Conditional gradient methods for stochastically constrained convex minimization

Maria-Luiza Vladarean 1 Ahmet Alacaoglu 1 Ya-Ping Hsieh 1 Volkan Cevher 1

Abstract

We propose two novel conditional gradient-based methods for solving structured stochastic convex optimization problems with a large number of linear constraints. Instances of this template naturally arise from SDP-relaxations of combinatorial problems, which involve a number of constraints that is polynomial in the problem dimension. The most important feature of our framework is that only a subset of the constraints is processed at each iteration, thus gaining a computational advantage over prior works that require full passes. Our algorithms rely on variance reduction and smoothing used in conjunction with conditional gradient steps, and are accompanied by rigorous convergence guarantees. Preliminary numerical experiments are provided for illustrating the practical performance of the methods.

1. Introduction

We study the following optimization template:

\[
\begin{align*}
\min_{x \in \mathcal{X}} \quad & f(x) := \mathbb{E}[f(x, \xi)] \\
\text{s.t.} \quad & A(\xi)x \in b(\xi) \quad \text{almost surely},
\end{align*}
\]

where \( f(x, \xi) : \mathbb{R}^d \to \mathbb{R} \) are random convex functions with \( L_f \)-Lipschitz gradient, \( \mathcal{X} \) is a convex and compact set of \( \mathbb{R}^d \), \( A(\xi) \) is an \( m \times d \) matrix-valued random variable, and \( b(\xi) \) is a closed and projectable random convex set in \( \mathbb{R}^m \).

Stochastically constrained convex optimization problems have recently gained interest in the machine learning community, as they provide a convenient and powerful framework for handling instances subject to a large, or even infinite number of constraints. For example, convex feasibility and optimal control problems have variables lying in a possibly infinite intersection of stochastic, projectable constraint sets, and hence are tackled through this lens by Patrascu & Necoara (2017). Xu (2018) also studies the minimization of a stochastic objective controlled by a very large number of stochastic functional constraints, with application to stochastic linear programming. Finally, put forth by Fercoq et al. (2019), extensions to situations where the number of constraints is unknown (e.g. online settings) can be modeled by a template highly similar to (1), thus addressing important applications such as online portfolio optimization.

In this paper, we are interested in a class of applications which can benefit from being cast under template (1), namely semidefinite programs (SDPs) with a large number of linear constraints, such as arise in combinatorial optimization. A prominent example in machine learning is the \( k \)-means clustering problem, whose SDP relaxation comprises \( \mathcal{O}(d^2) \) linear constraints where \( d \) is the number of data samples (Peng & Wei, 2007). Maximum a posteriori estimation (Huang et al., 2014), quadratic assignment (Burer & Monteiro, 2005), \( k \)-nearest neighbor classification (Weinberger & Saul, 2009) and Sparsest cut (Arora et al., 2009) are other relevant SDP instances with linear constraints of order \( \mathcal{O}(d^2) \) or \( \mathcal{O}(d^3) \). Coupled with large input dimensions, such SDPs become problematic for most existing methods, due to the high cost of processing the constraints in-full during optimization.

In contrast, casting such SDPs into (1) suggests a simple solution: treat the linear constraints stochastically by only accessing a random subset at each iteration, then solve (1) using cheap gradient methods. However, the bottleneck in executing this idea is that existing methods require the constraint \( \mathcal{X} \) to possess an efficient projection oracle, whereas projecting onto the semidefinite cone amounts to full singular value decompositions, an operation that is prohibitively expensive even when the problem dimension is moderate. We hence ask:

Does a scalable method exist for solving (1) when the set \( \mathcal{X} \) does not have an efficient projection oracle?

The present work resolves the above challenge in the pos-
ite. To this end, we borrow tools from the conditional gradient methods (CGM) (Frank & Wolfe, 1956; Jaggi, 2013), which rely on the generally cheaper linear minimization oracles (lmo), rather than their projection counterparts. In particular, as the Lanczos method enables an efficient lmo computation for the spectrahedron (Arora et al., 2005), CGMs have already been proposed for solving SDPs (Jaggi, 2013; Garber & Hazan, 2016; Yurtsever et al., 2018; Locatello et al., 2019). However, none of these methods can handle the constraints stochastically.

In a nutshell, our approach relies on homotopy smoothing of the stochastic constraints in conjunction with CGM steps and a carefully chosen variance reduction procedure. Our analysis gives rise to two fully stochastic algorithms for solving problem (1) without projections onto $\mathcal{X}$. The first of the methods, H-SFW1, relies on a single sample (or fixed batch size) for computing the variance-reduced gradient and converges at a cost of $O(\varepsilon^{-6})$ lmo calls and $O(\varepsilon^{-6})$ stochastic first-order oracle (sfo) calls. The second, H-SPIDERFW, uses batches of increasing size under the SPIDER variance reduction scheme (Fang et al., 2018) and attains a theoretical complexity of $O(\varepsilon^{-2})$ lmo calls and $O(\varepsilon^{-4})$ sfo calls. The difference in convergence rates emphasizes the trade-off between the computational cost per-iteration and the number of iterations required to reach the constrained optimum.

