# A Stochastic-Gradient-based Interior-Point Algorithm for Solving Smooth Bound-Constrained Optimization Problems

Frank E. Curtis

Department of Industrial and Systems Engineering, Lehigh University

Vyacheslav Kungurtsev

Department of Computer Science, Czech Technical University

Daniel P. Robinson

Department of Industrial and Systems Engineering, Lehigh University

Qi Wang

Department of Industrial and Systems Engineering, Lehigh University

COR@L Technical Report 23T-018

# A Stochastic-Gradient-based Interior-Point Algorithm for Solving Smooth Bound-Constrained Optimization Problems

Frank E. Curtis*[1], Vyacheslav Kungurtsev†[2], Daniel P. Robinson‡[3], and Qi Wang§[4]

[1]Department of Industrial and Systems Engineering, Lehigh University
[2]Department of Computer Science, Czech Technical University
[3]Department of Industrial and Systems Engineering, Lehigh University
[4]Department of Industrial and Systems Engineering, Lehigh University

April 28, 2023

**Abstract**

A stochastic-gradient-based interior-point algorithm for minimizing a continuously differentiable objective function (that may be nonconvex) subject to bound constraints is presented, analyzed, and demonstrated through experimental results. The algorithm is unique from other interior-point methods for solving smooth (nonconvex) optimization problems since the search directions are computed using stochastic gradient estimates. It is also unique in its use of inner neighborhoods of the feasible region—defined by a positive and vanishing neighborhood-parameter sequence—in which the iterates are forced to remain. It is shown that with a careful balance between the barrier, step-size, and neighborhood sequences, the proposed algorithm satisfies convergence guarantees in both deterministic and stochastic settings. The results of numerical experiments show that in both settings the algorithm can outperform a projected-(stochastic)-gradient method.

## 1 Introduction

The interior-point methodology is one of the most effective approaches for solving continuous constrained optimization problems. In the context of (deterministic) derivative-based algorithmic strategies, interior-point methods offer convergence guarantees from remote starting points [11, 21, 27], and in both convex and nonconvex settings such algorithms can offer good worst-case iteration complexity properties [7, 21]. Furthermore, many of the most popular software packages for solving large-scale continuous optimization problems are based on interior-point methods [1, 11, 24, 25, 26, 27], and these have been used to great effect for many years.

Despite the extensive literature on theoretical and practical benefits of interior-point methods in the context of (deterministic) derivative-based algorithms for solving (non)convex optimization problems, to the best of our knowledge there has not yet been one that has been shown rigorously to offer convergence guarantees when neither function nor derivative evaluations are available, and instead only stochastic gradient estimates are employed. (An interior-point stochastic-approximation method was proposed and tested in [12],

*E-mail: frank.e.curtis@lehigh.edu
†E-mail: kunguvya@fel.cvut.cz
‡E-mail: daniel.p.robinson@lehigh.edu
§E-mail: qiw420@lehigh.edu

but as we mention in Remark 3.3 on page 14, the claimed asymptotic-convergence guarantee in [12] overlooks a critical issue related to the step sizes.) In this paper, we propose, analyze, and present the results of experiments with such an algorithm. Randomized algorithms for minimizing a linear function over a convex set have been proposed [2, 19], but the setting and the techniques those algorithms employ are distinct from the ones considered in this paper.

For a straightforward presentation of our proposed strategy and its convergence guarantees, we focus on the case of constrained optimization with bound constraints only. That said, our algorithmic strategies have been designed so that they may be extended for solving problems with continuous (potentially nonlinear) equality and/or inequality constraints as well. For example, since interior-point methods handle inequality constraints through the introduction of an additional objective function term that is weighted by a barrier parameter and a continuation strategy that reduces the barrier parameter iteratively, one might consider extending our algorithmic ideas using the recently proposed stochastic algorithms for solving equality-constrained optimization presented in [4, 5, 3, 6, 15, 14, 16, 17, 18, 23]. The main challenge to address in such potential extensions is the one that we address in this paper, namely, that derivatives of the barrier function are not Lipschitz continuous in the interior of a set of bound constraints. We focus primarily on the setting of minimizing an objective that may be nonconvex. Upon seeing our algorithm, a reader may wonder if a simpler variant has convergence guarantees. However, we discuss in Section 4 why the challenges that we overcome in the general (potentially nonconvex) setting are not readily avoided with a simpler variant, even in the strongly convex setting.

## 1.1 Contributions

We propose, analyze, and provide the results of numerical experiments with a stochastic-gradient-based interior-point method for solving (potentially nonconvex) bound-constrained optimization problems. Although not considered in this paper, our proposed algorithm can form the basis for a variety of algorithms for solving problems with nonlinear equality and inequality constraints. Our algorithm involves multiple unique features compared to other derivative-based interior-point methods that have been proposed and analyzed in the literature. Overall, the main contributions of our proposed algorithm, analysis, and experiments are the following.

(i) Our algorithm employs a prescribed monotonically nonincreasing and vanishing barrier parameter sequence. In this manner, the algorithm does not rely on the ability to compute values for derivative-based stationarity tests, as is done in derivative-based interior-point methods for deciding whether to decrease the barrier parameter in a given iteration. This is significant since stationarity measures cannot be computed accurately in the stochastic setting that we consider.

(ii) Our algorithm does not employ a fraction-to-the-boundary rule. Such a rule is critical for convergence guarantees of other derivative-based interior-point methods for solving nonconvex problems (see, e.g., [11, 27]), since it ensures that—with respect to a threshold that depends on the barrier parameter value—the iterates do not get too close to the boundary of the feasible region. This, in turn, ensures that the iterates remain in a region in which derivatives of the barrier function are Lipschitz continuous. By contrast, in our proposed algorithm, we present a unique strategy that involves keeping each iterate within an inner neighborhood of the feasible region. Over a run of the algorithm, these neighborhoods are defined by a monotonically nonincreasing and vanishing sequence.

(iii) Our algorithm does not rely on step acceptance criteria (e.g., using line searches, trust regions, regularization, etc.) that in turn rely on exact objective function evaluations. This is significant since such evaluations are intractable in various settings of interest; see, e.g., [9]. That said, as is common for other stochastic optimization algorithms, our convergence guarantees rely on knowledge of problem-dependent quantities, including a Lipschitz constant for the gradient of the objective. (In practice, the problem-dependent quantities that our algorithm requires can be estimated using (stochastic) function and/or derivative values.)

(iv) We present general sets of conditions under which the algorithm's barrier, neighborhood, and step size sequences are balanced so as to ensure convergence guarantees in both deterministic and stochastic settings.

(v) We show by a representative comparison that, in deterministic and stochastic settings, the algorithm can outperform a projected-(stochastic)-gradient method.

One aspect that limits the applicability of our work is that, for the stochastic setting, we assume that the errors of the stochastic gradient estimators are bounded by a known value. Convergence guarantees have been established under looser assumptions in unconstrained settings [9], but since unique challenges arise in the constrained setting, we consider ours a significant first step in the design and analysis of interior-point algorithms for constrained stochastic optimization.

## 1.2 Notation

We use $\mathbb{R}$ to denote the set of real numbers, $\overline{\mathbb{R}}$ to denote the set of extended-real numbers (i.e., $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$), and $\mathbb{R}_{\geq a}$ (resp., $\mathbb{R}_{>a}$, $\mathbb{R}_{<a}$, or $\mathbb{R}_{\leq a}$) to denote the set of real numbers greater than or equal to (resp., greater than, less than, or less than or equal to) $a \in \mathbb{R}$. We append a superscript to such a set to denote the space of vectors or matrices whose elements are restricted to the indicated set; e.g., we use $\mathbb{R}^n$ to denote the set of $n$-dimensional real vectors and $\mathbb{R}^{m \times n}$ to denote the set of $m$-by-$n$-dimensional real matrices. We use $\mathbb{S}^n \subset \mathbb{R}^{n \times n}$ to denote the set of $n$-by-$n$-dimensional real symmetric matrices. We use $\mathbb{N} := \{1, 2, \dots\}$ to denote the set of positive integers and, given $n \in \mathbb{N}$, we denote $[n] := \{1, \dots, n\}$.

