

Lemma 2 *The functions $\mathscr{F}_k$, $k \in \mathbb{Z}_+$, are uniformly strongly convex.*

Proof From (6c) and Lemma 1, $\mathscr{F}_k$ can be expressed as

$$\mathscr{F}_k(x) = \mathscr{F}_0(x) + r_k^T x, \ \forall x \in \mathscr{X},$$

for $k \in \mathbb{Z}_+$, where r_k are uniformly bounded vectors with $r_0 = 0$. Assumption 1 gives that $\mathscr{F}_0$ is strongly convex, and therefore that $\mathscr{F}_k$ are uniformly strongly convex since their strong-convexity constant is independent of k. □

Lemma 3 *There exists an $M > 0$ such that $\|x_{k+1} - x_k\|_2 \leq \alpha_k M$ for all $k \in \mathbb{Z}_+$. For convenience, we say that $\|x_{k+1} - x_k\|_2$ is $\mathcal{O}(\alpha_k)$.*

Proof Equation (6a) implies that

$$\left(g_k(x_k) + J_k(x_k)^T \lambda_k\right)^T (x - x_k) \geq 0, \ \forall x \in \mathscr{X} \tag{10}$$

$$\left(g_{k+1}(x_{k+1}) + J_{k+1}(x_{k+1})^T \lambda_{k+1}\right)^T (x - x_{k+1}) \geq 0, \ \forall x \in \mathscr{X}, \tag{11}$$

for all $k \in \mathbb{Z}_+$. In particular, (10) holds with $x = x_{k+1}$, and (11) holds with $x = x_k$. Hence,

$$(g_{k+1}(x_{k+1}) - g_k(x_k))^T \delta_k \leq \left(J_k(x_k)^T \lambda_k - J_{k+1}(x_{k+1})^T \lambda_{k+1}\right)^T \delta_k,$$

where $\delta_k := x_{k+1} - x_k$. Using (6c) and (6d) gives

$$(g_k(x_{k+1}) - g_k(x_k))^T \delta_k \leq \left(J_k(x_k)^T \lambda_k - J_k(x_{k+1})^T \lambda_{k+1}\right)^T \delta_k - \alpha_k \left(\Delta_k^T \lambda_{k+1} + \eta_k\right)^T \delta_k, \tag{12}$$

where

$$\eta_k := \hat{g}_k - g_k(x_k) \tag{13}$$

$$\Delta_k := \hat{J}_k - J_k(x_k). \tag{14}$$

From Lemma 2, $\mathscr{F}_k$ is uniformly strongly convex so there exists a $C > 0$ independent of k such that

$$C\|\delta_k\|_2^2 \leq (g_k(x_{k+1}) - g_k(x_k))^T \delta_k$$

for all $k \in \mathbb{Z}_+$. It follows from this and (12) that

$$\begin{aligned} C\|\delta_k\|_2^2 &\leq \left(\lambda_k^T J_k(x_k) - \lambda_{k+1}^T J_k(x_{k+1})\right) \delta_k - \alpha_k \left(\Delta_k^T \lambda_{k+1} + \eta_k\right)^T \delta_k \\ &\leq \left(\lambda_k^T J_k(x_k) - \lambda_{k+1}^T J_k(x_{k+1})\right) \delta_k + \alpha_k M_1 \|\delta_k\|_2 \end{aligned} \tag{15}$$

for all $k \in \mathbb{Z}_+$, where $M_1 > 0$ exists because of the uniform boundedness of λ_k (by construction), and that of Δ_k and η_k (by Assumption 2 and Lemma 1). From the convexity of the component functions of $\mathscr{G}_k$ and the fact that $\lambda_k \geq 0$ and $\lambda_{k+1} \geq 0$, it holds that

$$\begin{aligned} \lambda_{k+1}^T (\mathscr{G}_k(x_k) - \mathscr{G}_k(x_{k+1})) &\geq -\lambda_{k+1}^T J_k(x_{k+1}) \delta_k \\ \lambda_k^T (\mathscr{G}_k(x_{k+1}) - \mathscr{G}_k(x_k)) &\geq \lambda_k^T J_k(x_k) \delta_k. \end{aligned}$$

Adding these inequalities gives

$$\begin{aligned} \left(\lambda_k^T J_k(x_k) - \lambda_{k+1}^T J_k(x_{k+1})\right) \delta_k &\leq (\lambda_{k+1} - \lambda_k)^T (\mathscr{G}_k(x_k) - \mathscr{G}_k(x_{k+1})) \\ &\leq \|\lambda_{k+1} - \lambda_k\|_2 M_2 \|\delta_k\|_2 \\ &\leq \alpha_k \|\hat{G}_k\|_2 M_2 \|\delta_k\|_2 \\ &\leq \alpha_k M_2 M_3 \|\delta_k\|_2, \end{aligned}$$

where $M_2 > 0$ and $M_3 > 0$ exist due to the uniform boundedness of $\hat{G}_k$ and Lemma 1. It follows from this and (15) that

$$C\|\delta_k\|_2^2 \leq \alpha_k(M_1 + M_2M_3)\|\delta_k\|_2,$$

or equivalently,

$$\|\delta_k\|_2 \leq \alpha_k(M_1 + M_2M_3)/C$$

for all $k \in \mathbb{Z}_+$. □

Corollary 1 *For each $\sigma > 0$, $\|x_{k+1}^\sigma - x_k^\sigma\|_2$ and $\|\lambda_{k+1}^\sigma - \lambda_k^\sigma\|_2$ are $\mathscr{O}(\alpha_k)$.*

Proof For each $\sigma > 0$ and $k \in \mathbb{Z}_+$, the optimality of x_k^σ implies that there exists a $g_k^\sigma \in \partial\mathscr{F}(x_k^\sigma)$ such that

$$\left(g_k^\sigma + \sigma(x_k^\sigma - x_k)\right)^T(x - x_k^\sigma) \geq 0$$

for all $x \in \mathscr{X}$ such that $\mathscr{G}(x) \leq 0$. It follows from this that

$$\begin{aligned}\left(g_k^\sigma + \sigma(x_k^\sigma - x_k)\right)^T(x_{k+1}^\sigma - x_k^\sigma) &\geq 0\\ \left(g_{k+1}^\sigma + \sigma(x_{k+1}^\sigma - x_{k+1})\right)^T(x_k^\sigma - x_{k+1}^\sigma) &\geq 0.\end{aligned}$$

Adding these inequalities, rearranging, and using $(g_{k+1}^\sigma - g_k^\sigma)^T(x_{k+1}^\sigma - x_k^\sigma) \geq 0$ gives

$$\begin{aligned}\sigma\|x_{k+1}^\sigma - x_k^\sigma\|_2^2 &\leq \sigma(x_{k+1} - x_k)^T(x_{k+1}^\sigma - x_k^\sigma) + (g_k^\sigma - g_{k+1}^\sigma)^T(x_{k+1}^\sigma - x_k^\sigma)\\ &\leq \sigma\|x_{k+1} - x_k\|_2\|x_{k+1}^\sigma - x_k^\sigma\|_2.\end{aligned}$$

It follows from this and Lemma 3 that $\|x_{k+1}^\sigma - x_k^\sigma\|_2$ is $\mathscr{O}(\alpha_k)$. Assumption 4 then gives that $\|\lambda_{k+1}^\sigma - \lambda_k^\sigma\|_2$ is also $\mathscr{O}(\alpha_k)$. □

Lemma 4 *For all $\sigma > 0$ small enough, there exists an $M > 0$ such that*

$$\frac{1}{2}\|\lambda_{k+1} - \lambda_{k+1}^\sigma\|_2^2 \leq \frac{1}{2}\|\lambda_k - \lambda_k^\sigma\|_2^2 + (\lambda_{k+1}^\sigma - \lambda_k^\sigma)^T\Psi_k + \alpha_k(\lambda_k - \lambda_k^\sigma)^T\hat{G}_k + \alpha_k^2M$$

for all $k \in \mathbb{Z}_+$, where Ψ_k is as defined in Assumption 5.

Proof From equation (6b), the fact that $\lambda_k^\sigma \in \Lambda$ for small enough σ, Corollary 1, and the uniform boundedness of $\hat{G}_k$, it follows that there exist positive scalars M_1 and M_2 independent of k such that

$$\begin{aligned}\frac{1}{2}\|\lambda_{k+1} - \lambda_{k+1}^\sigma\|_2^2 &= \frac{1}{2}\left\|\Pi_\lambda\left(\lambda_k + \alpha_k\hat{G}_k\right) - \lambda_{k+1}^\sigma\right\|_2^2\\ &\leq \frac{1}{2}\|\lambda_k + \alpha_k\hat{G}_k - \lambda_{k+1}^\sigma\|_2^2\\ &= \frac{1}{2}\|\lambda_k - \lambda_{k+1}^\sigma\|_2^2 + \alpha_k(\lambda_k - \lambda_{k+1}^\sigma)^T\hat{G}_k + \frac{1}{2}\|\alpha_k\hat{G}_k\|_2^2\\ &\leq \frac{1}{2}\|\lambda_k - \lambda_{k+1}^\sigma\|_2^2 + \alpha_k(\lambda_k - \lambda_k^\sigma)^T\hat{G}_k + \alpha_k^2M_1\\ &\leq \frac{1}{2}\|\lambda_k - \lambda_k^\sigma\|_2^2 + (\lambda_{k+1}^\sigma - \lambda_k^\sigma)^T\Psi_k + \alpha_k(\lambda_k - \lambda_k^\sigma)^T\hat{G}_k + \alpha_k^2M_2.\end{aligned}$$