2. Related Work

The present work lies at the intersection of several lines of research, whose relevant literature we describe in the following sections.

Proximal Methods for Almost Sure Constraints. Problems of similar formulation to (1) have been addressed in prior literature under the assumption of an efficient projection oracle over $\mathcal{X}$. Works such as (Patrascu & Necoara, 2017; Xu, 2018; Fercoq et al., 2019) solve these problems via stochastic proximal methods and attain a complexity of $O(\varepsilon^{-2})$ sfo calls, which is known to be optimal even for unconstrained stochastic optimization. In particular, Patrascu & Necoara (2017) study convex constrained optimization, where the constraints are expressed as a (possibly infinite) intersection of stochastic, closed, convex and projectable sets $X_\xi$. Problem (1) can be partly cast to this template, with $A(\xi)X \in b(\xi)$ being the homologues of $X_\xi$. However, our additional set $\mathcal{X}$ does not allow for efficient projections, making this framework inapplicable.

Xu (2018) solves a convex constrained optimization problem over a convex set $\mathcal{X}$, subject to a large number of convex functional constraints $f_j$, $j = 1 \ldots M$. The functions $f_j$ are sampled uniformly at random during optimization, which corresponds to a finitely sampled instance of problem (1) for affine $f_j$. However, we meet again with the limiting condition that projections onto $\mathcal{X}$ are computationally expensive in our setting.

Finally, Fercoq et al. (2019) study convex problems subject to a possibly infinite number of almost sure linear inclusion constraints, a template which closely resembles ours. The limitation, however, lies in their inclusion of a proximal-friendly component in the objective used to perform stochastic proximal gradient steps. This assumption does not hold for our problem formulation.

Conditional Gradient Methods for Constrained Optimization. CGM was first proposed in the seminal work of Frank & Wolfe (1956) and its academic interest has witnessed a resurgence in the past decade. The advantage of CGMs lies in the low per-iteration cost of the lmo, alongside their ability to produce sparse solutions. In comparison to projection-based approaches, the lmo is cheaper to compute for several important domains, amongst which the spectrahedron, polytopes emerging from combinatorial optimization, and $\ell_p$ norm-induced balls (Garber, 2016). Consequently, CG-type methods have been studied under varying assumptions in (Hazan, 2008; Clarkson, 2010; Hazan & Kale, 2012; Jaggi, 2013; Lan, 2013; Balasubramanian & Ghadimi, 2018), and have been incorporated as cheaper sub-solvers into algorithms which originally relied on projection oracles (Lan & Zhou, 2016; Liu et al., 2019).

CGMs have been further extended to the setting of convex composite minimization via the Augmented Lagrangian framework in (Gidel et al., 2018; Silvetti-Falls et al., 2019; Yurtsever et al., 2019a). Most relevant to our work, CGM-based quadratic penalty methods have been studied for convex problems with constraints of the form $Ax - b \in \mathcal{K}$, where $\mathcal{K}$ is a closed, convex set (Yurtsever et al., 2018; Locatello et al., 2019). We compare our methods against the latter two in Section 4.5.

Variance Reduction. Stochastic variance reduction (VR) methods have gained popularity in recent years following their initial study by (Roux et al., 2012; Johnson & Zhang, 2013; Mahdavi et al., 2013). The VR technique relies on averaging schemes to reduce the variance inherent to stochastic gradients, with several different flavors having emerged in the past decade: SAG (Schmidt et al., 2017), SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014), SVRG++ (Allen-Zhu & Yuan, 2016), SARAH (Nguyen et al., 2017) and SPIDER (Fang et al., 2018). Such methods outperform the classical SGD under the finite sum model, a fact which led to their widespread use in large-scale applications and their further inclusion into other stochastic optimization algorithms (see for example (Xiao & Zhang, 2014; Hazan & Luo, 2016)).
Relevant to our setting, VR has been studied in the context of CGMs for convex minimization by (Mokhtari et al., 2018; Hazan & Luo, 2016; Locatello et al., 2019; Yurtsever et al., 2019b; Zhang et al., 2019). The sfo complexity of these methods varies depending on the VR scheme, with the best guarantee being of order $O(\epsilon^{-2})$ (Zhang et al., 2019). For a thorough comparison of the complexities, we refer the reader to Section 6 of (Yurtsever et al., 2019b).

3. Preliminaries

Notation. We use $\| \cdot \|$ to express the Euclidean norm and $\langle \cdot , \cdot \rangle$ to denote the corresponding inner product. The distance between a point $x$ and a set $X$ is defined as $\text{dist}(x, X) := \inf_{y \in X} \| y - x \|$. The indicator function of a set $X$ is given by $\delta_X(x) = 0$, if $x \in X$, and $\delta_X(x) = +\infty$ otherwise. We denote by $D_X := \max_{(y, z) \in X \times X} \| y - z \|$ the diameter of a compact set $X$.