Given $\mathcal{B} \subseteq \mathbb{R}^n$, we use $\text{int}(\mathcal{B})$ to denote the interior of $\mathcal{B}$. We use $\mathbb{1}$ to denote a vector of ones whose length is determined by the context in which it appears. Given $(A, B) \in \mathbb{S}^n \times \mathbb{S}^n$, we write $A \succeq B$ (resp., $A \succ B$) to indicate that $A - B \in \mathbb{S}^n$ is positive semidefinite (resp., positive definite). Given $l \in \overline{\mathbb{R}}^n$, we use $\Psi(l)$ to denote the extended-real-valued diagonal matrix whose $(i,i)$-element is equal to $l_i$ for all $i \in [n]$. Given a sequence of real-valued vectors $\{\mu_k\}$ and $\mathcal{M} \subseteq \mathbb{R}^n$, we write $\{\mu_k\} \subset \mathcal{M}$ to indicate that $\mu_k \in \mathcal{M}$ for all $k \in \mathbb{N}$. Moreover, for a real-number sequence, we write $\{\mu_k\} \searrow 0$ to indicate that (a) $\{\mu_k\} \subset \mathbb{R}_{>0}$, (b) $\{\mu_k\}$ is monotonically nonincreasing, and (c) the limit of $\{\mu_k\}$ is zero. Given sequences $\{a_k\} \subset \mathbb{R}_{\geq 0}$ and $\{b_k\} \subset \mathbb{R}_{>0}$, we write $\{a_k\} = \mathcal{O}(b_k)$ (resp., $\{a_k\} = \Theta(b_k)$) to indicate that there exists $C \in \mathbb{R}_{>0}$ (resp., $(c, C) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$) such that $a_k \leq C b_k$ (resp., $c b_k \leq a_k \leq C b_k$) for all sufficiently large $k \in \mathbb{N}$. Given such sequences, we write $\{a_k\} = o(b_k)$ to indicate that $\{a_k / b_k\} \to 0$. Notice that in this paragraph and throughout the paper we use a subscript either to indicate an element index (of a vector or matrix) or an index of a sequence. In all such cases, the meaning of a subscript is clear from the context.

The algorithm that we propose is iterative in the sense that any given run of the algorithm produces an iterate sequence (of real-valued vectors) $\{x_k\} \subset \mathbb{R}^n$. Like for the iterate sequence, we append a positive integer as a subscript for a quantity to denote its value during an iteration of an algorithm. Multiple subscripts are used in some cases, as needed; e.g., the $i$th element of the $k$th iterate $x_k \in \mathbb{R}^n$ is denoted as $x_{k,i} \in \mathbb{R}$ and the $(i,i)$-element of a matrix $H_k \in \mathbb{S}^n$ is denoted as $H_{k,i,i}$.

At times, we express algebraic operations using quantities with infinite magnitude, namely, $-\infty$ and $\infty$. In such cases, we adopt natural conventions. In particular, given $a \in \mathbb{R}$, we let $\infty - a = \infty$ and $a - (-\infty) = \infty$, and, given $a \in \mathbb{R}_{>0}$, we let $a \cdot \infty = \infty$ and $a / \infty = 0$. Given a pair of nonnegative extended-real-number-valued vectors $(a, b) \in \overline{\mathbb{R}}^n_{\geq 0} \times \overline{\mathbb{R}}^n_{\geq 0}$, we write $a \perp b$ to indicate that $a_i = 0$ and/or $b_i = 0$ for all $i \in [n]$.

## 1.3 Organization

Our main problem of interest, namely, minimizing a potentially nonconvex continuous function over a set of bound constraints, is stated formally along with a presentation of our main algorithm in Section 2. Our convergence analyses for this algorithm are presented in Section 3. In Section 4, we discuss the obstacles of proving a convergence guarantee for a simpler variant of our algorithm, even in the strongly convex setting. The results of numerical experiments are presented in Section 5 and concluding remarks are provided in Section 6.

## 2 Algorithm

Our main problem of interest is to minimize an objective function over a feasible region that we denote as $\mathcal{B} := [l, u] \equiv \{x \in \mathbb{R}^n : l \leq x \leq u\}$, where $(l, u) \in \overline{\mathbb{R}}^n \times \overline{\mathbb{R}}^n$ with $l_i < u_i$ for all $i \in [n]$. We assume that at least one element of $(l, u)$ is finite, so the problem is indeed constrained. For the sake of generality, we only require that the objective has as its domain an open set $\mathcal{B}^+$ containing $\mathcal{B}$. Denoting this objective function as $f : \mathcal{B}^+ \to \mathbb{R}$, we write our problem of interest as

$$\min_{x \in \mathbb{R}^n} \ f(x) \ \text{ subject to } \ x \in \mathcal{B} := [l, u]. \tag{1}$$

We make the following assumption pertaining to this (potentially nonconvex) $f$.

**Assumption 2.1.** *The objective function $f : \mathcal{B}^+ \to \mathbb{R}$ is continuously differentiable over $\mathcal{B}^+$, bounded below by $f_{\inf} \in \mathbb{R}$ over $\mathcal{B}$, and bounded above by $f_{\sup} \in \mathbb{R}$ over $\mathcal{B}$. In addition, its gradient function $\nabla f : \mathcal{B}^+ \to \mathbb{R}^n$ is Lipschitz continuous with respect to the 2-norm over $\mathcal{B}$ with constant $\ell_{\nabla f,\mathcal{B}} \in \mathbb{R}_{>0}$ and is bounded in 2-norm (resp., $\infty$-norm) over $\mathcal{B}$ by $\kappa_{\nabla f,\mathcal{B},2} \in \mathbb{R}_{>0}$ (resp., $\kappa_{\nabla f,\mathcal{B},\infty} \in \mathbb{R}_{>0}$).*

Assumption 2.1 is mostly standard. The nonstandard aspect is the assumption that $f$ is bounded above over $\mathcal{B}$; this is a relatively loose assumption to handle extreme cases in the stochastic setting. The existence of $\kappa_{\nabla f,\mathcal{B},\infty}$ follows from that of $\kappa_{\nabla f,\mathcal{B},2}$, and vice versa, but we define both for the sake of notational convenience.

Under Assumption 2.1, specifically under the assumption that $f$ is continuously differentiable over an open set containing the feasible region $\mathcal{B}$, it follows that if $x \in \mathbb{R}^n$ is a minimizer of (1), then there must exist $(y, z) \in \mathbb{R}^n \times \mathbb{R}^n$ such that $(x, y, z)$ satisfies the Karush-Kuhn-Tucker (KKT) conditions given by

$$\nabla f(x) - y + z = 0, \quad 0 \leq (x - l) \perp y \geq 0, \quad \text{and} \quad 0 \leq (u - x) \perp z \geq 0. \tag{2}$$

Defining the index sets of finite bounds as $\mathcal{L} := \{i \in [n] : l_i > -\infty\}$ and $\mathcal{U} := \{i \in [n] : u_i < \infty\}$, $x \in \mathbb{R}^n$ implies that $x_i - l_i = \infty > 0$ for all $i \in [n] \setminus \mathcal{L}$ and $u_i - x_i = \infty > 0$ for all $i \in [n] \setminus \mathcal{U}$, meaning that (2) and our definition of the operator $\perp$ in Section 1.2 require that $y_i = 0$ for all $i \in [n] \setminus \mathcal{L}$ and $z_i = 0$ for all $i \in [n] \setminus \mathcal{U}$.

Central ideas of the interior-point methodology are to replace inequality constraints with a parameterized barrier function in the objective, and to solve the original constrained optimization problem through a continuation approach by driving the barrier parameter to zero. For example, using a so-called log-barrier in the context of (1), this amounts to introducing the barrier parameter $\mu \in \mathbb{R}_{>0}$ and using the log-barrier-augmented function $\phi : \text{int}(\mathcal{B}) \times \mathbb{R}_{>0} \to \mathbb{R}$ given by

$$\phi(x, \mu) = f(x) - \mu \sum_{i \in \mathcal{L}} \log(x_i - l_i) - \mu \sum_{i \in \mathcal{U}} \log(u_i - x_i). \tag{3}$$

(We use a log-barrier function throughout the paper, although one might extend our algorithm and analysis for other barrier functions as well.) Given $\mu \in \mathbb{R}_{>0}$ and letting $\nabla_x \phi$ denote the gradient operator of $\phi$ with respect to its first argument, a minimizer of the barrier-augmented function $\phi(\cdot, \mu)$ over $\text{int}(\mathcal{B})$ must satisfy

$$0 = \nabla_x \phi(x, \mu) \equiv \nabla f(x) - \mu \Psi(x - l)^{-1} \mathbb{1} + \mu \Psi(u - x)^{-1} \mathbb{1}. \tag{4}$$

A traditional interior-point method for derivative-based nonconvex optimization would involve fixing the barrier parameter at a value $\mu \in \mathbb{R}_{>0}$, employing an unconstrained optimization method to minimize $\phi(\cdot, \mu)$ until an approximate stationarity tolerance is satisfied (with a safeguard such as a fraction-to-the-boundary rule to ensure that the iterates remain within $\text{int}(\mathcal{B})$), then reducing the barrier parameter and repeating the procedure in an iterative manner. However, for our purposes, we avoid the need to check a stationarity tolerance, since this would require an evaluation of a gradient of the objective (see (4)), which we presume is intractable to obtain.