□

Lemma 5 *For all $\sigma > 0$ small enough, the sequence $\{T_k^\sigma\}_{k\in\mathbb{Z}_+}$ defined by*

$$T_k^\sigma := \mathscr{F}_k(x_k^\sigma) + \lambda_k^T \mathscr{G}_k(x_k^\sigma) - \mathscr{F}_k(x_k) - \lambda_k^T \mathscr{G}_k(x_k) + \frac{1}{2}\|\lambda_k - \lambda_k^\sigma\|_2^2$$

satisfies

$$\begin{aligned} T_{k+1}^\sigma - T_k^\sigma \leq\ & (\lambda_{k+1} - \lambda_k)^T \Gamma_k + (\lambda_{k+1}^\sigma - \lambda_k^\sigma)^T \Psi_k + (x_{k+1}^\sigma - x_k^\sigma)^T \Upsilon_k + \\ & \alpha_k \left(L(x_k^\sigma, \lambda_k^\sigma, w_k) - L(x_k, \lambda_k^\sigma, w_k) + \frac{\sigma}{2}\|x_k^\sigma - x_k\|_2^2 \right) + \\ & \alpha_k (\lambda_k - \lambda_k^\sigma)^T G(x_k^\sigma, w_k) + \\ & \mathscr{O}(\alpha_k^2) \end{aligned}$$

for all $k \in \mathbb{Z}_+$, where Γ_k, Ψ_k and Υ_k are as defined in Assumption 5.

Proof From the definition of T_k^σ, it follows that

$$\begin{aligned} T_{k+1}^\sigma - T_k^\sigma = & \mathscr{F}_{k+1}(x_{k+1}^\sigma) + \lambda_{k+1}^T \mathscr{G}_{k+1}(x_{k+1}^\sigma) - \mathscr{F}_{k+1}(x_{k+1}) - \lambda_{k+1}^T \mathscr{G}_{k+1}(x_{k+1}) - \\ & \mathscr{F}_k(x_k^\sigma) - \lambda_k^T \mathscr{G}_k(x_k^\sigma) + \mathscr{F}_k(x_k) + \lambda_k^T \mathscr{G}_k(x_k) + \\ & \frac{1}{2}\|\lambda_{k+1} - \lambda_{k+1}^\sigma\|_2^2 - \frac{1}{2}\|\lambda_k - \lambda_k^\sigma\|_2^2. \end{aligned}$$

Using equations (6c) and (6d), and adding and subtracting $\lambda_k^T \mathscr{G}_k(x_{k+1})$ as well as $\lambda_k^T \mathscr{G}_k(x_{k+1}^\sigma)$ to the right-had side of the above equation gives

$$\begin{aligned} T_{k+1}^\sigma - T_k^\sigma = & \mathscr{F}_k(x_k) + \lambda_k^T \mathscr{G}_k(x_k) - \mathscr{F}_k(x_{k+1}) - \lambda_k^T \mathscr{G}_k(x_{k+1}) + \\ & \mathscr{F}_k(x_{k+1}^\sigma) + \lambda_k^T \mathscr{G}_k(x_{k+1}^\sigma) - \mathscr{F}_k(x_k^\sigma) - \lambda_k^T \mathscr{G}_k(x_k^\sigma) + \\ & (\lambda_{k+1} - \lambda_k)^T (\mathscr{G}_k(x_{k+1}^\sigma) - \mathscr{G}_k(x_{k+1})) + \\ & \alpha_k \eta_k^T (x_{k+1}^\sigma - x_{k+1}) + \\ & \alpha_k \lambda_{k+1}^T \Delta_k (x_{k+1}^\sigma - x_{k+1}) + \\ & \frac{1}{2}\|\lambda_{k+1} - \lambda_{k+1}^\sigma\|_2^2 - \frac{1}{2}\|\lambda_k - \lambda_k^\sigma\|_2^2, \end{aligned}$$

where η_k and Δ_k are defined in (13) and (14), respectively. From (6a), it holds that

$$\mathscr{F}_k(x_k) + \lambda_k^T \mathscr{G}_k(x_k) - \mathscr{F}_k(x_{k+1}) - \lambda_k^T \mathscr{G}_k(x_{k+1}) \leq 0.$$

Using this, Lemma 4, the convexity of $\mathscr{L}_k(\cdot, \lambda_k)$, and the definitions of η_k and Δ_k gives

$$\begin{aligned} T_{k+1}^\sigma - T_k^\sigma \leq\ & (\lambda_{k+1} - \lambda_k)^T (\mathscr{G}_k(x_{k+1}^\sigma) - \mathscr{G}_k(x_{k+1})) + \\ & (\lambda_{k+1}^\sigma - \lambda_k^\sigma)^T (\lambda_k^\sigma - \lambda_k) + \\ & (x_{k+1}^\sigma - x_k^\sigma)^T \nabla \mathscr{L}_k(x_{k+1}^\sigma, \lambda_k) + \\ & \alpha_k \left(\hat{g}_k + \hat{f}_k^T \lambda_{k+1}\right)^T (x_{k+1}^\sigma - x_{k+1}) - \\ & \alpha_k \left(g_k(x_k) + J_k(x_k)^T \lambda_{k+1}\right)^T (x_{k+1}^\sigma - x_{k+1}) + \\ & \alpha_k (\lambda_k - \lambda_k^\sigma)^T \hat{G}_k + \\ & \mathscr{O}(\alpha_k^2). \end{aligned}$$

From equation (6b) and the uniform boundedness of $\hat{G}_k$, it holds that $\|\lambda_{k+1}-\lambda_k\|_2$ is $\mathscr{O}(\alpha_k)$. From Lemma 3 and Corollary 1, it holds that $\|x_{k+1}-x_k\|_2$, $\|x^\sigma_{k+1}-x^\sigma_k\|_2$ and $\|\lambda^\sigma_{k+1}-\lambda^\sigma_k\|_2$ are $\mathscr{O}(\alpha_k)$. From these bounds, Lemma 1 and Assumption 2, it follows that

$$\begin{aligned}T^\sigma_{k+1}-T^\sigma_k \leq\ & (\lambda_{k+1}-\lambda_k)^T\Gamma_k+(\lambda^\sigma_{k+1}-\lambda^\sigma_k)^T\Psi_k+(x^\sigma_{k+1}-x^\sigma_k)^T\Upsilon_k+\\ & \alpha_k\left(\hat{g}_k+\hat{J}_k^T\lambda_k\right)^T(x^\sigma_k-x_k)-\\ & \alpha_k\left(g_k(x_k)+J_k(x_k)^T\lambda_k\right)^T(x^\sigma_k-x_k)+\\ & \alpha_k(\lambda_k-\lambda^\sigma_k)^T\hat{G}_k+\\ & \mathscr{O}(\alpha_k^2).\end{aligned}$$

where Γ_k, Ψ_k and Υ_k are as defined in Assumption 5. From equation (6a), it holds that

$$\left(g_k(x_k)+J_k(x_k)^T\lambda_k\right)^T(x^\sigma_k-x_k)\geq 0.$$

Using this and the fact that $\hat{g}_k+\hat{J}_k^T\lambda_k$ is a subgradient of $L(x,\lambda_k,w_k)+\frac{\sigma}{2}\|x-x_k\|_2^2$ at x_k gives

$$\begin{aligned}T^\sigma_{k+1}-T^\sigma_k \leq\ & (\lambda_{k+1}-\lambda_k)^T\Gamma_k+(\lambda^\sigma_{k+1}-\lambda^\sigma_k)^T\Psi_k+(x^\sigma_{k+1}-x^\sigma_k)^T\Upsilon_k+\\ & \alpha_k\left(L(x^\sigma_k,\lambda_k,w_k)-L(x_k,\lambda_k,w_k)+\frac{\sigma}{2}\|x^\sigma_k-x_k\|_2^2\right)+\\ & \alpha_k(\lambda_k-\lambda^\sigma_k)^T\hat{G}_k+\\ & \mathscr{O}(\alpha_k^2).\end{aligned}$$

Then, adding and subtracting $\alpha_k\lambda^\sigma_k G(x^\sigma_k,w_k)$, and using $\hat{G}_k=G(x_k,w_k)$ and the definition of $L(\cdot,\cdot,w_k)$ gives

$$\begin{aligned}T^\sigma_{k+1}-T^\sigma_k \leq\ & (\lambda_{k+1}-\lambda_k)^T\Gamma_k+(\lambda^\sigma_{k+1}-\lambda^\sigma_k)^T\Psi_k+(x^\sigma_{k+1}-x^\sigma_k)^T\Upsilon_k+\\ & \alpha_k\left(L(x^\sigma_k,\lambda^\sigma_k,w_k)-L(x_k,\lambda^\sigma_k,w_k)+\frac{\sigma}{2}\|x^\sigma_k-x_k\|_2^2\right)+\\ & \alpha_k(\lambda_k-\lambda^\sigma_k)^TG(x^\sigma_k,w_k)+\\ & \mathscr{O}(\alpha_k^2).\end{aligned}$$

□

Lemma 6 *For all $\sigma>0$ small enough, the sequence $\{x_k-x^\sigma_k\}_{k\in\mathbb{Z}_+}$ has a subsequence that converges to zero almost surely.*