For the probabilistic setting, we denote by $\xi$ an element of our sample space and by $P(\xi)$ its probability measure. Unless stated otherwise, expectations will be taken with respect to $\xi$. We use $[n]$ to denote $\{1, 2, \ldots , n\}$. Given a function $f : \mathbb{R}^d \to \mathbb{R}$ and $L > 0$, we say that $f$ is $L$-smooth if $\nabla f$ is Lipschitz continuous, which is defined as $\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|, \forall x, y \in \mathbb{R}^d$.

Following the same setup as in (Fercoq et al., 2019), the space of random variables used in this work is

$$\mathcal{H} = \{ y(\xi) \in \mathbb{R}^m \mid \xi \in \mathbb{R}^n, \mathbb{E} \left[ \| y(\xi) \|_2^2 \right] < +\infty \},$$

where the associated scalar product is given by $\langle x, z \rangle := \mathbb{E} \left[ x^T z(\xi) \right] = \int x^T z(\xi) dP(\xi)$.

Smoothing. Nesterov (2005) proposes a technique for obtaining smooth approximations parametrized by $\beta$, of a nonsmooth and convex function $g$. The resulting smoothed approximations take the following form:

$$g_\beta(x) = \max_y \langle y, x \rangle - g^*(y) - \frac{\beta}{2} \| y \|^2,$$

where $g^*(y) = \sup_z \langle z, y \rangle - g(z)$ is the Fenchel conjugate of $g$. Note that $g_\beta$ is convex and $\frac{\beta}{2}$-smooth. The present work focuses on the case when $g(\cdot, \xi) = \delta_{b(\xi)}(\cdot)$. Smoothing the indicator function is studied in the context of proximal methods by Tran-Dinh et al. (2018); Fercoq et al. (2019) and for deterministic CGM by Yurtsever et al. (2018). Of particular note is that when $g(x) = \delta_X(x)$, the smoothed function becomes $g_\beta(x) = \frac{1}{2\beta} \text{dist}(x, X)^2$.

Optimality Conditions. We denote by $x^*$ a solution to problem (1) and say that $x$ is an $\epsilon$-solution for (1) if it satisfies

$$\mathbb{E} \left[ \| f(x, \xi) - f(x^*) \| \right] \leq \epsilon, \quad \sqrt{\mathbb{E} \left[ \text{dist}(A(\xi)x, b(\xi))^2 \right]} \leq \epsilon. \quad (2)$$

Oracles. Our complexity results are given relative to the following oracles:

- **Stochastic first order oracle (sfo):** For a stochastic function $\mathbb{E} [ f(x, \xi) ]$ with $\xi \sim P$, the sfo returns a pair $(f(x, \xi), \nabla f(x, \xi))$ where $\xi$ is an i.i.d. sample from $P$ (Nemirovsky & Yudin, 1983).
- **Incremental first order oracle (ifo):** For finite-sum problems, the ifo takes an index $i \in [n]$ and returns a pair $(f_i(x), \nabla f_i(x))$.
- **Linear minimization oracle (lmo):** The linear minimization oracle of set $\mathcal{X}$ is given by $\text{lm}o_{\mathcal{X}}(y) = \arg \min_{x \in \mathcal{X}} \langle x, y \rangle$ and is assumed to be efficient to compute throughout this paper. This is the main projection-free oracle model for CGM-type methods.

4. Algorithms & Convergence

We now describe our proposed methods for solving (1), H-1SFW and H-SPIIDER-FW, and provide their theoretical convergence guarantees.

4.1. Challenges and High-Level Ideas

Problem (1) can be rewritten equivalently as:

$$\min_{x \in \mathcal{X}} F(x) := \mathbb{E} \left[ f(x, \xi) + \delta_{b(\xi)}(A(\xi)x) \right]. \quad (3)$$

Note that, in this form, our objective is non-smooth due to the indicator function. In order to leverage the conditional gradient framework, we smooth $\delta_{b(\xi)}(A(\xi)x)$ through the technique described in Section 3, thus obtaining a surrogate objective $F_{\beta}$. For notational simplicity, we refer to the smoothed stochastic indicator as:

$$g_\beta(A(\xi)x) = \frac{1}{2\beta} \text{dist}(A(\xi)x, b(\xi))^2. \quad (4)$$

The minimization problem in terms of the smoothed objective thus becomes:

$$\min_{x \in \mathcal{X}} F_{\beta}(x) := \mathbb{E} \left[ f(x, \xi) + g_\beta(A(\xi)x) \right], \quad (5)$$

with $\lim_{\beta \to 0} F_{\beta}(x) = F(x)$. A natural idea is to optimize smooth approximations $F_{\beta}$ which are progressively more accurate representations of $F$. To this end, we apply conditional gradient steps in conjunction with decreasing the smoothness parameter $\beta$, practically emulating a homotopy algorithm.
transformation. As the iterations unfold our algorithms in fact approach the optimum of the original objective function \( F(x) \), as stated theoretically in Sections 4.3.2 and 4.4.2.

However, the aforementioned idea faces a technical challenge: decreasing the smoothing parameter \( \beta \) impacts the variance of the stochastic gradients \( \nabla_x g_\beta(A(\xi)x) \), which increases proportionally. This issue has previously been signaled in the work of (Fercoq et al., 2019), where the authors address a similar setting using stochastic proximal gradient steps. Here, the problem is further aggravated by the use of lmo calls over \( X \), as it is well-known that CGMs are sensitive to non-vanishing gradient noise (Mokhtari et al., 2018).