Our proposed algorithm, by contrast, employs a prescribed positive barrier parameter sequence $\{\mu_k\} \searrow 0$ and a line-search-free strategy for generating the positive step size sequence $\{\alpha_k\} \subset \mathbb{R}_{>0}$. We state the algorithm in a generic manner, but in our analyses in Section 3 we reveal specific requirements that these sequences must satisfy to yield convergence guarantees in deterministic and stochastic settings. In addition, rather than rely on a safeguard such as a fraction-to-the-boundary rule—which presents challenges in terms of ensuring convergence guarantees in a stochastic setting since such a rule would enforce an iterate-dependent bound on the steps—our algorithm employs a rule that ensures that, for all $k \in \mathbb{N}$, the subsequent iterate remains sufficiently within $\text{int}(\mathcal{B})$ by a prescribed margin. For this, we introduce

$$\mathcal{N}_{[l,u]}(\theta) := \{x \in \mathbb{R}^n : l + \theta \leq x \leq u - \theta\}, \tag{5}$$

and, for all $k \in \mathbb{N}$, have the algorithm ensure $x_{k+1} \in \mathcal{N}_{[l,u]}(\theta_k)$ for some $\theta_k \in \mathbb{R}_{>0}$. The positive sequence $\{\theta_k\} \searrow 0$, base value $\theta_0 \in \mathbb{R}_{>\theta_1}$, and initial point $x_1 \in \mathbb{R}^n$ must be prescribed for each run of the algorithm, and the latter two must satisfy

$$x_1 \in \mathcal{N}_{[l,u]}(\theta_0) \ \text{ and } \ \theta_0 < \tfrac{\Delta}{2}, \ \text{ where } \ \Delta := \min\left\{\bar{\Delta}, \min_{i \in [n]}(u_i - l_i)\right\} \in \mathbb{R}_{>0} \tag{6}$$

for some $\bar{\Delta} \in \mathbb{R}_{>0}$, where $\bar{\Delta}$ is introduced merely to ensure that $\Delta$ is finite.

The search direction computation in our proposed algorithm is the main aspect that distinguishes between the deterministic and stochastic settings. Specifically, letting $g_k$ denote a stochastic gradient estimate computed with respect to $x_k$ (see Section 3.3), we denote the gradient (estimate) for the barrier-augmented function by

$$q_k := \begin{cases} \nabla f(x_k) - \mu_k \Psi(x_k - l)^{-1}\mathbb{1} + \mu_k \Psi(u - x_k)^{-1}\mathbb{1} & \text{(deterministic)} \\ g_k - \mu_k \Psi(x_k - l)^{-1}\mathbb{1} + \mu_k \Psi(u - x_k)^{-1}\mathbb{1} & \text{(stochastic).} \end{cases}$$

Then, for $(\lambda_{k,\min}, \lambda_{k,\max}) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ with $\lambda_{k,\min} \leq \lambda_{k,\max}$ and diagonal $H_k \in \mathbb{S}^n$ with $\lambda_{k,\max} I \succeq H_k \succeq \lambda_{k,\min} I$, the search direction $d_k \in \mathbb{R}^n$ is $d_k = -H_k^{-1} q_k$.

A complete statement of our proposed interior-point method (IPM) for solving problem (1) with prescribed parameter sequences (i.e., barrier-parameter sequence $\{\mu_k\}$, neighborhood-parameter sequence $\{\theta_k\}$, step-size-bound sequences $\{\alpha_{k,\max}\}$ and $\{\gamma_{k,\max}\}$, and eigenvalue-bound sequences $\{\lambda_{k,\min}\}$ and $\{\lambda_{k,\max}\}$) is stated as Algorithm 1. We have written Algorithm 1 in a generic manner that demonstrates flexibility in the required parameter sequences. Our analyses in the next section prescribe additional rules for these sequences that lead to convergence guarantees.

**Algorithm 1** IPM with Prescribed Parameter Sequences

```
Require: $\{\mu_k\} \searrow 0$; $\{\theta_k\} \searrow 0$; $\{\alpha_{k,\max}\} \subset \overline{\mathbb{R}}_{>0}$; $\{\gamma_{k,\max}\} \subset (0,1]$; $\{\lambda_{k,\min}\} \subset \mathbb{R}_{>0}$ and $\{\lambda_{k,\max}\} \subset \mathbb{R}_{>0}$
    such that $\lambda_{k,\min} \leq \lambda_{k,\max}$ for all $k \in \mathbb{N}$; and $x_1 \in \mathcal{N}_{[l,u]}(\theta_0)$ for some $\theta_0 \in \mathbb{R}_{>\theta_1}$ satisfying (6)
for $k = 1, 2, \dots$ do
    choose diagonal $H_k \in \mathbb{S}^n$ such that $\lambda_{k,\max} I \succeq H_k \succeq \lambda_{k,\min} I$
    compute $d_k \leftarrow -H_k^{-1} q_k$
    choose $\alpha_k \in (0, \alpha_{k,\max}]$
    compute $\gamma_k \leftarrow \max\{\gamma \in (0, \gamma_{k,\max}] : x_k + \gamma \alpha_k d_k \in \mathcal{N}_{[l,u]}(\theta_k)\}$
    set $x_{k+1} \leftarrow x_k + \gamma_k \alpha_k d_k$
end for
```

**Remark 2.1.** *One could extend our algorithm and analysis to allow, for all $k \in \mathbb{N}$, the employment of non-diagonal $H_k$ and/or the option to set $x_{k+1}$ by searching further along the piecewise linear path defined by the projections of $x_k + \gamma \alpha_k d_k$ onto $\mathcal{N}_{[l,u]}(\theta_k)$ over $\gamma \in (0, \gamma_{k,\max}]$. In such a setting, one needs to ensure that the barrier-augmented function decrease lemmas that appear in our analyses in Section 3.2 and 3.3 (namely, Lemmas 3.5 and 3.12, respectively) guarantee decreases of the same order in terms of the algorithmic parameters. This can be done, for example, by ensuring that the angle between the resulting direction and $-q_k$ is acute and bounded away from 90° by a threshold that is independent of $k$ and that the norm of the direction and $-q_k$ are proportional uniformly over all $k \in \mathbb{N}$. However, since such extensions would only obfuscate* *our analysis (specifically, Lemma 3.6) without adding significant value to our conclusions, we consider the simpler procedures in Algorithm 1, which has the features that would drive convergence in such algorithm variants as well.*

**Remark 2.2.** *Algorithm 1 is written as a primal interior-point method in the sense that each search direction is computed from an n-by-n "Newton-like" system. One could also consider a primal-dual interior-point method where the sequence of Lagrange multiplier estimates, say $\{(y_k, z_k)\}$ (see (2)), act as independent components of the iterates. To ensure convergence guarantees for a deterministic version of such a method, one can employ our strategies for the parameter sequences as long as safeguards are included to ensure that $y_k$ and $z_k$ remain within appropriately defined neighborhoods of $\mu_k \Psi(x_k - l)^{-1}\mathbb{1}$ and $\mu_k \Psi(u - x_k)^{-1}\mathbb{1}$, respectively, for all $k \in \mathbb{N}$. Similar safeguards have been used in the literature; see, e.g., [27, Section 2.2]. However, convergence guarantees for such a method in the stochastic setting do not follow readily from our analysis for Algorithm 1; hence, overall, we do not consider a primal-dual interior-point variant of Algorithm 1 in this paper.*

# 3 Convergence Analyses

We analyze the behavior of Algorithm 1 under Assumption 2.1 as well as the following assumption.

**Assumption 3.1.** *The iterate sequence $\{x_k\}$ of Algorithm 1 is contained in an open set $\mathcal{X} \subseteq \text{int}(\mathcal{B})$ over which distances of iterate components to finite bounds are bounded in the sense that, for some $\chi \in \mathbb{R}_{>1}$, one has for all $k \in \mathbb{N}$ that $x_{k,i} - l_i \leq \chi$ for all $i \in \mathcal{L}$ and $u_i - x_{k,i} \leq \chi$ for all $i \in \mathcal{U}$.*

The bounds required for Assumption 3.1 to hold are not restrictive for practical purposes. Indeed, while Assumption 3.1 requires that $\chi \in \mathbb{R}_{>0}$ exists, it can be arbitrarily large and *knowledge of it is not required by the algorithm*; see upcoming Lemma 3.1.

## 3.1 Preliminary Results

In this subsection, we provide preliminary results that are required for our analyses of Algorithm 1 for the deterministic and stochastic settings, which are considered separately in the subsequent subsections.