Proof From Lemma 5, for all $\sigma>0$ small enough it holds that

$$\begin{aligned}T^\sigma_{k+1}-T^\sigma_k \leq\ & (\lambda_{k+1}-\lambda_k)^T\Gamma_k+(\lambda^\sigma_{k+1}-\lambda^\sigma_k)^T\Psi_k+(x^\sigma_{k+1}-x^\sigma_k)^T\Upsilon_k+\\ & \alpha_k\left(L(x^\sigma_k,\lambda^\sigma_k,w_k)-L(x_k,\lambda^\sigma_k,w_k)+\frac{\sigma}{2}\|x^\sigma_k-x_k\|_2^2\right)+\\ & \alpha_k(\lambda_k-\lambda^\sigma_k)^TG(x^\sigma_k,w_k)+\\ & \mathscr{O}(\alpha_k^2)\end{aligned}$$

for all $k \in \mathbb{Z}_+$. Hence, summing over k from 0 to K gives

$$\begin{aligned} T_{K+1}^{\sigma} - T_0^{\sigma} \leq & \sum_{k=0}^{K} (\lambda_{k+1} - \lambda_k)^T \Gamma_k + \sum_{k=1}^{K} (\lambda_{k+1}^{\sigma} - \lambda_k^{\sigma})^T \Psi_k + \sum_{k=1}^{K} (x_{k+1}^{\sigma} - x_k^{\sigma})^T \Upsilon_k + \\ & \sum_{k=0}^{K} \alpha_k \left(L(x_k^{\sigma}, \lambda_k^{\sigma}, w_k) - L(x_k, \lambda_k^{\sigma}, w_k) + \frac{\sigma}{2} \|x_k^{\sigma} - x_k\|_2^2 \right) + \\ & \sum_{k=0}^{K} \alpha_k (\lambda_k - \lambda_k^{\sigma})^T G(x_k^{\sigma}, w_k) + \\ & \sum_{k=0}^{K} \mathscr{O}(\alpha_k^2). \end{aligned}$$

From this, Assumptions 3 and 5, there exists an $M > 0$ independent of K such that

$$\begin{aligned} T_{K+1}^{\sigma} - T_0^{\sigma} - M \leq & \sum_{k=0}^{K} \alpha_k \left(L(x_k^{\sigma}, \lambda_k^{\sigma}, w_k) - L(x_k, \lambda_k^{\sigma}, w_k) + \frac{\sigma}{2} \|x_k^{\sigma} - x_k\|_2^2 \right) + \\ & \sum_{k=0}^{K} \alpha_k (\lambda_k - \lambda_k^{\sigma})^T G(x_k^{\sigma}, w_k) \end{aligned}$$

for all K almost surely. By rearranging, taking expectation, and using the properties $(\lambda_k^{\sigma})^T \mathscr{G}(x_k^{\sigma}) = 0$ and $\lambda_k^T \mathscr{G}(x_k^{\sigma}) \leq 0$, which follow from the fact that the primal-dual point $(x_k^{\sigma}, \lambda_k^{\sigma})$ solves (7), it follows that

$$\sum_{k=0}^{K} \mathbb{E}\left[\alpha_k \left(\mathscr{L}(x_k, \lambda_k^{\sigma}) - \mathscr{L}(x_k^{\sigma}, \lambda_k^{\sigma}) - \frac{\sigma}{2} \|x_k^{\sigma} - x_k\|_2^2 \right) \right] \leq M + \mathbb{E}\left[T_0^{\sigma}\right] - \mathbb{E}\left[T_{K+1}^{\sigma}\right].$$

Taking $K \to \infty$ and using the boundedness of the terms on the right-hand side gives

$$\sum_{k=0}^{\infty} \mathbb{E}\left[\alpha_k \left(\mathscr{L}(x_k, \lambda_k^{\sigma}) - \mathscr{L}(x_k^{\sigma}, \lambda_k^{\sigma}) - \frac{\sigma}{2} \|x_k^{\sigma} - x_k\|_2^2 \right) \right] < \infty. \tag{16}$$

Again, since $(x_k^{\sigma}, \lambda_k^{\sigma})$ solves (7),

$$\begin{aligned} \mathscr{L}(x_k, \lambda_k^{\sigma}) &= \mathscr{F}(x_k) + (\lambda_k^{\sigma})^T \mathscr{G}(x_k) \\ &= \mathscr{F}(x_k) + (\lambda_k^{\sigma})^T \mathscr{G}(x_k) + \frac{\sigma}{2} \|x_k - x_k\|_2^2 \\ &\geq \mathscr{F}(x_k^{\sigma}) + (\lambda_k^{\sigma})^T \mathscr{G}(x_k^{\sigma}) + \frac{\sigma}{2} \|x_k^{\sigma} - x_k\|_2^2 \\ &= \mathscr{L}(x_k^{\sigma}, \lambda_k^{\sigma}) + \frac{\sigma}{2} \|x_k^{\sigma} - x_k\|_2^2, \end{aligned}$$

and hence

$$\mathscr{L}(x_k, \lambda_k^{\sigma}) - \mathscr{L}(x_k^{\sigma}, \lambda_k^{\sigma}) - \frac{\sigma}{2} \|x_k^{\sigma} - x_k\|_2^2 \geq 0.$$

It follows from this last inequality, (16), and Assumption 3 that there exists a subsequence $\{(x_{n_k}, x_{n_k}^{\sigma}, \lambda_{n_k}^{\sigma})\}_{k \in \mathbb{Z}_+}$ such that

$$\mathscr{L}(x_{n_k}, \lambda_{n_k}^{\sigma}) - \mathscr{L}(x_{n_k}^{\sigma}, \lambda_{n_k}^{\sigma}) - \frac{\sigma}{2} \|x_{n_k}^{\sigma} - x_{n_k}\|_2^2 \to 0 \quad \text{as} \quad k \to \infty \tag{17}$$

almost surely. Since $(x_k^\sigma, \lambda_k^\sigma)$ solves (7), there exists a $\eta_k^\sigma \in \partial \mathscr{L}(x_k^\sigma, \lambda_k^\sigma)$ such that

$$\left(\eta_k^\sigma + \sigma(x_k^\sigma - x_k)\right)^T (x - x_k^\sigma) \geq 0$$

for all $x \in \mathscr{X}$, and hence that

$$(\eta_{n_k}^\sigma)^T (x_{n_k} - x_{n_k}^\sigma) \geq \sigma \|x_{n_k}^\sigma - x_{n_k}\|_2^2$$

for all $k \in \mathbb{Z}_+$. It follows from this that

$$\begin{aligned}
\mathscr{L}(x_{n_k}, \lambda_{n_k}^\sigma) - \mathscr{L}(x_{n_k}^\sigma, \lambda_{n_k}^\sigma) - \frac{\sigma}{2}\|x_{n_k}^\sigma - x_{n_k}\|_2^2 &\geq (\eta_{n_k}^\sigma)^T (x_{n_k} - x_{n_k}^\sigma) - \frac{\sigma}{2}\|x_{n_k}^\sigma - x_{n_k}\|_2^2 \\
&\geq \sigma \|x_{n_k}^\sigma - x_{n_k}\|_2^2 - \frac{\sigma}{2}\|x_{n_k}^\sigma - x_{n_k}\|_2^2 \\
&= \frac{\sigma}{2}\|x_{n_k}^\sigma - x_{n_k}\|_2^2.
\end{aligned}$$

This and (17) then give that $(x_{n_k}^\sigma - x_{n_k}) \to 0$ as $k \to \infty$ almost surely. □

Theorem 1 *The sequence $\{x_k\}_{k\in\mathbb{Z}_+}$ of primal iterates produced by algorithm* (6) *has a subsequence that converges almost surely to an optimal point of problem* (1).

Proof Since x_k^σ is an optimal primal point of (7), it follows that $\mathscr{G}(x_k^\sigma) \leq 0$ and

$$\begin{aligned}
\mathscr{F}(x_k^\sigma) &\leq \mathscr{F}(x^*) + \frac{\sigma}{2}\left(\|x^* - x_k\|_2^2 - \|x_k^\sigma - x_k\|_2^2\right) \\
&\leq \mathscr{F}(x^*) + \sigma M_1
\end{aligned}$$

for all $k \in \mathbb{Z}_+$ and $\sigma > 0$, where x^* is an optimal primal point of problem (1), and $M_1 > 0$ exists due to x^*, x_k^σ and x_k all being inside the compact convex set $\mathscr{X}$. Letting $m_{-1} := 0$, it follows from the above inequalities, Lemma 6, and the continuity of $\mathscr{F}$ and $\mathscr{G}$ in $\mathscr{X}$, that for each $k \in \mathbb{Z}_+$ there exist $m_k > m_{k-1}$ and $\sigma_k > 0$ small enough with probability one such that

$$\begin{aligned}
|\mathscr{F}(x_{m_k}) - \mathscr{F}(x_{m_k}^{\sigma_k})| &\leq 1/k \\
\|\mathscr{G}(x_{m_k}) - \mathscr{G}(x_{m_k}^{\sigma_k})\|_2 &\leq 1/k \\
\mathscr{F}(x_{m_k}^{\sigma_k}) &\leq \mathscr{F}(x^*) + M_1/k.
\end{aligned}$$