Our solution is to simply perform VR on the stochastic gradients and theoretically establish a rate for \( \beta \to 0 \) in order to counteract the exploding variance. Precisely, we show how two different VR schemes can be successfully used within the homotopy framework:

- **H-1SFW** uses one stochastic sample to update a gradient estimator at every iteration, following the technique introduced in (Mokhtari et al., 2018). Depending on computational resources, the single-sample model can be extended to a fixed batch size with the same convergence guarantees.

- **H-SPIDER-FW** uses stochastic minibatches of increasing size to compute the gradient estimator, using the technique proposed in (Fang et al., 2018).

Theoretical results characterizing our algorithms are presented in sections refsec:1sfw and 4.4. First, we state the rate at which the \( \beta \)-dependent stochastic gradient noise vanishes under each VR scheme in lemmas 4.1 and 4.2. The main convergence theorems 4.1 and 4.2 then describe the performance of our algorithms in terms of the quantity \( \mathbb{E} \left[ S_{\beta_k}(x_k, \xi) \right] := \mathbb{E} \left[ F_{\beta_k}(x_k, \xi) - F(x^*) \right] \), called the *smoothed gap*. Finally, in corollaries 4.1 and 4.2 we translate the aforementioned results into guarantees over the objective residual and constraint feasibility. All proofs are deferred to the appendix due to lack of space.

### 4.2. Technical Assumptions

**Assumption 4.1.** The stochastic functions \( f(\cdot, \xi) \) are convex and \( L_f \)-smooth. This further implies that \( f(x) \) is \( L_f \)-smooth.

**Assumption 4.2.** The stochastic gradients \( \nabla f(x, \xi) \) are unbiased and have a uniform variance bound \( \sigma_f^2 \). Formally,

\[
\mathbb{E} \left[ \nabla f(x, \xi) \right] = \nabla f(x)
\]

\[
\mathbb{E} \left[ ||\nabla f(x, \xi) - \nabla f(x)||^2 \right] \leq \sigma_f^2 < +\infty.
\]

**Algorithm 1 H-1SFW**

**Input:** \( x_1 \in X, \rho_0 > 0, P(\xi) \)

**for** \( k = 1, 2, \ldots \)**

- Set \( \rho_k, \beta_k \) and \( \gamma_k \); sample \( \xi_k \sim P(\xi) \)
- \( d_k = (1 - \rho_k)d_{k-1} + \rho_k \nabla F_{\beta_k}(x_k, \xi_k) \)
- \( w_k = \text{lmo}_X(d_k) \)
- \( x_{k+1} = x_k + \gamma_k (w_k - x_k) \)

**end for**

**Assumption 4.3.** The domain \( X \) is convex and compact, with diameter \( \mathcal{D}_X \).

**Assumption 4.4.** Slater’s condition holds for problem (3).

Specifically, letting \( G : \mathcal{H} \to \mathbb{R} \cup \{ \infty \}, G(Ax) := \mathbb{E} \left[ \delta_{[\xi]}(A(\xi)x) \right] \), with the linear operator \( A : \mathbb{R}^d \to \mathcal{H} \)

defined as \( (Ax)(\xi) := A(\xi)x, \forall x \), we require that

\[
0 \in sri(\text{dom}(G) - A \text{dom}(f)),
\]

where \( sri \) is the strong relative interior of the set (Bauschke et al.).

**Assumption 4.5.** The spectral norm of the stochastic linear operator \( A(\xi) \) is uniformly bounded by a constant \( L_A \):

\[
L_A := \sup_{\xi} \| A(\xi) \|^2 < +\infty.
\]

This assumption is also made in (Fercoq et al., 2019).

### 4.3. Homotopy-1SFW

We now describe our first algorithm which relies on the VR scheme proposed in (Mokhtari et al., 2018), and whose advantage lies in a simple update rule and single-loop structure.

#### 4.3.1. Gradient Estimator Model

We denote the gradient estimator by \( d_k \), and remark that it is biased with respect to the true gradient \( \nabla F_{\beta_k}(x_k) \) and exhibits a vanishing variance. This scheme achieves VR while conveniently considering only one stochastic constraint at a time. The estimator update rule is given by

\[
d_k = (1 - \rho_k)d_{k-1} + \rho_k \nabla F_{\beta_k}(x_k, \xi_k),
\]

where \( \nabla F_{\beta_k}(x_k, \xi_k) = \nabla f(x_k, \xi_k) + \nabla g_{\beta_k}(A(\xi_k)X_k) \), and \( \rho_k \) is a decaying convex combination parameter. The proposed method is provided via pseudocode in Algorithm 1.

#### 4.3.2. Convergence Results

Before stating the results, we remark that Lemma 4.1 is the counterpart of Lemma 1 in (Mokhtari et al., 2018) and its
proof follows a similar route, up to bounding \( \beta \)-dependent quantities. It is worth noting that in our case, handling the stochastic linear inclusion constraints results in a rate surcharge factor of \( \mathcal{O}(k^{1/3}) \).