Our first lemma essentially shows that derivatives of the barrier-augmented function are unaffected by scaling of the displacements from finite bounds that appear in the barrier function. For the lemma, we introduce the function $\tilde{\phi} : \text{int}(\mathcal{B}) \times \mathbb{R}_{>0} \to \mathbb{R}$ that one obtains by scaling the barrier terms by $\chi^{-1}$ (see Assumption 3.1), namely,

$$\tilde{\phi}(x, \mu) = f(x) - \mu \sum_{i \in \mathcal{L}} \log\left((x_i - l_i)/\chi\right) - \mu \sum_{i \in \mathcal{U}} \log\left((u_i - x_i)/\chi\right). \tag{7}$$

Important relationships between $\phi$ and $\tilde{\phi}$ and the derivatives of these functions with respect to their first argument are the subject of this first lemma.

**Lemma 3.1.** *For all $(x, \mu) \in \mathcal{X} \times \mathbb{R}_{>0}$, one finds $\tilde{\phi}(x, \mu) = \phi(x, \mu) + \mu M \geq f_{\inf}$, so $\nabla_x \phi(x, \mu) = \nabla_x \tilde{\phi}(x, \mu)$, where $M \in \mathbb{R}_{>0}$ is independent of $x$ and $\mu$. Moreover, for any $(\mu, \bar{\mu}) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ with $\bar{\mu} < \mu$, one has that $\tilde{\phi}(x, \bar{\mu}) < \tilde{\phi}(x, \mu)$ for all $x \in \mathcal{X}$.*

*Proof.* The first desired equation follows from the definitions of $\phi(\cdot, \mu)$ and $\tilde{\phi}(\cdot, \mu)$ in (3) and (7), respectively, and the fact that, for any $\delta \in \mathbb{R}_{>0}$, one finds (since $\chi \in \mathbb{R}_{>1}$) that $-\log(\delta/\chi) = -\log(\delta) + \log(\chi)$. Then, the fact that $\tilde{\phi}(x, \mu) \geq f_{\inf}$ for all $(x, \mu) \in \mathcal{X} \times \mathbb{R}_{>0}$ follows by Assumption 2.1 and the fact that $(x_i - l_i)/\chi \in [0, 1]$ for all $i \in \mathcal{L}$ and $(u_i - x_i)/\chi \in [0, 1]$ for all $i \in \mathcal{U}$. Next, the desired conclusions pertaining to the derivatives of $\phi(\cdot, \mu)$ and $\tilde{\phi}(\cdot, \mu)$ follow from the first conclusion. The final desired conclusion follows from the fact that, for all $x \in \mathcal{X}$, one finds that $(x_i - l_i)/\chi \in [0, 1]$ for all $i \in \mathcal{L}$ and $(u_i - x_i)/\chi \in [0, 1]$ for all $i \in \mathcal{U}$. □

A consequence of Lemma 3.1 is that, for any $k \in \mathbb{N}$ such that the true gradient of the objective $\nabla f(x_k)$ is used in the definition of $q_k$, the search direction computation in Algorithm 1 produces a descent direction for $\tilde{\phi}(\cdot, \mu_k)$ from $x_k$ (recall that $H_k \succ 0$) *even without explicit knowledge of the bound $\chi$ defined in Assumption 3.1.*

We now prove that the scaled barrier-augmented function has a gradient that satisfies a Lipschitz-continuity property over the line segment between any two points in a neighborhood of the type that is defined for the algorithm. The result also shows that the corresponding Lipschitz constant depends on how close the elements of the points are to their corresponding lower and/or upper bounds, which, as shown in our subsequent analysis, is a fact that can be exploited by our algorithm.

**Lemma 3.2.** *For any $(\mu, \theta, \bar{\theta}) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ with $\bar{\theta} \in (0, \theta]$, $(x, \bar{x}) \in \mathcal{N}_{[l,u]}(\theta) \times \mathcal{N}_{[l,u]}(\bar{\theta})$, and $\gamma \in [0, 1]$, one finds that*

$$\|\nabla_x \tilde{\phi}(x + \gamma(\bar{x} - x), \mu) - \nabla_x \tilde{\phi}(x, \mu)\|_2 \leq \gamma \ell_{\nabla f, \mathcal{B}, \mu, x, \bar{x}} \|\bar{x} - x\|_2, \tag{8}$$

*where $\ell_{\nabla f, \mathcal{B}, \mu, x, \bar{x}} := \ell_{\nabla f, \mathcal{B}} + \mu a(x, \bar{x})^{-1} + \mu b(x, \bar{x})^{-1}$ with*

$$a_i(x) := x_i - l_i; \qquad a_i^+(x, \bar{x}) := \min\{x_i - l_i, \bar{x}_i - l_i\}; \qquad a(x, \bar{x}) := \min_{i \in [n]}\{a_i(x) a_i^+(x, \bar{x})\}$$

$$b_i(x) := u_i - x_i; \qquad b_i^+(x, \bar{x}) := \min\{u_i - x_i, u_i - \bar{x}_i\}; \qquad b(x, \bar{x}) := \min_{i \in [n]}\{b_i(x) b_i^+(x, \bar{x})\}.$$

*Moreover, one finds that $\ell_{\nabla f, \mathcal{B}, \mu, x, \bar{x}} \leq \ell_{\nabla f, \mathcal{B}, \mu, \theta, \bar{\theta}} := \ell_{\nabla f, \mathcal{B}} + 2\mu \theta^{-1} \bar{\theta}^{-1} \in \mathbb{R}_{>0}$.*

*Proof.* For arbitrary such $(\mu, \theta, \bar{\theta}, x, \bar{x}, \gamma)$, (4) and Lemma 3.1 imply

$$\begin{aligned}
&\|\nabla_x \tilde{\phi}(x + \gamma(\bar{x} - x), \mu) - \nabla_x \tilde{\phi}(x, \mu)\|_2 \\
&\leq \|\nabla f(x + \gamma(\bar{x} - x)) - \nabla f(x)\|_2 \\
&\quad + \mu \|((\Psi(x - l) + \gamma \Psi(\bar{x} - x))^{-1} - \Psi(x - l)^{-1}) \mathbb{1}\|_2 \\
&\quad + \mu \|((\Psi(u - x) - \gamma \Psi(\bar{x} - x))^{-1} - \Psi(u - x)^{-1}) \mathbb{1}\|_2.
\end{aligned}$$

Considering the latter two terms, for arbitrary $i \in [n]$, one has

$$\tfrac{1}{x_i + \gamma(\bar{x}_i - x_i) - l_i} - \tfrac{1}{x_i - l_i} = \tfrac{\gamma(x_i - \bar{x}_i)}{(x_i + \gamma(\bar{x}_i - x_i) - l_i)(x_i - l_i)} \leq \tfrac{\gamma(x_i - \bar{x}_i)}{a_i(x) a_i^+(x, \bar{x})}$$

and similarly that $(u_i - x_i - \gamma(\bar{x}_i - x_i))^{-1} - (u_i - x_i)^{-1} \leq \gamma(x_i - \bar{x}_i) b_i(x)^{-1} b_i^+(x, \bar{x})^{-1}$. By Assumption 2.1 and these bounds over all $i \in [n]$, the desired conclusion follows. □

**Remark 3.1.** *Lemma 3.2 and all of our subsequent analysis could be based on the simpler, but more conservative $\ell_{\nabla f, \mathcal{B}, \mu, \theta, \bar{\theta}}$ rather than $\ell_{\nabla f, \mathcal{B}, \mu, x, \bar{x}}$ in (8). Similarly, the step-size rule that we present in upcoming Parameter Rule 3.1 could be based on the simpler, but more conservative $2\theta^{-1}\bar{\theta}^{-1}$ rather than $a(x, \bar{x})^{-1} + b(x, \bar{x})^{-1}$. However, these tighter bounds have the effect of allowing larger step sizes, which is beneficial in the numerical experiments presented in Section 5. Hence, we make these choices to have consistency between our analysis and numerical experimentation.*

The following consequence of Lemma 3.2 is central to our analysis.