It follows that the resulting sequence $\{x_{m_k}\}_{k\in\mathbb{Z}_+}$, which is a subsequence of $\{x_k\}_{k\in\mathbb{Z}_+}$, satisfies

$$\begin{aligned}
\mathscr{F}(x_{m_k}) &= \mathscr{F}(x_{m_k}^{\sigma_k}) + \mathscr{F}(x_{m_k}) - \mathscr{F}(x_{m_k}^{\sigma_k}) \\
&\leq \mathscr{F}(x_{m_k}^{\sigma_k}) + |\mathscr{F}(x_{m_k}) - \mathscr{F}(x_{m_k}^{\sigma_k})| \\
&\leq \mathscr{F}(x_{m_k}^{\sigma_k}) + 1/k \\
&\leq \mathscr{F}(x^*) + (M_1 + 1)/k
\end{aligned}$$

and

$$\begin{aligned}\mathscr{G}(x_{m_k})^T e_i &= \mathscr{G}(x_{m_k}^{\sigma_k})^T e_i + (\mathscr{G}(x_{m_k}) - \mathscr{G}(x_{m_k}^{\sigma}))^T e_i \\ &\leq \mathscr{G}(x_{m_k}^{\sigma_k})^T e_i + \|\mathscr{G}(x_{m_k}) - \mathscr{G}(x_{m_k}^{\sigma})\|_2 \\ &\leq 1/k\end{aligned}$$

for each k, where e_i is any standard basis vector. These inequalities and the boundedness of $\{x_{m_k}\}_{k\in\mathbb{Z}_+}$ imply that the sequence $\{x_k\}_{k\in\mathbb{Z}_+}$ has a subsequence that convergences almost surely to a point $\bar{x} \in \mathscr{X}$ that satisfies $\mathscr{G}(\bar{x}) \leq 0$ and $\mathscr{F}(\bar{x}) = \mathscr{F}(x^*)$. □

5.2 Drift Assumption

Assumption 5 states that for all $\sigma > 0$ small enough, the sequences of partial sums given in (8) are almost-surely bounded above. This assumption is only needed for Lemma 6 and Theorem 1. To provide a practical justification for this assumption, it is enough to consider the scalar case, *i.e.*, $\lambda_k \in \mathbb{R}$ and $x_k \in \mathbb{R}$.

By the compactness of Λ and $\mathscr{X}$, and Assumption 4, the sequences $\{\lambda_k\}$, $\{\lambda_k^{\sigma}\}$ and $\{x_k^{\sigma}\}$ are bounded. From this and Lemma 1, it holds that the sequences $\{\Gamma_k\}$, $\{\Psi_k\}$ and $\{\Upsilon_k\}$ are also bounded. Furthermore, equation (6b), the uniform boundedness of $\hat{G}_k$, and Corollary 1 give that there exists an $M_1 > 0$ such that

$$|\lambda_{k+1} - \lambda_k| \leq \alpha_k M_1, \quad |\lambda_{k+1}^{\sigma} - \lambda_k^{\sigma}| \leq \alpha_k M_1, \quad |x_{k+1}^{\sigma} - x_k^{\sigma}| \leq \alpha_k M_1$$

for all $k \in \mathbb{Z}_+$. Similarly, these inequalities, equations (6c) and (6d), the boundedness of $\{\lambda_k\}$, $\{\lambda_k^{\sigma}\}$ and $\{x_k^{\sigma}\}$, Lemma 1, and Assumption 2 imply that there exists an $M_2 > 0$ such that

$$|\Gamma_{k+1} - \Gamma_k| \leq \alpha_k M_2, \quad |\Psi_{k+1} - \Psi_k| \leq \alpha_k M_2, \quad |\Upsilon_{k+1} - \Upsilon_k| \leq \alpha_k M_2$$

for all $k \in \mathbb{Z}_+$. Hence, the partial sums in (8) (in the scalar case) are all of the form

$$S_n := \sum_{k=1}^{n} (a_{k+1} - a_k) b_k,$$

where $\{a_k\}$ and $\{b_k\}$ are bounded, and $|a_{k+1} - a_k| \leq \alpha_k M$ and $|b_{k+1} - b_k| \leq \alpha_k M$ for all $k \in \mathbb{Z}_+$ for some $M > 0$. Unfortunately, these properties alone of a_k and b_k, which can be assumed to be non-negative without loss of generality, are not enough to ensure that the partial sums S_n are bounded above in the general case, as counterexamples for this can be constructed. However, due to Assumption 3, all counterexamples require a high degree of synchronization between a_k and b_k in order to make b_k favor consistently (and infinitely often) increases in a_k, *i.e.*, $a_{k+1} - a_k > 0$, over decreases. Due to this strong synchronization requirement, it seems reasonable to expect that in practical applications the sequences of partial sums in (8) are bounded above and hence that Assumption 5 holds.

5.3 Illustrative Examples

Two simple deterministic two-dimensional examples are used to illustrate the key properties of algorithm (6). In particular, they are used to show the effects of the quality of the initial function approximation $\mathscr{G}_0$ and of the dual bound $\lambda^{\max}$ (used for defining Λ in Section 2) on the performance of the algorithm, and to illustrate the geometric ideas behind the algorithm. The effects of the quality of $\mathscr{F}_0$ are similar or less severe than those of $\mathscr{G}_0$ and hence are not shown. Both examples considered are problems of the form

$$\begin{aligned} \underset{x_1,x_2}{\text{minimize}} \quad & \mathscr{F}(x_1,x_2) && \text{(18a)} \\ \text{subject to} \quad & \mathscr{G}(x_1,x_2) \leq 0 && \text{(18b)} \\ & x^{\min} \leq x_i \leq x^{\max},\ i \in \{1,2\}. && \text{(18c)} \end{aligned}$$

The first example consists of the Quadratic Program (QP) obtained using

$$\begin{aligned} \mathscr{F}(x_1,x_2) &= x_1^2 + x_2^2 \\ \mathscr{G}(x_1,x_2) &= -0.5x_1 - x_2 + 2 \\ (x^{\min},x^{\max}) &= (-6,6). \end{aligned}$$

The primal optimal point of the resulting problem is $(x_1^*,x_2^*) = (0.8,1.6)$, and the dual optimal point associated with constraint (18b) is $\lambda^* = 3.2$. For the objective function, the initial approximation

$$\mathscr{F}_0(x_1,x_2) = 0.3(x_1-1)^2 + 2(x_2-0.5)^2$$

is used. For the constraint function, the initial approximations

$$\begin{aligned} \mathscr{G}_0^1(x_1,x_2) &= -1.5x_1 - 2x_2 \\ \mathscr{G}_0^2(x_1,x_2) &= 3x_1 + x_2 \\ \mathscr{G}_0^3(x_1,x_2) &= 4x_1 + 9x_2, \end{aligned}$$

are used, which can be considered "good", "poor", and "very poor", respectively, according to how their gradients compare against that of the exact function at the optimal point. Note that for this case, the function approximations are of the same form as the exact functions, *i.e.*, the objective function approximation is quadratic and the constraint function approximation is linear. The algorithm parameters used in this example are $\lambda_0 = \lambda^*/8$ and $\alpha_k = 1/(k+3)$.

Figure 1 shows the primal error, dual error, and drift as a function of the iterations of the algorithm for the different initial constraint approximations and dual bounds. The drift shown here is the maximum of the partial sums in (8) using $\sigma = 1$. As the figure shows, the primal and dual errors approach zero relatively fast and the drift remains small when using $\mathscr{G}_0^1$ and $\mathscr{G}_0^2$. On the other hand, when using the "very poor" approximation $\mathscr{G}_0^3$ and a loose dual bound, the primal error, dual error and drift all become large during the initial iterations of the algorithm. In this case, a very large number of iterations is required by the algorithm to approach the solution. Compared

to this case, the performance of the algorithm is significantly improved when using a tighter dual bound. This is because the better bound prevents the dual iterates from continuing to grow, and this allows the slope corrections to "catch up" and make the constraint approximation more consistent with the exact function.

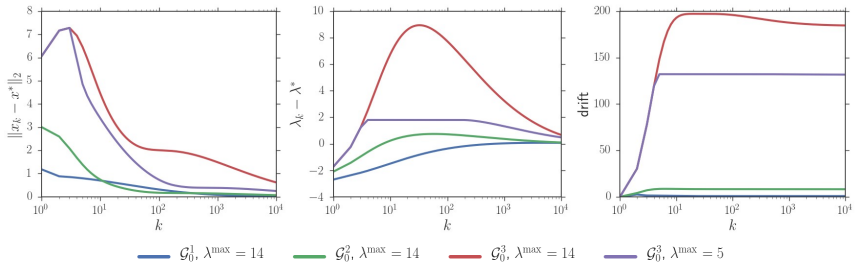

Fig. 1 Algorithm performance on QP

Figure 2 shows the contours of the constraint and objective function approximations during certain iterations using $\mathscr{G}_0^2$ as well as those of the exact functions. The gradients of these functions are shown at the optimal primal points associated with the functions. As the figure shows, the initial approximations (top left) are poor since their gradients point in opposite directions as those of the exact functions (bottom right). As the number of iterations increases, the slope corrections performed by the algorithm improve the direction of these gradients and hence the quality of the iterates. An important observation is that the shape of the contours does not change since only the slope and not the curvature of the function approximations is affected. For example, the gradient of the objective function approximation at the solution matches well that of the exact function but its contours are elliptical and not circular.