**Lemma 4.1.** Let \( \rho_k = \frac{3}{(k+5)^{2/3}}, \quad \gamma_k = \frac{2}{k+1}, \quad \beta_k = \frac{\beta_0}{(k+1)^{1/\sigma}}, \quad \beta_0 > 0 \) in Algorithm 1. Then, for all \( k \),

\[
\mathbb{E} \left[ \| \nabla F_{\beta_k}(x_k) - d_k \|^2 \right] \leq \frac{C_1}{(k+5)^{1/3}},
\]

where \( C_1 = \max \left\{ 6^{1/3}\| \nabla F_{\beta_0}(x_0) - d_0 \|^2, 2 \left[ 18\gamma_1^2 + \frac{112\gamma_1^2 \mathcal{D}_X^2}{\beta_0^2} \right] \right\} \).

**Theorem 4.1.** Consider Algorithm 1 with parameters \( \rho_k = \frac{3}{(k+5)^{2/3}}, \quad \gamma_k = \frac{2}{k+1}, \quad \beta_k = \frac{\beta_0}{(k+1)^{1/\sigma}}, \quad \beta_0 > 0 \) (identical to Lemma 4.1). Then, for all \( k \),

\[
\mathbb{E} \left[ S_{\beta_k}(x_{k+1}) \right] \leq \frac{C_2}{k^{1/6}},
\]

where \( C_2 = \max \left\{ S_0(x_1), b = 2\mathcal{D}_X \sqrt{C_1} + 2\mathcal{D}_X \left( L_f + \frac{L_f}{\beta_0^2} \right) \right\} \) and \( C_1 \) is defined in Lemma 4.1.

**Corollary 4.1.** The expected convergence in terms of objective suboptimality and feasibility of Algorithm 1 is, respectively,

\[
\mathbb{E} \left[ \| f(x_k, \xi) - f(x^*) \| \right] \in \mathcal{O}(k^{-1/6})
\]

\[
\mathbb{E} \left[ \text{dist}(A(\xi)x_k, b(\xi))^2 \right] \in \mathcal{O}(k^{-1/6}).
\]

Consequently, the oracle complexity is \#(sfo) \( \in \mathcal{O}(e^{-6}) \) and \#(imo) \( \in \mathcal{O}(e^{-6}) \).

### 4.4. H(omotopy)-SPIDER-FW

Our second algorithm presents a more complex VR scheme, which improves on the complexity of H-1SFW. The method relies on the SPIDER estimator originally proposed under the framework of Normalized Gradient Descent in (Fang et al., 2018) and further studied for CGMs in (Yurtsever et al., 2019b). Different from Section 4.3.2, the results that follow distinguish two scenarios: the first is customary to VR methods such as SVRG (Johnson & Zhang, 2013) or SARAH (Nguyen et al., 2017) and assumes a finite-sum form of \( f \); the second, different from most other VR schemes, caters to objectives of the form \( f(x) = \mathbb{E}[f(x, \xi)] \) where \( \xi \sim P(\xi) \), and can handle a potentially infinite number of stochastic functions of \( f \).

**Algorithm 2** H-SPIDER-FW

**Input:** \( x_1 \in \mathcal{X}, \beta_0 > 0, P(\xi) \)

**for** \( t = 1, 2, \ldots, T \) **do**

\( x_{t+1} = \bar{x}_t \)

Compute \( \gamma_{t,1}, \beta_{t,1}, K_t \); sample \( \zeta \mathcal{Q}_t \sim P(\xi) \)

\( v_{t,1} = \tilde{\nabla} F_{\beta_{t,1}}(x_{t,1}, \zeta \mathcal{Q}_t) \)

\( w_{t,1} \in \text{imo}(v_{t,1}) \)

\( x_{t,2} = x_{t,1} + \gamma_{t,1}(w_{t,1} - x_{t,1}) \)

**for** \( k = 2, \ldots, K_t \) **do**

Compute \( \gamma_{t,k}, \beta_{t,k}; \) sample \( \zeta \mathcal{S}_{t,k} \sim P(\xi) \)

\( v_{t,k} = v_{t,k-1} - \tilde{\nabla} F_{\beta_{t,k-1}}(x_{t,k-1}, \zeta \mathcal{S}_{t,k}) \)

\( w_{t,k} \in \text{imo}(v_{t,k}) \)

\( x_{t,k+1} = x_{t,k} + \gamma_{t,k}(w_{t,k} - x_{t,k}) \)

**end for**

Set \( \bar{x}_{t+1} = x_{t,K_t+1} \)

**end for**

#### 4.4.1. GRADIENT ESTIMATOR MODEL

We denote the SPIDER gradient estimator by \( v_{t,k} \), and remark that it is also biased relative to \( \nabla F_{\beta_k}(x_k) \) and exhibits a vanishing variance. This scheme achieves VR through the use of increasing-size minibatches. The estimator update rule is given by

\[
v_{t,k} = v_{t,k-1} - \tilde{\nabla} F_{\beta_{t,k-1}}(x_{t,k-1}, \zeta \mathcal{S}_{t,k}) + \tilde{\nabla} F_{\beta_{t,k}}(x_{t,k}, \zeta \mathcal{S}_{t,k})
\]

where

\[
\tilde{\nabla} F_{\beta_{t,k}}(x_{t,k}, \zeta \mathcal{S}_{t,k}) = \tilde{\nabla} f(x_k, \zeta \mathcal{S}_{t,k}) + \nabla g_{\beta_{t,k}}(A(\mathcal{S}_{t,k})x_{t,k})
\]

defines the averaged gradient over a minibatch of size \( |\mathcal{S}_{t,k}| \).