**Lemma 3.3.** *For any $(\mu, \theta, \bar{\theta}) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ with $\bar{\theta} \in (0, \theta]$ and $(x, \bar{x}) \in \mathcal{N}_{[l,u]}(\theta) \times \mathcal{N}_{[l,u]}(\bar{\theta})$, one finds with $\ell_{\nabla f, \mathcal{B}, \mu, x, \bar{x}} \in \mathbb{R}_{>0}$ is defined in Lemma 3.2 that*

$$\tilde{\phi}(\bar{x}, \mu) \leq \tilde{\phi}(x, \mu) + \nabla_x \tilde{\phi}(x, \mu)^T (\bar{x} - x) + \tfrac{1}{2} \ell_{\nabla f, \mathcal{B}, \mu, x, \bar{x}} \|\bar{x} - x\|_2^2.$$

*Proof.* For arbitrary $(\mu, \theta, \bar{\theta}, x, \bar{x})$ satisfying the conditions of the lemma, it follows with the Fundamental Theorem of Calculus and the Cauchy-Schwarz inequality that

$$\tilde{\phi}(\bar{x}, \mu) - \tilde{\phi}(x, \mu)$$

$$
\begin{aligned}
&= \int_0^1 \tfrac{\partial \tilde{\phi}(x+\gamma(\overline{x}-x),\mu)}{\partial \gamma} \mathrm{d}\gamma = \int_0^1 \nabla_x \tilde{\phi}(x+\gamma(\overline{x}-x),\mu)^T(\overline{x}-x)\mathrm{d}\gamma \\
&= \nabla_x \tilde{\phi}(x,\mu)^T(\overline{x}-x) + \int_0^1 (\nabla_x \tilde{\phi}(x+\gamma(\overline{x}-x),\mu) - \nabla_x \tilde{\phi}(x,\mu))^T(\overline{x}-x)\mathrm{d}\gamma \\
&\leq \nabla_x \tilde{\phi}(x,\mu)^T(\overline{x}-x) + \|\overline{x}-x\|_2 \int_0^1 \|\nabla_x \tilde{\phi}(x+\gamma(\overline{x}-x),\mu) - \nabla_x \tilde{\phi}(x,\mu)\|_2 \mathrm{d}\gamma.
\end{aligned}
$$

Hence, the desired conclusion follows along with Lemma 3.2 and since $\int_0^1 \gamma \mathrm{d}\gamma = \frac{1}{2}$. □

The prior lemma motivates the following parameter rule that we make going forward. We remark at this stage that, for our analysis of the deterministic setting in the next subsection, one can consider $\alpha_{k,\max} \leftarrow \infty$ for all $k \in \mathbb{N}$ so that the step size is always set to be the first term in the minimum in (9). However, for the stochastic setting, our analysis requires a more conservative choice for $\{\alpha_{k,\max}\}$; see Section 3.3. Hence, we introduce $\{\alpha_{k,\max}\}$ at this stage, and carry it through our analysis, to maintain consistency between the deterministic and stochastic settings.

**Parameter Rule 3.1.** *For all $k \in \mathbb{N}$, the algorithm has*

$$\alpha_{k,\max} \geq \alpha_{k,\min}, \quad \textit{where} \quad \alpha_{k,\min} := \tfrac{\lambda_{k,\min}}{\ell_{\nabla f,\mathcal{B}} + 2\mu_k \theta_k^{-2}},$$

*and the algorithm sets*

$$
\begin{aligned}
\alpha_k &\leftarrow \min\left\{\tfrac{\lambda_{k,\min}}{\ell_{\nabla f,\mathcal{B},k}}, \alpha_{k,\max}\right\}, \qquad (9) \\
\textit{where} \quad \alpha_{k,\text{pre}} &\leftarrow \tfrac{\lambda_{k,\min}}{\ell_{\nabla f,\mathcal{B}} + \mu_k a(x_k,x_k)^{-1} + \mu_k b(x_k,x_k)^{-1}}, \\
\bar{\gamma}_k &\leftarrow \max\{\gamma \in (0,\gamma_{k,\max}] : x_k + \gamma \alpha_{k,\text{pre}} d_k \in \mathcal{N}_{[l,u]}(\theta_k)\}, \\
\textit{and} \quad \ell_{\nabla f,\mathcal{B},k} &\leftarrow \ell_{\nabla f,\mathcal{B}} + \mu_k a(x_k, x_k + \bar{\gamma}_k \alpha_{k,\text{pre}} d_k)^{-1} + \mu_k b(x_k, x_k + \bar{\gamma}_k \alpha_{k,\text{pre}} d_k)^{-1}.
\end{aligned}
$$

The following lemma shows that the step-size rule in Parameter Rule 3.1 employs a denominator, namely, $\ell_{\nabla f,\mathcal{B},k}$, that serves as an upper bound for the Lipschitz constant seen in Lemmas 3.2 and 3.3. An implication of this fact is that the inequalities in these lemmas hold with that constant replaced by $\ell_{\nabla f,\mathcal{B},k}$, and another implication, stated in the lemma, is that the step size is contained in a prescribed interval. (The proof reveals that an important property of $\alpha_{k,\text{pre}}$ is that it can be computed prior to $\alpha_k$, yet is ensured to be an upper bound for the value of $\alpha_k$ that will be computed.)

**Lemma 3.4.** *For all $k \in \mathbb{N}$, with $\ell_{\nabla f,\mathcal{B},\mu_k,x_k,x_{k+1}}$ defined as in Lemma 3.2 and $\ell_{\nabla f,\mathcal{B},k}$ defined as in Parameter Rule 3.1, one finds that*

$$\ell_{\nabla f,\mathcal{B},\mu_k,x_k,x_{k+1}} \leq \ell_{\nabla f,\mathcal{B},k} \leq \ell_{\nabla f,\mathcal{B},\mu_k,\theta_{k-1},\theta_k} \leq \ell_{\nabla f,\mathcal{B}} + 2\mu_k \theta_k^{-2}, \qquad (10)$$

*from which it follows that the step size in Parameter Rule 3.1 has $\alpha_k \in [\alpha_{k,\min}, \alpha_{k,\max}]$.*

*Proof.* Consider arbitrary $k \in \mathbb{N}$. To prove the first inequality in (10), it follows from the definitions in Lemma 3.2 that it is sufficient to prove that, for all $i \in [n]$,

$$
\begin{aligned}
a_i^+(x_k, x_{k+1}) &\geq a_i^+(x_k, x_k + \bar{\gamma}_k \alpha_{k,\text{pre}} d_k) \qquad (11\text{a}) \\
\text{and} \quad b_i^+(x_k, x_{k+1}) &\geq b_i^+(x_k, x_k + \bar{\gamma}_k \alpha_{k,\text{pre}} d_k). \qquad (11\text{b})
\end{aligned}
$$

Toward this end, let us first show that $\gamma_k \alpha_k \leq \bar{\gamma}_k \alpha_{k,\text{pre}}$. Denoting (with $\min \emptyset = \infty$)

$$
\begin{aligned}
\delta_k^l(\alpha_k) &:= \min\left\{\tfrac{l_i + \theta_k - x_{k,i}}{\alpha_k d_{k,i}} : d_{k,i} < 0, i \in [n]\right\} \\
\text{and} \quad \delta_k^u(\alpha_k) &:= \min\left\{\tfrac{u_i - \theta_k - x_{k,i}}{\alpha_k d_{k,i}} : d_{k,i} > 0, i \in [n]\right\},
\end{aligned}
$$

the definition of $\gamma_k$ in line 5 of Algorithm 1 yields $\gamma_k = \min\{\gamma_{k,\max}, \delta_k^l(\alpha_k), \delta_k^u(\alpha_k)\}$. Thus, $\gamma_k\alpha_k = \min\{\gamma_{k,\max}\alpha_k, \delta_k^l(1), \delta_k^u(1)\}$. Similarly, by the definition of $\bar{\gamma}_k$ in Parameter Rule 3.1, one finds that $\bar{\gamma}_k\alpha_{k,\text{pre}} = \min\{\gamma_{k,\max}\alpha_{k,\text{pre}}, \delta_k^l(1), \delta_k^u(1)\}$. Hence, since $a(x_k, x_k + \bar{\gamma}_k\alpha_{k,\text{pre}}d_k) \leq a(x_k, x_k)$ and $b(x_k, x_k + \bar{\gamma}_k\alpha_{k,\text{pre}}d_k) \leq b(x_k, x_k)$ imply $\alpha_k \leq \alpha_{k,\text{pre}}$, one finds $\gamma_k\alpha_k \leq \bar{\gamma}_k\alpha_{k,\text{pre}}$, as desired. Now, for $i \in [n]$ with $d_{k,i} < 0$,

$$\begin{aligned} a_i^+(x_k, x_{k+1}) &= x_{k+1,i} - l_i = x_{k,i} + \gamma_k\alpha_k d_{k,i} - l_i \\ &\geq x_{k,i} + \bar{\gamma}_k\alpha_{k,\text{pre}}d_{k,i} - l_i = a_i^+(x_k, x_k + \bar{\gamma}_k\alpha_{k,\text{pre}}d_k), \end{aligned}$$

while for $i \in [n]$ with $d_{k,i} \geq 0$, $a_i^+(x_k, x_{k+1}) = x_{k,i} - l_i = a_i^+(x_k, x_k + \bar{\gamma}_k\alpha_{k,\text{pre}}d_k)$. Therefore, (11a) holds. One finds that (11b) holds with a similar derivation. Consequently, the first desired inequality in (10) holds. The remaining desired inequalities in (10) follow by the definitions in Lemma 3.2, Parameter Rule 3.1, and $\{\theta_k\} \searrow 0$. □

The step-size choice in Parameter Rule 3.1 depends on the Lipschitz constant $\ell_{\nabla f,\mathcal{B}}$ (amongst other quantities prescribed and/or computed in Algorithm 1), which can be any Lipschitz constant for $\nabla f$ over $\mathcal{B}$ (i.e., it does not need to be the minimal such Lipschitz constant). This choice is reasonable for practical purposes since such a value can be computed or estimated in practice. Overall, this choice of step size in Parameter Rule 3.1 can be viewed as a generalization of the $\Theta(1/\ell_{\nabla f,\mathcal{B}})$-type step-size rules common in unconstrained (deterministic and stochastic) optimization.