The second example consists of the Quadratically Constrained Quadratic Program (QCQP) obtained using

$$\begin{aligned}\mathscr{F}(x_1,x_2) &= x_1^2 + x_2^2\\ \mathscr{G}(x_1,x_2) &= (x_1-1.8)^2 + (x_2-1.4)^2 - 1\\ (x^{\min},x^{\max}) &= (-6,6).\end{aligned}$$

The primal optimal point for this problem is $(x_1^*,x_2^*) = (1,0.8)$, and the dual optimal point associated with constraint (18b) is $\lambda^* = 1.3$. The initial objective function approximation used is

$$\mathscr{F}_0(x_1,x_2) = 0.3(x_1-1)^2 + 2(x_2-0.5)^2,$$

and the initial constraint function approximations used are

$$\begin{aligned}\mathscr{G}_0^1(x_1,x_2) &= -1.5x_1 - 2x_2\\ \mathscr{G}_0^2(x_1,x_2) &= -1.5x_1 + 3x_2\\ \mathscr{G}_0^3(x_1,x_2) &= 2x_1 + 4x_2,\end{aligned}$$

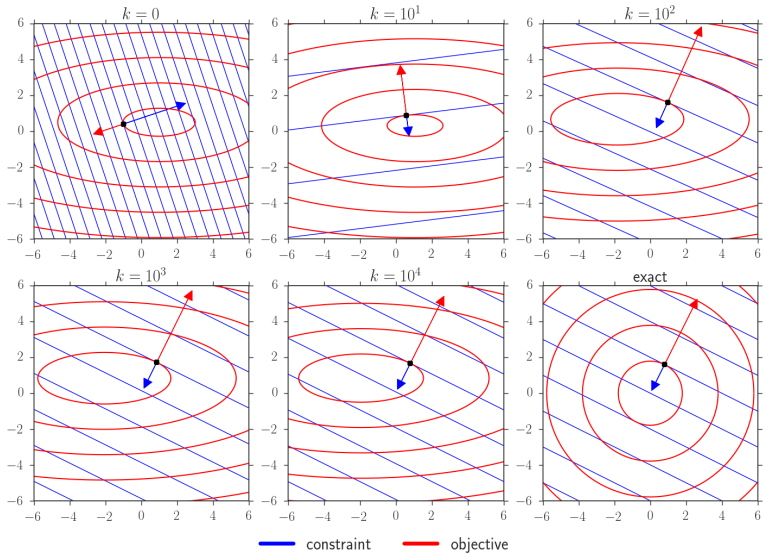

Fig. 2 Contours of objective and constraint approximations of QP using $\mathscr{G}_0^2$ and $\lambda^{\max} = 14$

which again can be considered "good", "poor", and "very poor", respectively. Note that for this case, only the objective function approximation is of the same form as the exact function. The initial constraint approximation is not since it is linear while the exact function is quadratic. The algorithm parameters used in this example are $\lambda_0 = \lambda^*/6$ and $\alpha_k = 1/(k+3)$.

Figures 3 and 4 show the performance of the algorithm and the evolution of the function approximations, respectively, on the second example. The results obtained are similar to those obtained on the first example. That is, the algorithm can perform well with reasonable initial function approximations. With very poor initial function approximations, the primal error, dual error and drift can all increase significantly during the initial iterations until the approximations become reasonable, resulting in the algorithm requiring a large number of iterations to approach the solution. Furthermore, when the initial function approximation is very poor, the quality of the dual bound plays an important role in the performance of the algorithm.

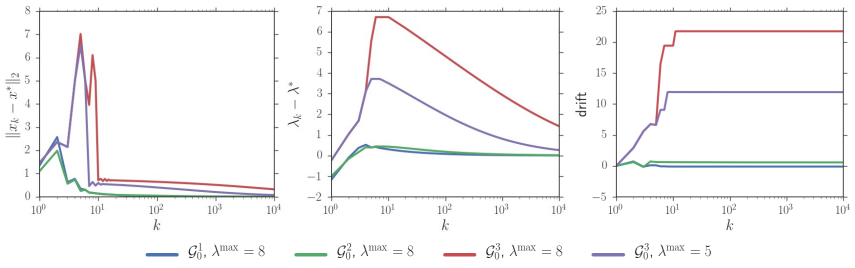

Fig. 3 Algorithm performance on QCQP

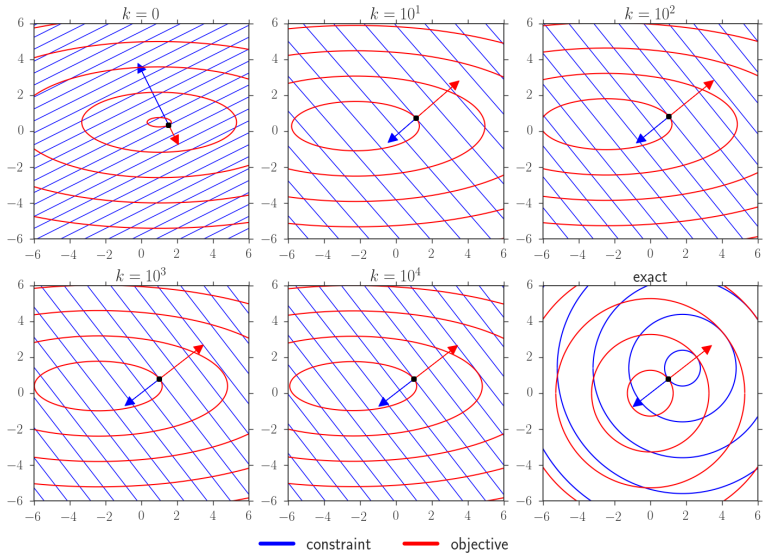

Fig. 4 Contours of objective and constraint approximations of QCQP using $\mathscr{G}_0^2$ and $\lambda^{\max} = 8$

6 Application: Optimal Risk-Averse Generator Dispatch

In this section, the performance of the algorithms described in Sections 3, 4 and 5 is compared on a stochastic optimization problem that contains expectation functions in both the objective and constraints. This problem originates in power systems operations planning under high penetration of renewable energy, where optimal risk-averse power generation policies are sought. Mathematically, the problem is formulated as a two-stage stochastic optimization problem with quadratic objectives and a CVaR constraint that limits the risk of high second-stage costs.

In electric power systems, which are also known as power networks or grids, the output powers of generators are typically planned in advance, *e.g.*, hours ahead. The reason for this is that low-cost generators are typically slow and hence require some time to change output levels. Uncertainty in the operating conditions of a system, such as power injections from renewable energy sources, must be taken into account during operations planning. This is because real-time imbalances in power production and consumption are undesirable as they must be resolved with resources that typically have a high cost, such as fast-ramping generators or load curtailments. Furthermore, due to the large quantities of money involved in the operation of a power system and society's dependence on reliable electricity supply, risk-averse operation policies are desirable. The problem of determining optimal risk-averse power generation policies hence consists of determining generator output powers in advance that minimize expected operation cost, while ensuring that the risk of high balancing costs is below an acceptable limit.

6.1 Problem Formulation

Mathematically, a simplified version of the problem of determining optimal risk-averse power generation policies can be formulated as

$$\underset{p\in\mathscr{P}}{\text{minimize}} \quad \varphi_0(p)+\mathbb{E}\,[Q(p,r)] \tag{19a}$$

$$\text{subject to} \quad \text{CVaR}_\gamma(Q(p,r)-Q^{\max})\leq 0, \tag{19b}$$

where $p\in\mathbb{R}^{n_p}$ is a vector of planned generator output powers, n_p is the number of generators, $r\in\mathbb{R}^{n_r}_+$ is a vector of uncertain renewable powers available in the near future, say one hour ahead, n_r is the number of renewable energy sources, φ_0 is a strongly convex separable quadratic function that quantifies planned generation cost, $Q(p,r)$ quantifies the real-time power balancing cost for a given plan of generator outputs and realization of the uncertainty, $Q^{\max}$ is a positive constant that defines a desirable limit on the balancing cost, the set

$$\mathscr{P}:=\{\ \bar{p}\mid p^{\min}\leq\bar{p}\leq p^{\max}\ \}$$

represents power limits of generators, and CVaR_γ is conditional value-at-risk with parameter $\gamma\in[0,1]$. CVaR is a coherent risk measure that quantifies dangers beyond the Value-at-Risk (VaR) [22], and is given by the formula

$$\text{CVaR}_\gamma(Z)=\inf_{t\in\mathbb{R}}\left(t+\frac{1}{1-\gamma}\mathbb{E}\,[(Z-t)_+]\right), \tag{20}$$

for any random variable Z, where $(\,\cdot\,)_+:=\max\{\,\cdot\,,0\}$ [1] [23]. An important property of this risk measure is that constraint (19b) is a convex conservative approximation (when $Q(\,\cdot\,,r)$ is convex) of the chance constraint

$$\text{Prob}\,(Q(p,r)\leq Q^{\max})\geq\gamma,$$

and hence it ensures that the balancing cost is below the predefined limit with a desired probability [23]. The function $Q(p,r)$ is given by the optimal objective value of a second-stage optimization problem that represents real-time balancing of the power system, which is given by

$$\underset{q,\theta,s}{\text{minimize}} \quad \varphi_1(q) \tag{21a}$$

$$\text{subject to} \quad Y(p+q)+Rs-A\theta-Dd=0 \tag{21b}$$

$$p^{\min}\leq p+q\leq p^{\max} \tag{21c}$$

$$z^{\min}\leq J\theta\leq z^{\max} \tag{21d}$$

$$0\leq s\leq r. \tag{21e}$$

Here, φ_1 is a strongly convex separable quadratic function that quantifies the cost of power balancing adjustments, $q\in\mathbb{R}^{n_p}$ are the power adjustments, which for simplicity are treated as changes in planned generator powers, $\theta\in\mathbb{R}^{n_b}$ are node voltage phase angles, n_b is the number of non-slack nodes, $s\in\mathbb{R}^{n_r}$ are powers from renewable sources after possible curtailments, $d\in\mathbb{R}^{n_d}$ are load power consumptions, n_d

is the number of loads, $\{Y,R,A,D,J\}$ are sparse matrices, constraint (21b) enforces conservation of power using a DC power flow model [24], constraint (21c) enforces generator power limits, constraint (21d) enforces edge flow limits due to thermal ratings of transmission lines and transformers, and constraint (21e) enforces renewable power curtailment limits. Under certain assumptions, which are made here, the functions $Q(\cdot,r)$ and $\mathbb{E}[Q(\cdot,r)]$ are convex and differentiable [2].