The double indexing used in (7) hints at the double-loop structure of the algorithm, a format similar to most VR-based methods. The method is structured similarly to SPIDER-FW from (Yurtsever et al., 2019b), and proceeds in two steps: the outer loop computes an accurate gradient estimator and sets the batch size for the inner iterations. The inner-loop then iteratively ‘refreshes’ this gradient according to (7) and performs homotopy steps on \( \beta \) using a theoretically-determined schedule. The proposed method is provided via pseudocode in Algorithm 2.

#### 4.4.2. CONVERGENCE RESULTS

Again, we remark that Lemma 4.2 is the counterpart of Lemma 4, Appendix C in (Yurtsever et al., 2019b). However in this case, our proof takes a different, more tedious route, as the latter result does not accommodate homotopy steps.
In comparison, the bound we obtain depends linearly on the total iteration count, whereas the lemma of (Yurtsever et al., 2019b) depends only on the outer loop counter \( K_t \).

**Lemma 4.2** (Estimator variance for finite-sum problems). Consider Algorithm 2, and let \( \xi \) be finitely sampled from set \([n]\), \( \xi_{Q_t} = [n] \) and \( \xi_{S_t,k} \), such that \( |S_{t,k}| = K_t = 2^{t-1} \). Also, let \( \gamma_t, k = \frac{2}{K_t + \tau}, \beta_t, k = \frac{\beta_0}{\sqrt{K_t + \tau}}, \beta_0 > 0 \). Then, for a fixed \( t \) and for all \( k \leq K_t \),

\[
\mathbb{E} \left[ \| \nabla F_{\beta_t, k}(x_{t,k}) - v_{t,k} \|^2 \right] \leq \frac{C_1}{K_t + k},
\]

where \( C_1 = 2D_X^2 \left( 8L_f^2 + \frac{98L_f^2}{\beta_0^2} \right) \).

**Lemma 4.3** (Estimator variance for general expectation problems). Consider Algorithm 2 and let \( \xi \sim P(\xi) \) and \( \xi_{Q_t} \), such that \( |Q_t| = \left[ \frac{2K_t}{\beta_0} \right] + 1 \). Also, let \( \xi_{S_t,k} \), such that \( |S_{t,k}| = K_t = 2^{t-1} \), \( \gamma_t, k = \frac{2}{K_t + \tau}, \beta_t, k = \frac{\beta_0}{\sqrt{K_t + \tau}}, \beta_0 > 0 \). Then, for a fixed \( t \) and for all \( k \leq K_t \),

\[
\mathbb{E} \left[ \| \nabla F_{\beta_t, k}(x_{t,k}) - v_{t,k} \|^2 \right] \leq \frac{C_2}{K_t + k},
\]

where \( C_2 = 16L_f^2D_X^2 + 2L_f^2D_X^2 \left( \frac{98}{\beta_0^2} + 1 \right) + 2\beta_0^2 \sigma_f^2 \).

**Theorem 4.2.** Consider Algorithm 2 with parameters

\[
\gamma_t, k = \frac{2}{K_t + \tau}, \beta_t, k = \frac{\beta_0}{\sqrt{K_t + \tau}}, \beta_0 > 0, \quad \text{and} \quad \xi_{S_t,k} \), such that \( |S_{t,k}| = K_t = 2^{t-1} \). Then,

- For \( \xi \) be finitely sampled from set \([n]\), \( \xi_{Q_t} = [n] \) and \( \forall t \in \mathbb{N}, 1 \leq k \leq 2^{t-1} \),

\[
\mathbb{E} \left[ |S_{\beta_t, k}(x_{t,k+1})| \right] \leq \frac{C_3}{\sqrt{K_t + k + 1}},
\]

where \( C_3 = \max \left\{ S_{\beta_{1,0}}(x_{1,1}), 2D_X^2L_f + 2D_X^2 \sqrt{16L_f^2 + \frac{196L_f^2}{\beta_0^2} + \frac{2D_X^2L_f}{\beta_0}} \right\} \).