**Remark 3.2.** *Observe that Lemma 3.2 reveals that convergence guarantees in the context of an interior-point method do not follow readily from the standard arguments for a (stochastic-)gradient-based method for unconstrained optimization. In particular, even though the lemma shows that the gradient of the (scaled) barrier-augmented function is Lipschitz continuous over $\mathcal{N}_{[l,u]}(\theta)$ for any $\theta \in \mathbb{R}_{>0}$, the Lipschitz constant (for a given $\mu \in \mathbb{R}_{>0}$) can diverge as $\theta \searrow 0$. Hence, ensuring convergence requires a careful balance between the barrier-parameter sequence, step-size sequence, and neighborhood-parameter sequence, as revealed in our subsequent analyses.*

## 3.2 Deterministic Setting

We now focus on convergence guarantees that can be shown for Algorithm 1 in the deterministic setting, i.e., when $q_k$ is computed using $\nabla f(x_k)$ for all $k \in \mathbb{N}$. We begin by proving a set of generic results, then conclude with observations about specific choices for the parameter sequences that yield convergence guarantees with respect to stationarity measures.

We first provide a decrease lemma for the shifted barrier-augmented function, which is reminiscent of a standard decrease lemma for a gradient-based method in the context of unconstrained optimization. The particular form of this result is a consequence of the choice of the step size stated in Parameter Rule 3.1.

**Lemma 3.5.** *For all $k \in \mathbb{N}$, one finds that*

$$\tilde{\phi}(x_{k+1}, \mu_{k+1}) - \tilde{\phi}(x_k, \mu_k) \leq -\tfrac{1}{2}\gamma_k\alpha_k\|\nabla_x\tilde{\phi}(x_k, \mu_k)\|_{H_k^{-1}}^2.$$

*Proof.* For all $k \in \mathbb{N}$, Lemmas 3.3 and 3.4, the value of $x_{k+1}$ in line 6 of Algorithm 1, and the conditions on $H_k$ in line 2 of Algorithm 1 imply

$$\begin{aligned} &\tilde{\phi}(x_{k+1}, \mu_k) - \tilde{\phi}(x_k, \mu_k) \\ \leq\ &\nabla_x\tilde{\phi}(x_k, \mu_k)^T(x_{k+1} - x_k) + \tfrac{1}{2}\ell_{\nabla f,\mathcal{B},\mu_k,x_k,x_{k+1}}\|x_{k+1} - x_k\|_2^2 \\ \leq\ &-\nabla_x\tilde{\phi}(x_k, \mu_k)^T(\gamma_k\alpha_k H_k^{-1}\nabla_x\tilde{\phi}(x_k, \mu_k)) + \tfrac{1}{2}\ell_{\nabla f,\mathcal{B},k}\|\gamma_k\alpha_k H_k^{-1}\nabla_x\tilde{\phi}(x_k, \mu_k)\|_2^2 \\ \leq\ &-\gamma_k\alpha_k\|\nabla_x\tilde{\phi}(x_k, \mu_k)\|_{H_k^{-1}}^2 + \tfrac{1}{2}\gamma_k^2\alpha_k^2\lambda_{k,\min}^{-1}\ell_{\nabla f,\mathcal{B},k}\|\nabla_x\tilde{\phi}(x_k, \mu_k)\|_{H_k^{-1}}^2 \\ =\ &-\gamma_k\alpha_k(1 - \tfrac{1}{2}\gamma_k\alpha_k\lambda_{k,\min}^{-1}\ell_{\nabla f,\mathcal{B},k})\|\nabla_x\tilde{\phi}(x_k, \mu_k)\|_{H_k^{-1}}^2. \end{aligned}$$

Now, one finds under Parameter Rule 3.1 that the parameter sequences yield

$$\alpha_k \ell_{\nabla f,\mathcal{B},k} \leq \lambda_{k,\min} \implies \gamma_k \alpha_k \ell_{\nabla f,\mathcal{B},k} \leq \lambda_{k,\min} \iff \tfrac{1}{2} \leq 1 - \tfrac{1}{2}\gamma_k \alpha_k \lambda_{k,\min}^{-1} \ell_{\nabla f,\mathcal{B},k}.$$

Thus, one finds from above that $\tilde{\phi}(x_{k+1},\mu_k) - \tilde{\phi}(x_k,\mu_k) \leq -\frac{1}{2}\gamma_k\alpha_k \|\nabla_x \tilde{\phi}(x_k,\mu_k)\|_{H_k^{-1}}^2$. Combining this inequality with Lemma 3.1, which since $\mu_{k+1} < \mu_k$ shows that $\tilde{\phi}(x_{k+1},\mu_{k+1}) < \tilde{\phi}(x_{k+1},\mu_k)$, one reaches the desired conclusion. □

We now prove a critical lower bound on each element of the sequence $\{\gamma_k\}$.

**Lemma 3.6.** *For all $k \in \mathbb{N}$, define*

$$\gamma_{k,\min} := \min\left\{1, \frac{\lambda_{k,\min}\left(\frac{\frac{1}{2}\mu_k\Delta}{\mu_k + \frac{1}{2}\kappa_{\nabla f,\mathcal{B},\infty}\Delta} - \theta_k\right)}{\alpha_k(\kappa_{\nabla f,\mathcal{B},\infty} + \mu_k\theta_{k-1}^{-1})}\right\} \tag{12}$$

*and suppose $\gamma_{k,\max} \in [\gamma_{k,\min}, 1]$. Then, for all $k \in \mathbb{N}$, $\gamma_k \geq \gamma_{k,\min}$.*

*Proof.* Recall that the algorithm ensures, for all $k \in \mathbb{N}$, that

$$x_k \in \mathcal{N}_{[l,u]}(\theta_{k-1}) \iff l + \theta_{k-1} \leq x_k \leq u - \theta_{k-1}. \tag{13}$$

For all $k \in \mathbb{N}$ and $i \in [n]$, let $\gamma_{k,i} := \max\{\gamma \in (0, \gamma_{k,\max}] : x_{k,i} + \gamma\alpha_k d_{k,i} \in [l_i + \theta_k, u_i - \theta_k]\}$ so that $\gamma_k \leftarrow \min_{i\in[n]} \gamma_{k,i}$. Considering arbitrary $k \in \mathbb{N}$ and $i \in [n]$, let us suppose that $d_{k,i} < 0$ and prove a lower bound on $\gamma_{k,i}$ that is independent of the index $i \in [n]$. One would find—although we omit details for brevity—that the same lower bound for $\gamma_{k,i}$ can be proved when $d_{k,i} > 0$ in a similar manner. All of this, along with the fact that $\gamma_{k,i} = \gamma_{k,\max}$ when $d_{k,i} = 0$, leads to the desired conclusion.