Using the definition (20) of CVaR and adding bounds for the variable t, the first-stage problem (19) can be reformulated as

$$\begin{aligned} \underset{p,t}{\text{minimize}} \quad & \mathbb{E}[\varphi_0(p)+Q(p,r)] && \text{(22a)}\\ \text{subject to} \quad & \mathbb{E}\left[(1-\gamma)t+(Q(p,r)-Q^{\max}-t)_+\right] \leq 0 && \text{(22b)}\\ & (p,t) \in \mathscr{P} \times \mathscr{T}, && \text{(22c)} \end{aligned}$$

where $\mathscr{T} := [t^{\min}, t^{\max}]$. In order to avoid affecting the problem, $t^{\max}$ can be set to zero while $t^{\min}$ can be set to a sufficiently large negative number. The resulting problem is clearly of the form of (1).

6.2 Uncertainty

As already noted, the random vector r represents available powers from renewable energy sources in the near future, say one hour ahead. It is modeled here by the expression

$$r = \Pi_r(r_0+\delta),$$

where r_0 is a vector of base powers, $\delta \sim \mathscr{N}(0,\Sigma)$ are perturbations, Π_r is the projection operator on the set

$$\mathscr{R} := \{\, \bar{r} \mid 0 \leq \bar{r} \leq r^{\max} \,\},$$

and the vector $r^{\max}$ represents the capacities of the renewable energy sources.

The use of Gaussian random variables for modeling renewable energy uncertainty is common in the power systems operations planning literature [25] [26]. To model spatial correlation between nearby renewable sources, a sparse non-diagonal covariance matrix Σ is considered here. The off-diagonal entries of this matrix that correspond to pairs of renewable energy sources that are "nearby" are such that the correlation coefficient between their powers equals a pre-selected value ρ. In the absence of geographical information, a pair of sources are considered as being "nearby" if the network shortest path between the nodes where they are connected is less than or equal to some pre-selected number of edges N. This maximum number of edges is referred here as the correlation distance.

6.3 Noisy Functions and Subgradients

By comparing problems (1) and (22), the functions $F(\cdot,w)$ and $G(\cdot,w)$ are given by

$$\begin{aligned} F(x,w) &= \varphi_0(p)+Q(p,r)\\ G(x,w) &= (1-\gamma)t+(Q(p,r)-Q^{\max}-t)_+ \end{aligned}$$

for all $x = (p,t)$, where $w = r$. Subgradients of these functions are given by the expressions

$$\begin{bmatrix} \nabla\varphi_0(p) + \nabla_p Q(p,r) \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} 0 \\ 1-\gamma \end{bmatrix} + 1\{Q(p,r) - Q^{\max} - t \geq 0\} \begin{bmatrix} \nabla_p Q(p,r) \\ -1 \end{bmatrix},$$

respectively, where $1\{\cdot\}$ is the indicator function defined by

$$1\{A\} = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise,} \end{cases}$$

for any event or condition A. As shown in [2], $\nabla_p Q(p,r) = -\nabla\varphi_1(q^*)$, where q^* is part of the optimal point of problem (21).

6.4 Function Approximations

In order to apply the primal-dual stochastic hybrid approximation algorithm (6), initial function approximations $\mathscr{F}_0$ and $\mathscr{G}_0$ need to be defined. $\mathscr{F}_0$ needs to be strongly convex and continuously differentiable, while $\mathscr{G}_0$ needs to be component-wise convex and continuously differentiable. Motivated by the results obtained in [2], the approximations used here are also based on the certainty-equivalent formulation of problem (1), namely, the problem obtained by replacing $\mathbb{E}[f(p,r)]$ with $f(p,\bar{r})$, where $\bar{r} := \mathbb{E}[r]$, for any stochastic function f. More specifically, $\mathscr{F}_0$ is defined by

$$\mathscr{F}_0(x) := \varphi_0(p) + Q(p,\bar{r}) + \frac{1}{2}\varepsilon t^2, \tag{23}$$

for all $x = (p,t)$, where $\varepsilon > 0$, and $\mathscr{G}_0$ is defined by

$$\mathscr{G}_0(x) := (1-\gamma)t + (Q(p,\bar{r}) - Q^{\max} - t)_+, \tag{24}$$

for all $x = (p,t)$. This choice of function $\mathscr{G}_0$ is differentiable everywhere except when $Q(p,\bar{r}) - Q^{\max} - t = 0$. As will be seen in the experiments, this deficiency does not cause problems on the test cases considered. For cases for which this may cause problems, *e.g.*, for cases for which $Q(p,\bar{r}) - Q^{\max} - t = 0$ occurs near or at the solution, the smooth function

$$(1-\gamma)t + \frac{1}{\nu}\log\left(1 + e^{\nu(Q(p,\bar{r}) - Q^{\max} - t)}\right)$$

may be used instead, where ν is a positive parameter.

During the k-th iteration of algorithm (6), the function approximations can be expressed as

$$\begin{aligned} \mathscr{F}_k(x) &= \mathscr{F}_0(x) + \zeta_k^T x \\ \mathscr{G}_k(x) &= \mathscr{G}_0(x) + \upsilon_k^T x, \end{aligned}$$

for all $x = (p,t)$, where the vectors ζ_k and υ_k follow the recursive relations

$$\zeta_{k+1} = \zeta_k + \alpha_k \left(\hat{g}_k - g_0(x_k) - \zeta_k\right)$$
$$\upsilon_{k+1} = \upsilon_k + \alpha_k \left(\hat{J}_k - J_0(x_k) - \upsilon_k^T\right)^T$$

with $\zeta_0 = \upsilon_0 = 0$, and $\hat{g}_k$, g_0, $\hat{J}_k$ and J_0 are as defined in Section 5.

6.5 Subproblems

To perform step (6a) of the primal-dual stochastic hybrid approximation algorithm, the subproblem

$$\underset{x \in \mathscr{X}}{\text{minimize}} \quad \mathscr{F}_k(x) + \lambda_k^T \mathscr{G}_k(x)$$

needs to be solved, where $\lambda_k \geq 0$. For problem (22) with initial function approximations (23) and (24), this subproblem is given by

$$\begin{array}{ll} \underset{p,t,z}{\text{minimize}} & \varphi_0(p) + Q(p,\bar{r}) + \dfrac{1}{2}\varepsilon t^2 + \lambda_k(1-\gamma)t + \lambda_k z + (\zeta_k + \lambda_k \upsilon_k)^T \begin{bmatrix} p \\ t \end{bmatrix} \\ \text{subject to} & Q(p,\bar{r}) - Q^{\max} - t - z \leq 0 \\ & (p,t) \in \mathscr{P} \times \mathscr{T} \\ & 0 \leq z. \end{array}$$

By embedding the definition of $Q(p,\bar{r})$, this subproblem can be reformulated as

$$\begin{array}{ll} \underset{p,t,q,\theta,s,z}{\text{minimize}} & \varphi_0(p) + \varphi_1(q) + \dfrac{1}{2}\varepsilon t^2 + \lambda_k(1-\gamma)t + \lambda_k z + (\zeta_k + \lambda_k \upsilon_k)^T \begin{bmatrix} p \\ t \end{bmatrix} \\ \text{subject to} & Y(p+q) + Rs - A\theta - Dd = 0 \\ & \varphi_1(q) - Q^{\max} - t - z \leq 0 \\ & p^{\min} \leq p + q \leq p^{\max} \\ & z^{\min} \leq J\theta \leq z^{\max} \\ & p^{\min} \leq p \leq p^{\max} \\ & t^{\min} \leq t \leq t^{\max} \\ & 0 \leq s \leq \bar{r} \\ & 0 \leq z, \end{array} \tag{25}$$

which is a sparse QCQP.

6.6 Implementation

The benchmark primal-dual stochastic approximation algorithm (4) and the new primal-dual stochastic hybrid approximation algorithm (6) were implemented in Python using the scientific computing packages `Numpy` and `Scipy` [27] [28]. The code for these algorithms has been included in the Python package `OPTALG` [1].

[1] https://github.com/ttinoco/OPTALG

The optimal risk-averse generation planning problem (22) was also constructed using Python and included in the Python package `GRIDOPT` [2]. In particular, the C library `PFNET` [3] was used through its Python wrapper for representing power networks and for constructing the required constraints and functions of the problem. For solving the sparse QPs (21) and sparse QCQPs (25), the modeling library `CVXPY` [29] was used together with the second-order cone solver `ECOS` [30]. The SAA-based benchmark algorithms of Section 3, namely the one that solves (2) directly and algorithm (3) that uses decomposition and cutting planes were also implemented using `CVXPY` and `ECOS`.