- For \( \xi \sim P(\xi), \xi_{Q_t} \), such that \( |Q_t| = \left[ \frac{2K_t}{\beta_0} \right] + 1 \) and \( \forall t \in \mathbb{N}, 1 \leq k \leq 2^{t-1} \),

\[
\mathbb{E} \left[ |S_{\beta_t, k}(x_{t,k+1})| \right] \leq \frac{C_4}{\sqrt{K_t + k + 1}},
\]

where \( C_4 = \max \left\{ S_{\beta_{1,0}}(x_{1,1}), 2D_X^2L_f + \frac{2D_X^2L_A}{\beta_0} + 2D_X^2L_f + \frac{2D_X^2L_A}{\beta_0} \right\} \) and \( + 2D_X \sqrt{16L_f^2D_A^2 + 2L_f^2D_A^2 \left( \frac{98}{\beta_0^2} + 1 \right) + 2\beta_0^2 \sigma_f^2} \).
bounds of Lemmas 4.1 and 4.2-4.3, respectively: $\mathcal{O}(k^{-1/3})$ for $d_k$ vs. $\mathcal{O}(k^{-1})$ for $v_{t,k}$. This signals the existence of a trade-off: a more intricate way of handling stochastic penalty-type constraints can ensure the better convergence guarantees of H-SPIDER-FW, while a simpler VR scheme comes at the cost of the rather pessimistic ones of H-1SFW. Fortunately, as shown in the Section 5, the simple H-1SFW greatly outperforms its worst-case guarantees.

5. Numerical Experiments

For demonstrating the empirical efficiency of our algorithms, we apply them to three problem instances: synthetically-generated SDPs, the K-means clustering SDP relaxation and the Sparsest Cut-associated SDP.

Evaluation Metrics: Our experiments subscribe to a finite-sum template, where we define $f(x) := \sum_{i=1}^{n} f_i(x)$ and $g_i(Ax) = \sum_{j=1}^{n_{2}} g_{i,j}(A^T x)$. The objective convergence is recorded as $|f(x) - f^*|$, with $f^* := f(x^*)$. Due to imperfect feasibility, the value of $f(x)$ can overshoot $f^*$, since the constrained optimum is not the global one. This usually appears as the increase of $|f(x) - f^*|$ immediately after a significant drop when the quantity $f(x) - f^*$ becomes negative; then the decreasing trend restarts, as the objective and constraints re-balance. Such a phenomenon is common for homotopy-based methods, see for instance (Yurtsever et al., 2018). Lastly, the feasibility is recorded as $\|Ax - b\|$.  

Baseline: To the best of our knowledge, the HCGM (Yurtsever et al., 2018) and the SHCGM (Locatello et al., 2019) are the only algorithms which tackle SDPs under the conditional gradient framework. The latter represents the empirical state-of-the-art and we choose it as the baseline for our experiments.

5.1. Synthetic SDP Problems

This proof-of-concept experiment aims to show the performance of our fully stochastic methods, given a fixed problem dimension and an increasing set of constraints. We consider the synthetic SDP:

$$\begin{align*}
\min_{\mathbf{x} \in \mathbb{R}^d} & \quad \langle C, \mathbf{x} \rangle \\
\text{subject to} & \quad \mathbf{A} \mathbf{x} - \mathbf{b} \leq \frac{1}{2} \\
\end{align*}$$

where the entries of $\mathbf{A}_i$ and $\mathbf{C}$ are generated from $\mathcal{U}(0, 1)$, and $b_i = \langle A_i, X^* \rangle$ for a fixed $X^*$. We perform uniform sampling on the pairs $(A_i, b_i)$ for computing their stochastic gradients in our algorithms. We fix the dimension to be $d = 20$ and vary the size of constraints with $n = 5e2$ and $5e3$.

For a fair comparison, we sweep the parameter $\beta_0$ for the three algorithms in the range $[1e-7, 1e1]$. We settle for $1e-7, 1e-7$ and $1e-5$ for SHCGM, H-1SFW and H-SPIDER-FW, respectively. For H-1SFW and SHCGM, we choose the batchsize to be 1% of the data.

Figure 1 illustrates the outcome of the experiments, where we observe a clear improvement of the stochastic algorithms over the baseline with a stable margin throughout the test cases.

Interestingly, H-1SFW exhibits strong empirical performance on the synthetic data, much better than its theoretical worst-case bound. A possible explanation is that the entries of $C$ and $A_i$ are generated from a “benign” distribution and concentrate around their mean (Ledoux, 2001). In such scenarios, even a small subset of constraints allows for effective variance reduction. For comparison, we provide an additional set of results for synthetic SDPs generated from a less well-behaved distribution in Appendix A.2. Nevertheless, we observe the same good performance of H-1SFW even with real data, in the next sections.

Regarding H-SPIDER-FW, we observe that the suboptimality and feasibility decrease at the rate $k^{-\frac{1}{2}}$ and $k^{-\frac{3}{4}}$, respectively, which is better than the worst-case bounds in Theorem 4.2.

5.2. The K-means Clustering Relaxation

We consider the unsupervised learning task of partitioning $d$ data points into $k$ clusters. We adopt the SDP formulation in (Peng & Wei, 2007), which amounts to solving:

$$\begin{align*}
\min_{\mathbf{X} \in \mathbb{R}^{d \times d}} & \quad \langle C, \mathbf{X} \rangle \\
\text{subject to} & \quad \mathbf{X} \mathbf{1} = \mathbf{1}, \quad 1 \leq i, j \leq d.
\end{align*}$$

Here, $C \in \mathbb{R}^{d \times d}$ is the Euclidean distance matrix of the $d$ data points, $\mathcal{X} = \{X \in \mathbb{R}^{d \times d} : X \succeq 0, \text{tr}(X) \leq k\}$, $\mathbf{1}$ is the all 1’s vector. Notice that the number of linear constraints in (8) is $O(d^2)$.