Consider arbitrary $k \in \mathbb{N}$ and $i \in [n]$ and, as previously stated, suppose that $d_{k,i} < 0$. If $i \notin \mathcal{L}$, then it follows that $\gamma_{k,i} = \gamma_{k,\max}$. Hence, we proceed under the assumption that $i \in \mathcal{L}$. If $x_{k,i}+\gamma_{k,\max}\alpha_k d_{k,i} \geq l_i+\theta_k$, then $\gamma_{k,i} = \gamma_{k,\max}$. Otherwise, the algorithm ensures $x_{k,i} + \gamma_{k,i}\alpha_k d_{k,i} = l_i + \theta_k$, and (13) and Assumption 2.1 give

$$\begin{aligned}\gamma_{k,i} &= \tfrac{x_{k,i}-l_i-\theta_k}{\alpha_k [H_k]_{i,i}^{-1}(\nabla_i f(x_k) - \mu_k(x_{k,i}-l_i)^{-1} + \mu_k(u_i - x_{k,i})^{-1})} \\ &\geq \tfrac{x_{k,i}-l_i-\theta_k}{\alpha_k [H_k]_{i,i}^{-1}(\nabla_i f(x_k) + \mu_k(u_i - x_{k,i})^{-1})} \geq \tfrac{\lambda_{k,\min}(x_{k,i}-l_i-\theta_k)}{\alpha_k(|\nabla_i f(x_k)| + \mu_k\theta_{k-1}^{-1})}.\end{aligned} \tag{14}$$

The remainder of our analysis in this case hinges on providing a positive lower bound for $x_{k,i} - l_i$. First, if $i \in \mathcal{L}$ and $i \notin \mathcal{U}$, then one finds that $d_{k,i} < 0$ means

$$\begin{aligned}&[H_k]_{i,i}^{-1}\left(-\nabla_i f(x_k) + \tfrac{\mu_k}{x_{k,i}-l_i}\right) < 0 \\ \iff &\left(-\nabla_i f(x_k) + \tfrac{\mu_k}{x_{k,i}-l_i}\right) < 0 \iff \nabla_i f(x_k) > 0 \ \text{ and } \ x_{k,i} - l_i > \tfrac{\mu_k}{|\nabla_i f(x_k)|}.\end{aligned}$$

Second, if $i \in \mathcal{L} \cup \mathcal{U}$, then one finds with $\Delta_i := u_i - l_i$ that

$$\begin{aligned}& & d_{k,i} = [H_k]_{i,i}^{-1}\left(-\nabla_i f(x_k) + \tfrac{\mu_k}{x_{k,i}-l_i} - \tfrac{\mu_k}{u_i - x_{k,i}}\right) &< 0 \\ &\iff & -\nabla_i f(x_k) + \tfrac{\mu_k}{x_{k,i}-l_i} - \tfrac{\mu_k}{\Delta_i - (x_{k,i}-l_i)} &< 0 \\ &\iff & \mu_k\left(\tfrac{\Delta_i - 2(x_{k,i}-l_i)}{(x_{k,i}-l_i)(\Delta_i-(x_{k,i}-l_i))}\right) &< \nabla_i f(x_k) \\ &\iff & \nabla_i f(x_k)(x_{k,i}-l_i)^2 - (2\mu_k + \nabla_i f(x_k)\Delta_i)(x_{k,i}-l_i) + \mu_k\Delta_i &< 0.\end{aligned}$$

Given this inequality, there are three subcases to consider depending on $\nabla_i f(x_k)$.

(i) Suppose $\nabla_i f(x_k) = 0$. Then, $x_{k,i} - l_i > \frac{1}{2}\Delta_i$.

(ii) Suppose $\nabla_i f(x_k) < 0$. Then, by the quadratic formula, it follows that

$$x_{k,i} - l_i < \frac{\mu_k}{\nabla_i f(x_k)} + \frac{\Delta_i}{2} - \sqrt{\frac{\mu_k^2}{(\nabla_i f(x_k))^2} + \frac{\Delta_i^2}{4}} \tag{15a}$$

$$\text{or}\ \ x_{k,i} - l_i > \frac{\mu_k}{\nabla_i f(x_k)} + \frac{\Delta_i}{2} + \sqrt{\frac{\mu_k^2}{(\nabla_i f(x_k))^2} + \frac{\Delta_i^2}{4}}. \tag{15b}$$

In fact, the upper bound on $x_{k,i} - l_i$ stated in (15a) is not possible since

$$\begin{aligned} &\frac{\mu_k}{\nabla_i f(x_k)} + \frac{\Delta_i}{2} - \sqrt{\frac{\mu_k^2}{(\nabla_i f(x_k))^2} + \frac{\Delta_i^2}{4}} \\ &\leq \frac{\mu_k}{\nabla_i f(x_k)} + \frac{\Delta_i}{2} - \sqrt{\frac{\mu_k^2}{(\nabla_i f(x_k))^2} + 2\frac{\mu_k}{\nabla_i f(x_k)}\frac{\Delta_i}{2} + \frac{\Delta_i^2}{4}} \\ &\leq \frac{\mu_k}{\nabla_i f(x_k)} + \frac{\Delta_i}{2} - \left|\frac{\mu_k}{\nabla_i f(x_k)} + \frac{\Delta_i}{2}\right| \leq 0 \end{aligned}$$

while the algorithm ensures $x_{k,i} - l_i > 0$. Hence, (15b) must hold in this case, from which it follows (by dropping the $\frac{1}{4}\Delta_i^2$ term) that $x_{k,i} - l_i > \frac{1}{2}\Delta_i$.

(iii) Suppose $\nabla_i f(x_k) > 0$. Then, by the quadratic formula, it follows that

$$x_{k,i} - l_i > \frac{\mu_k}{\nabla_i f(x_k)} + \frac{\Delta_i}{2} - \sqrt{\frac{\mu_k^2}{(\nabla_i f(x_k))^2} + \frac{\Delta_i^2}{4}}. \tag{16}$$

Define $a_{k,i} := \frac{\mu_k^2}{(\nabla_i f(x_k))^2} + 2\frac{\mu_k}{\nabla_i f(x_k)}\frac{\Delta_i}{2} + \frac{\Delta_i^2}{4}$ and $b_{k,i} := \frac{\mu_k^2}{(\nabla_i f(x_k))^2} + \frac{\Delta_i^2}{4}$, and observe that $a_{k,i} > b_{k,i} > 0$ while the right-hand side of (16) is equal to $s(a_{k,i}) - s(b_{k,i})$, where $s(\cdot) := \sqrt{\cdot}$ is the square root function. By the mean value theorem, there exists a real number $c \in [b_{k,i}, a_{k,i}]$ such that one finds

$$s(a_{k,i}) - s(b_{k,i}) = s'(c)(a_{k,i} - b_{k,i}), \quad \text{where} \quad s'(c) = \frac{1}{2\sqrt{c}} \geq \frac{1}{2\sqrt{a_{k,i}}}.$$

Hence, one has from (16) that

$$x_{k,i} - l_i > s(a_{k,i}) - s(b_{k,i}) \geq \frac{a_{k,i} - b_{k,i}}{2\sqrt{a_{k,i}}} = \frac{\frac{\mu_k}{\nabla_i f(x_k)}\frac{\Delta_i}{2}}{\frac{\mu_k}{\nabla_i f(x_k)} + \frac{\Delta_i}{2}} = \frac{\frac{1}{2}\mu_k\Delta_i}{\mu_k + \frac{1}{2}\nabla_i f(x_k)\Delta_i}.$$

Combining the results above when $i \in \mathcal{L}$, one finds that $d_{k,i} < 0$ implies

$$x_{k,i} - l_i \geq \min\left\{\frac{\mu_k}{|\nabla_i f(x_k)|}, \frac{\Delta_i}{2}, \frac{\frac{1}{2}\mu_k\Delta_i}{\mu_k + \frac{1}{2}|\nabla_i f(x_k)|\Delta_i}\right\} = \frac{\frac{1}{2}\mu_k\Delta_i}{\mu_k + \frac{1}{2}|\nabla_i f(x_k)|\Delta_i}.$$

Combining this inequality, (14), the facts that $\max_{i\in[n]} |\nabla_i f(x_k)| \leq \kappa_{\nabla f,\mathcal{B},\infty}$ and $\Delta \leq \min_{i\in[n]} \Delta_i$, and the monotonicity of $\frac{\rho z}{\tau + \omega z}$ with respect to $z$ when $\rho$, $\tau$, and $\omega$ are positive, one reaches the desired conclusion. □

The prior lemma motivates the following rule that we make going forward. Similarly as for the choice of the step size (recall Parameter Rule 3.1), the remainder of our analysis for the deterministic setting can use $\gamma_{k,\max} \leftarrow 1$ for all $k \in \mathbb{N}$, but for the stochastic setting our analysis requires a more conservative choice for $\{\gamma_{k,\max}\}$.

**Parameter Rule 3.2.** *For all $k \in \mathbb{N}$, with $\gamma_{k,\min}$ from* (12)*, $\gamma_{k,\max} \in [\gamma_{k,\min}, 1]$.*

We now prove a generic convergence theorem for Algorithm 1 in a deterministic setting. We follow this theorem with a corollary that provides specific choices of the parameter sequences that ensure that the conditions of the theorem hold.