The covariance matrix Σ of the renewable power perturbations defined in Section 6.2 was constructed using routines available in `PFNET`. For obtaining $\Sigma^{1/2}$ to sample $\delta \sim \mathcal{N}(0,\Sigma)$, the sparse Cholesky code `CHOLMOD` [31] was used via its Python wrapper [4].

6.7 Numerical Experiments

The stochastic optimization algorithms described in Sections 3, 4 and 5 were applied to two instances of the optimal risk-averse generation planning problem (22) in order to test and compare their performance. The test cases used were constructed from real power networks from North America (Canada) and Europe (Poland). Table 1 shows important information about each of the cases, including name, number of nodes in the network, number of edges, number of first-stage variables (dimension of (p,t) in (22)), number of second-stage variables (dimension of (q,θ,s) in (21)), and dimension of the uncertainty (dimension of the vector r in (22)).

Table 1 Information about test cases

name	nodes	edges	dim (p,t)	dim (q,θ,s)	dim r
Case A	2454	2781	212	2900	236
Case B	3012	3572	380	3688	298

The separable generation cost function φ_0 in (22) was constructed using uniformly random coefficients in $[0.01, 0.05]$ for the quadratic terms, and uniformly random coefficients in $[10, 50]$ for the linear terms. These coefficients assume that power quantities are in units of MW and are consistent with those found in several `MATPOWER` cases [32]. For the separable generation adjustment cost function φ_1 in (21), the coefficients for the quadratic terms were set to equal those of φ_0 scaled by a factor greater than one, and the coefficients of the linear terms were set to zero. The scaling factor was set large enough to make balancing costs with adjustments higher than with planned powers. The coefficients for the linear terms were set to zero to pe-

[2] https://github.com/ttinoco/GRIDOPT

[3] https://github.com/ttinoco/PFNET

[4] http://pythonhosted.org/scikits.sparse/

nalize both positive and negative power adjustments equally. For the t-regularization added in (23) to make $\mathscr{F}_0$ strongly convex, $\varepsilon = 10^{-6}$ was used.

The renewable energy sources used to construct the test cases of Table 1 were added manually to each of the power networks at the nodes with adjustable or fixed generators. The capacity $r_i^{\max}$ of each source was set to $1^T d/n_r$, where d is the vector of load power consumptions and n_r is the number of renewable energy sources. The base powers were set using $r_0 = 0.5 r^{\max}$ so that the base renewable energy penetration was 50% of the load, which is a high penetration scenario. The standard deviations of the renewable power perturbations, *i.e.*, $\operatorname{diag} \Sigma^{1/2}$, were also set to $0.5 r^{\max}$, which corresponds to a high-variability scenario. For the off-diagonals of Σ, a correlation coefficient ρ of 0.05 and a correlation distance N of 5 edges were used.

For the risk-limiting constraint (22b), the desirable balancing cost limit $Q^{\max}$ was set to $0.8Q_0$, where $Q_0 := \mathbb{E}[Q(p_0, r)]$ and p_0 is the optimal power generation policy for the certainty-equivalent problem without risk-limiting constraint, *i.e.*,

$$p_0 := \operatorname*{arg\,min}_{p \in \mathscr{P}} \ \varphi_0(p) + Q(p, \mathbb{E}[r]).$$

The CVaR parameter γ defined in Section 6.1, which is related to the probability of satisfying the inequality $Q(p,r) \leq Q^{\max}$, was set to 0.95. For bounding the variable t in problem (22), $t^{\max}$ and $t^{\min}$ were set to 0 and $-0.1Q_0$, respectively. For choosing this value of $t^{\min}$, some experimentation was required. In particular, executing algorithms (4) and (6) for a few iterations was needed in order to check that $t^{\min}$ was sufficiently negative to allow for feasibility, but not too large in magnitude. The latter was needed to avoid encountering excessively large values of $G(x_k, w_k)$ (and hence large increases in λ_k) due to poor values of t in the initial iterations.

Each of the algorithms was applied to the test cases shown in Table 1 and its performance was recorded. In particular, the SAA cutting-plane algorithm (3) with 1000 scenarios (`SAA CP 1000`), the primal-dual stochastic approximation algorithm (4) (`PD SA`), and the primal-dual stochastic hybrid approximation algorithm (6) considering and ignoring the risk-limiting constraint (`PD SHA` and `SHA`, respectively) were executed for a fixed time period. For the `PD SA` algorithm, $(p_0, t^{\min})$ was used as the initial primal solution estimate. For both `PD SHA` and `PD SA`, 0 was used as the initial dual solution estimate λ_0, and the step lengths α_k used were of the form $(k_0 + k)^{-1}$, where k_0 is a constant. A k_0 value of 50 was used for the `PD SHA` algorithm, while a k_0 value of 350 was used for the `PD SA` algorithm. These values were chosen to reduce the susceptibility of the algorithms to noise during the initial iterations. Only for the `SAA CP 1000` algorithm, 24 parallel processes were used. This was done to compute values and subgradients of the sample-average second-stage cost function efficiently. For evaluating the progress of the algorithms as a function of time, the quantities $\mathbb{E}[F(x,w)]$, $\mathbb{E}[G(x,w)]$ and $\text{Prob}(Q(p,r) \leq Q^{\max})$ were evaluated using 2000 fixed scenarios at fixed time intervals. Figures 5 and 6 show the results obtained on a computer cluster at the Swiss Federal Institute of Technology (ETH) with computers equipped with Intel®Xeon®E5-2697 CPU (2.70 GHz), and running the operating system CentOS 6.8. In the figures, the objective values shown were normalized by $\varphi_0(p_0) + Q_0$, while the constraint function values shown were normalized by Q_0, which was defined in the previous paragraph. The dashed lines are meant to represent

references, and correspond to the results obtained by solving the certainty-equivalent problem ignoring the risk-limiting constraint (`CE`), and the SAA-based algorithm that consists of solving problem (2) directly (without decomposition) using 1000 scenarios (`SAA 1000`). The times needed to compute these references are show in Table 2. Table 3 shows the (maximum) memory requirements of the algorithms on each of the test cases.

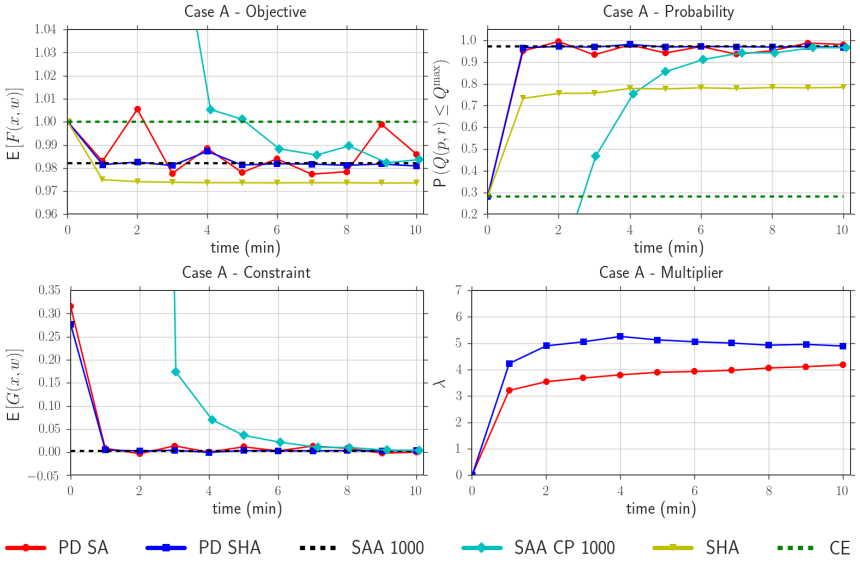

Fig. 5 Performance of algorithms on Case A

Table 2 Computation times of references in minutes

algorithm	Case A	Case B
`CE`	0.0075	0.018
`SAA 1000`	15	92

Table 3 Memory requirements of algorithms in gigabytes

algorithm	Case A	Case B
`CE`	0.14	0.16
`SAA 1000`	8.4	13
`SAA CP 1000`	1.5	1.8
`PD SA`	0.12	0.20
`PD SHA`	0.14	0.16