In order to compare against existing work, we adopt the MNIST dataset ($k = 10$) (LeCun & Cortes, 2010) with $d = 10^3$ samples and perform data preprocessing as in (Mixon et al., 2016). The very same setup appeared in several works (Mixon et al., 2016; Yurtsever et al., 2018; Locatello et al., 2019), with SHCGM (Locatello et al., 2019) showing the best practical performance.

We perform parameter sweeping on $\beta_0 \in [1e-7, 1e2]$ for H-1SFW and H-SPIDER-FW, and settle for $5e-2$ and $6e0$, respectively. For SHCGM, we adopt the same hyperparameter as in (Locatello et al., 2019). The batchsize for H-1SFW and SHCGM is set to 5%.

The comparison of our algorithms against SHCGM is reported in Figure 2. H-1SFW and H-SPIDER-FW converge...
Conditional gradient methods for stochastically constrained convex minimization

Figure 1. Synthetic SDPs, with each column showing the convergence in objective suboptimality (top) and in feasibility (bottom) for a specific problem. The left hand-side column corresponds to a problem with $5e2$ constraints, while the right hand-side one to a problem with $5e3$ constraints.

Figure 2. The K-means SDP relaxation, with convergence in objective suboptimality (top) and in feasibility (bottom).

Figure 3. The Sparsest Cut-associated SDP relaxation, where each column shows the convergence in objective suboptimality (top) and feasibility (bottom) for a specific problem. From left to right, the results correspond to graphs mammalia-primate-association-13, insecta-ant-colony1-day37 and insecta-ant-colony4-day10, sorted by increasing size.
Conditional gradient methods for stochastically constrained convex minimization

### Table 1. Details of the Network Repository (Rossi & Ahmed, 2015) graphs used in the experiments.

<table>
<thead>
<tr>
<th>Graph name</th>
<th></th>
<th></th>
<th>Avg. node degree</th>
<th>Max. node degree</th>
<th>USC SDP dimension</th>
<th>USC SDP # constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>mammalia-primate-association-13</td>
<td>25</td>
<td>181</td>
<td>14</td>
<td>19</td>
<td>$X \in \mathbb{R}^{25 \times 25}$</td>
<td>$\sim 6.90e3$</td>
</tr>
<tr>
<td>insecta-ant-colony1-day37</td>
<td>55</td>
<td>1k</td>
<td>42</td>
<td>53</td>
<td>$X \in \mathbb{R}^{55 \times 55}$</td>
<td>$\sim 7.87e4$</td>
</tr>
<tr>
<td>insecta-ant-colony4-day10</td>
<td>102</td>
<td>4k</td>
<td>79</td>
<td>99</td>
<td>$X \in \mathbb{R}^{102 \times 102}$</td>
<td>$\sim 5.15e5$</td>
</tr>
</tbody>
</table>

at a comparable rate, with both clearly overtaking the baseline with regards to objective suboptimality and feasibility convergence.

### 5.3. Computing an $\ell_2^2$ Embedding for the Uniform Sparsest Cut Problem

The Uniform Sparsest Cut problem (USC) aims to find a bipartition $(S, \bar{S})$ of the nodes of a graph $G = (V, E)$, $|V| = d$, which minimizes the quantity

$$\frac{E(S, \bar{S})}{|S||\bar{S}|},$$

where $E(S, \bar{S})$ is the number of edges connecting $S$ and $\bar{S}$. This problem is of broad interest, with applications in areas such as VLSI layout design, topological design of communication networks and image segmentation, to name a few. Relevant to machine learning, it appears as a subproblem in hierarchical clustering algorithms (Dasgupta, 2016; Chatziafratis et al., 2018).

Computing such a bipartition is NP-hard and intense research has gone into designing efficient approximation algorithms for this problem. In the seminal work of Arora et al. (2009) an $O(\sqrt{\log d})$ approximation algorithm is proposed for solving USC, which relies on finding a well-spread $\ell_2^2$ geometric representation of $G$ where each node $i \in V$ is mapped to a vector $v_i$ in $\mathbb{R}^d$. In this experimental section we focus on solving the SDP that computes this geometric embedding, as its high number of triangle inequality constraints ($O(n^3)$) vs $O(n^2)$, H-SPIDER-FW’s increasing minibatches readily reach an adequate balance between feasibility enforcement and objective minimization.

**Acknowledgements**

The authors are grateful to Mehmet Fatih Sahin and Alp Yurtsever for the helpful discussions throughout the development of this paper.

This work was partially supported by the Swiss National Science Foundation (SNSF) under grant number 200021_178865/1; the Army Research Office under Grant Number W911NF-19-1-0404; the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no 725594 - time-data).

**References**


Conditional gradient methods for stochastically constrained convex minimization


Locatello, F., Yurtsever, A., Fercoq, O., and Cevher, V. Stochastic conditional gradient method for composite
Conditional gradient methods for stochastically constrained convex minimization