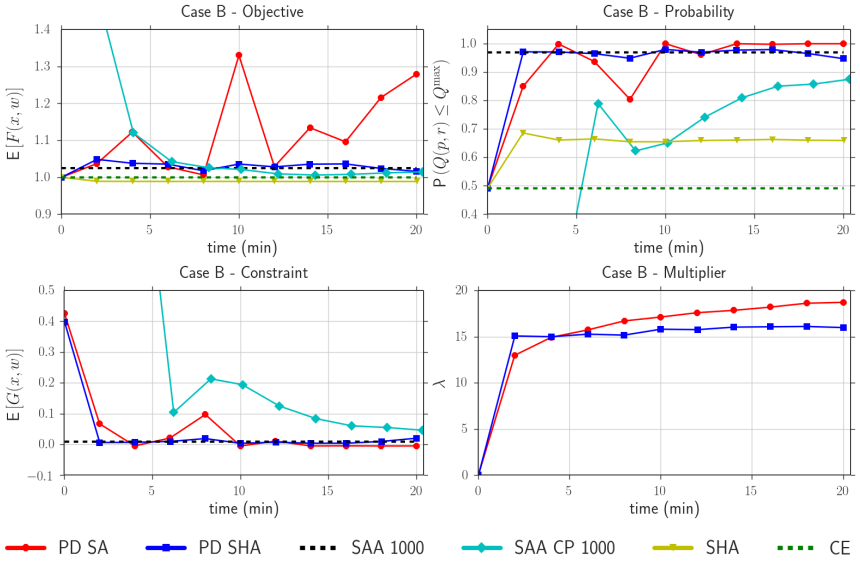

Fig. 6 Performance of algorithms on Case B

From the plots shown in Figures 5 and 6, several important observations can be made: First, the `CE` solution can result in both poor expected cost and high risk of violating $Q(p,r) \leq Q^{\max}$. On Case A, for example, all algorithms find power generation policies that result in lower expected cost and in $Q(p,r) \leq Q^{\max}$ being satisfied with significantly higher probability. Second, by ignoring the risk-limiting constraint, the `SHA` algorithm converges fast, *i.e.*, in a few minutes, and finds a power generation policy that results in the lowest expected cost. The probability of satisfying $Q(p,r) \leq Q^{\max}$, however, is only around 0.7 and 0.8 for both cases, and hence the policy found is not enough risk-averse. Third, by considering the risk-limiting constraint, the `PD SHA` and `PD SA` algorithms produce in only a few minutes power generation policies that are risk-averse, *i.e.*, $\text{Prob}(Q(p,r) \leq Q^{\max}) \geq \gamma$. This is, of course, at the expense of higher expected costs compared with those obtained with the `SHA` algorithm. The iterates produced by the `PD SA` algorithm experience significant variations and do not appear to settle during the execution period. This is particularly visible in the plot of expected cost for both test cases, and in the plot of probability for Case B. The `PD SHA` algorithm, on the other hand, produces iterates that experience much less variations (despite using larger step lengths) and converge during the execution period. The expected cost and risk associated with the policy found by the `PD SHA` algorithm match those obtained with the reference policy found with the `SAA 1000` algorithm, which requires more time and much more computer memory, as seen on Tables 2 and 3. Lastly, the other SAA-based algorithm, namely `SAA CP 1000`, also produces policies whose expected cost and risk approach those obtained using the `SAA 1000` algorithm. However, it requires more computer memory and, despite using 24 parallel processes, more time compared to the `PD SHA` algorithm.

The power generation policies associated with the last iterates produced by the primal-dual stochastic hybrid approximation algorithm considering and ignoring the risk-limiting constraint (`PD SHA` and `SHA`, respectively) were compared. The results are shown in Figure 7. The plots show the (sorted) power output differences in MW of the generators with respect to p_0, which is the solution of the certainty-equivalent problem without risk-limiting constraint (`CE`). As the plots show, the power generation policy obtained when ignoring the risk-limiting constraint shows a consistent negative bias on both test cases with respect to the risk-averse policy. This negative bias corresponds to lower planned generator powers, which is expected. The reason is that higher planned generator powers help reduce the risk of requiring generator output increases in real time when renewable power injections are low. On the other hand, when renewable power injections are high, renewable power curtailments can be used as a means for balancing instead of generator output adjustments.

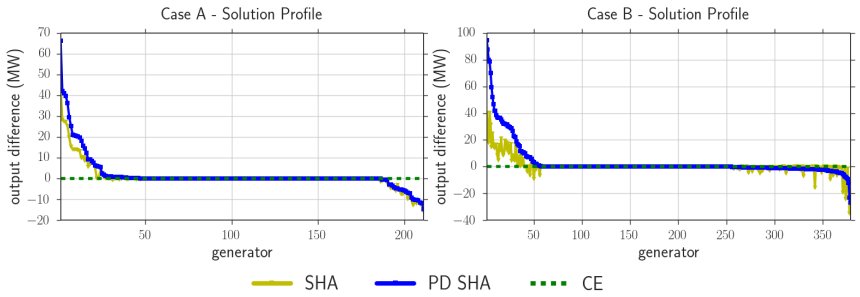

Fig. 7 Power generation policies found by algorithms

7 Conclusions

In this work, a new algorithm for solving convex stochastic optimization problems with expectation functions in both the objective and constraints has been described and evaluated. The algorithm combines a stochastic hybrid procedure, which was originally designed to solve problems with expectation only in the objective, with dual stochastic gradient ascent. More specifically, the algorithm generates primal iterates by minimizing deterministic approximations of the Lagrangian that are updated using noisy subgradients, and dual iterates by applying stochastic gradient ascent to the true Lagrangian. A proof that the sequence of primal iterates produced by this algorithm has a subsequence that converges almost surely to an optimal point under certain conditions has been provided. The proof relies on a "bounded drift" assumption for which we have provided an intuitive justification for why it may hold in practice as well as examples that illustrate its validity using different initial function approximations. Furthermore, experimental results obtained from applying the new and benchmark algorithms to instances of a stochastic optimization problem that originates in power systems operations planning under high penetration of renewable

energy have been reported. In particular, the performance of the new algorithm has been compared with that of a primal-dual stochastic approximation algorithm and algorithms based on sample-average approximations with and without decomposition. The results obtained showed that the new algorithm had superior performance compared to the benchmark algorithms on the test cases considered. In particular, the new algorithm was able to leverage accurate initial function approximations and produce iterates that approached a solution fast without requiring large computer memory or parallel computing resources.

Next research steps for this work include testing the new algorithm on problems with multiple expected-value constraints, extending the theoretical analysis, exploring new techniques for improving performance, and extending or modifying the algorithm to handle more complicated problems. Promising techniques for improving the performance of the algorithm include importance sampling, parallel computing, and the incorporation of curvature updates. Interesting directions for extending or modifying the algorithm include handling non-convex stochastic problems, and multi-stage problems. With regards to the theory, simple and easy-to-check conditions that guarantee the bounded drift condition are needed, and also an extension of the convergence analysis to show when the sequence of iterates and not just a subsequence converges and at what rate. Lastly, reliable termination conditions need to be investigated and tested in order to make the algorithm effective and easy-to-use in practice.

Acknowledgment

We thank the coordinating editor and reviewers for their constructive feedback and suggestions for improving the paper. We also acknowledge the support provided by the ETH Zurich Postdoctoral Fellowship FEL-11 15-1, which made this work possible.

References

1. W. Wang and S. Ahmed. Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters*, 36(5):515–519, 2008.
2. T. Tinoco De Rubira and G. Hug. Adaptive certainty-equivalent approach for optimal generator dispatch under uncertainty. In *European Control Conference*, June 2016.
3. S. Bhatnagar, N. Hemachandra, and V.K. Mishra. Stochastic approximation algorithms for constrained optimization via simulation. *ACM Transactions on Modeling and Computer Simulation*, 21(3):1–22, February 2011.
4. J. Atlason, M.A. Epelman, and S.G. Henderson. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, 127(1-4): 333–358, 2004.
5. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19 (4):1574–1609, 2009.

6. T. Homem-de-Mello and G. Bayraksan. Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.
7. J.R. Birge and H. Tang. L-shaped method for two stage problems of stochastic convex programming. Technical report, University of Michigan, Department of Industrial and Operations Engineering, August 1993.
8. R. Van Slyke and R. Wets. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17 (4):638–663, 1969.
9. A.J. King and R.T. Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18(1):148–162, February 1993.
10. A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.
11. H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.
12. J. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45(10):1839–1853, October 2000.
13. H. Kushner and D. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Applied Mathematical Sciences. Springer-Verlag, 1978.
14. H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer-Verlag, 2003.
15. J. Zhang and D. Zheng. A stochastic primal-dual algorithm for joint flow control and MAC design in multi-hop wireless networks. In *Conference on Information Sciences and Systems*, pages 339–344, March 2006.
16. R.K. Cheung and W.B. Powell. SHAPE - A stochastic hybrid approximation procedure for two-stage stochastic programs. *Operations Research*, 48(1):73–79, 2000.
17. W.B. Powell. The optimizing-simulator: Merging simulation and optimization using approximate dynamic programming. In *Winter Simulation Conference*, pages 96–109, 2005.
18. W.B. Powell, A. Ruszczyński, and H. Topaloglu. Learning algorithms for separable approximations of discrete stochastic optimization problems. *Mathematics of Operations Research*, 29(4):814–836, 2004.
19. W.B. Powell and H. Topaloglu. Stochastic programming in transportation and logistics. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 555–635. Elsevier, 2003.
20. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.
21. S. Bhatnagar, H. Prasad, and L. Prashanth. *Stochastic Approximation Algorithms*, pages 17–28. Springer London, 2013.
22. R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, pages 1443–1471, 2002.

23. A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.
24. J. Zhu. *Optimization of Power System Operation*. John Wiley & Sons, Inc., 2009.
25. D. Phan and S. Ghosh. Two-stage stochastic optimization for optimal power flow under renewable generation uncertainty. *ACM Transactions on Modeling and Computer Simulation*, 24(1):1–22, January 2014.
26. L. Roald, F. Oldewurtel, T. Krause, and G. Andersson. Analytical reformulation of security constrained optimal power flow with probabilistic constraints. In *IEEE PowerTech Conference*, pages 1–6, June 2013.
27. E. Jones, T. Oliphant, and P. Peterson. SciPy: Open source scientific tools for Python. `http://www.scipy.org`, 2001.
28. S. Van der Walt, S.C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2): 22–30, March 2011.
29. S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.
30. A. Domahidi, E. Chu, and S. Boyd. ECOS: An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076, 2013.
31. Y. Chen, T.A. Davis, W.W. Hager, and S. Rajamanickam. Algorithm 887: CHOLMOD, supernodal sparse cholesky factorization and update/downdate. *ACM Transactions on Mathematical Software*, 35(3):1–14, October 2008.
32. R. Zimmerman, C. Murillo-Sánchez, and R. Thomas. MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, February 2011.