**Theorem 3.1.** *Suppose that Assumptions 2.1 and 3.1 and Parameter Rules 3.1 and 3.2 hold. If, further, the parameter sequences of Algorithm 1 yield*

$$\sum_{k=1}^{\infty} \gamma_k \alpha_k = \infty, \tag{17}$$

*then*

$$\liminf_{k\to\infty} \|\nabla_x \tilde{\phi}(x_k, \mu_k)\|^2_{H_k^{-1}} = 0, \tag{18}$$

*meaning that if there exists $\overline{r} \in \mathbb{R}_{>0}$ such that $\lambda_{k,\max} \leq \overline{r}$ for all $k \in \mathbb{N}$, then*

$$\liminf_{k\to\infty} \|\nabla_x \tilde{\phi}(x_k, \mu_k)\|^2_2 = 0. \tag{19}$$

*Additionally, if such $\overline{r}$ exists, the sequence $\{\mu_k \theta_{k-1}^{-1}\}$ is bounded, and there exists a set $\mathcal{K} \subseteq \mathbb{N}$ of infinite cardinality such that $\{\nabla\tilde{\phi}(x_k, \mu_k)\}_{k\in\mathcal{K}} \to 0$ and $\{x_k\}_{k\in\mathcal{K}} \to \overline{x}$ for some $\overline{x} \in \mathcal{B}$, then the limit point $\overline{x}$ is a KKT point for* (1) *in the sense that there exists $(\overline{y}, \overline{z}) \in \mathbb{R}^n \times \mathbb{R}^n$ such that $(\overline{x}, \overline{y}, \overline{z})$ satisfies* (2)*.*

*Proof.* It follows from Lemma 3.1 that $\tilde{\phi}$ is bounded below by $f_{\inf}$ over $\mathcal{X} \times \mathbb{R}_{>0}$. Then, one finds by summing the expression in Lemma 3.5 over $k \in \mathbb{N}$ that

$$\infty > \tilde{\phi}(x_1, \mu_1) - f_{\inf} \geq \sum_{k=1}^{\infty} (\tilde{\phi}(x_k, \mu_k) - \tilde{\phi}(x_{k+1}, \mu_{k+1})) \geq \sum_{k=1}^{\infty} \tfrac{\gamma_k \alpha_k}{2} \|\nabla_x \tilde{\phi}(x_k, \mu_k)\|^2_{H_k^{-1}}.$$

If there exists $\epsilon \in \mathbb{R}_{>0}$ and $k_\epsilon \in \mathbb{N}$ such that $\|\nabla_x \tilde{\phi}(x_k, \mu_k)\|^2_{H_k^{-1}} \geq \epsilon$ for all $k \in \mathbb{N}$ with $k \geq k_\epsilon$, then the conclusion above contradicts (17). Hence, it follows that such $\epsilon$ and $k_\epsilon$ do not exist, meaning that (18) holds, as desired. Now, if there exists $\overline{r} \in \mathbb{R}_{>0}$ such that $\lambda_{k,\max} \leq \overline{r}$ for all $k \in \mathbb{N}$, then $0 = \liminf_{k\to\infty} \|\nabla_x \tilde{\phi}(x_k, \mu_k)\|^2_{H_k^{-1}} \geq \liminf_{k\to\infty} \overline{r}^{-1} \|\nabla_x \tilde{\phi}(x_k, \mu_k)\|^2_2$, from which (19) holds, as desired. Now suppose that such $\overline{r}$ exists, $\{\mu_k \theta_{k-1}^{-1}\}$ is bounded, and there exists an infinite-cardinality set $\mathcal{K} \subseteq \mathbb{N}$ as described in the theorem. By Lemma 3.1, it follows that $\{\nabla\phi(x_k, \mu_k)\}_{k\in\mathcal{K}} \to 0$ as well. Using this limit and, for all $k \in \mathcal{K}$, defining the auxiliary sequences

$$y_k := \mu_k \Psi(x_k - l)^{-1} \mathbb{1} \ \text{ and } \ z_k := \mu_k \Psi(u - x_k)^{-1} \mathbb{1}, \tag{20}$$

it follows that $\{(x_k, y_k, z_k)\}_{k\in\mathcal{K}} \subset \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$ satisfies

$$\{x_k\}_{k\in\mathcal{K}} \to \overline{x} \ \text{ and } \ \{\|\nabla f(x_k) - y_k + z_k\|_2\}_{k\in\mathcal{K}} \to 0. \tag{21}$$

Next, for all $k \in \mathbb{N}$, it follows from $x_k \in \mathcal{N}_{[l,u]}(\theta_{k-1})$ (see Algorithm 1) that one has $0 \leq y_{k,i} = \frac{\mu_k}{x_{k,i} - l_i} \leq \frac{\mu_k}{\theta_{k-1}}$ and $0 \leq z_{k,i} = \frac{\mu_k}{u_i - x_{k,i}} \leq \frac{\mu_k}{\theta_{k-1}}$. Since $\{\mu_k \theta_{k-1}^{-1}\}$ is bounded by assumption, it follows that $\{y_k\}_{k\in\mathcal{K}}$ and $\{z_k\}_{k\in\mathcal{K}}$ are bounded. Then, the Bolzano-Weierstrass Theorem gives the existence of an infinite subsequence of indices $K_{y,z} \subseteq K$ and vectors $\overline{y} \in \mathbb{R}^n$ and $\overline{z} \in \mathbb{R}^n$ such that

$$\{y_k\}_{k\in K_{y,z}} \to \overline{y} \ \text{ and } \ \{z_k\}_{k\in K_{y,z}} \to \overline{z}. \tag{22}$$

Using these limits, (20), and $\{\mu_k\} \searrow 0$, it follows that

$$\begin{aligned} \overline{y}_i &= 0 \ \text{ for all } i \in [n] \text{ with } \overline{x}_i \neq l_i \\ \text{and } \ \overline{z}_i &= 0 \ \text{ for all } i \in [n] \text{ with } \overline{x}_i \neq u_i. \end{aligned} \tag{23}$$

Combining $\overline{x} \in \mathcal{B}$, $K_{y,z} \subseteq K$, and (20)–(23), it follows that $\overline{x}$ is a KKT point for (1) since the tuple $(\overline{x}, \overline{y}, \overline{z})$ satisfies (2), thus completing the proof. □

The following corollary shows that there exist choices of the parameter sequences such that the conditions of Theorem 3.1 hold.

**Corollary 3.1.** *Suppose that Assumptions 2.1 and 3.1 and Parameter Rules 3.1 and 3.2 hold. Then, there exist parameter choices for Algorithm 1 such that the infinite series in* (17) *is unbounded and* $\{\mu_k\theta_{k-1}^{-1}\}$ *is bounded; e.g., these consequences follow if for some* $\underline{r} \in \mathbb{R}_{>0}$*,* $t \in [-1, 0)$*, and* $\mu_1 \in \mathbb{R}_{>0}$ *with* $\mu_1 > \frac{\frac{1}{2}\theta_0\kappa_{\nabla f,\mathcal{B},\infty}\Delta}{\frac{1}{2}\Delta - \theta_0}$ *the algorithm has* $\mu_k = \mu_1 k^t$*,* $\theta_{k-1} = \theta_0 k^t$*,* $\alpha_{k,\max} \leftarrow \infty$*,* $\gamma_{k,\max} \leftarrow 1$*, and* $\underline{r} \le \lambda_{k,\min} \le \lambda_{k,\max}$ *for all* $k \in \mathbb{N}$*. Thus, with these choices, the lower limit in* (18) *holds, and if there exists* $\overline{r} \in \mathbb{R}_{\ge \underline{r}}$ *such that* $\lambda_{k,\max} \le \overline{r}$ *for all* $k \in \mathbb{N}$*, then the lower limit in* (19) *holds. Finally, if all of the aforementioned choices of the parameter sequences are made and there exists an infinite-cardinality set* $\mathcal{K} \subseteq \mathbb{N}$ *such that* $\{\nabla\tilde{\phi}(x_k, \mu_k)\}_{k\in\mathcal{K}} \to 0$ *and* $\{x_k\}_{k\in\mathcal{K}} \to \overline{x}$ *for some* $\overline{x} \in \mathcal{B}$*, then the limit point* $\overline{x}$ *is a KKT point for* (1).

*Proof.* Under Parameter Rules 3.1 and 3.2, Lemmas 3.4 and 3.6 imply that with the parameter choices given in the corollary, one finds that

$$\begin{aligned}\gamma_k\alpha_k &\ge \underline{r}\min\left\{\frac{1}{\ell_{\nabla f,\mathcal{B}} + 2\mu_1\theta_0^{-2}k^t(k+1)^{-2t}}, \frac{\frac{\frac{1}{2}\mu_1\Delta k^t}{\mu_1 k^t + \frac{1}{2}\kappa_{\nabla f,\mathcal{B},\infty}\Delta} - \theta_0(k+1)^t}{\kappa_{\nabla f,\mathcal{B},\infty} + \mu_1\theta_0^{-1}}\right\} \\ &=: \underline{r}\min\{\beta_k, \eta_k\}. \end{aligned} \tag{24}$$

With respect to the sequence $\{\beta_k\}$, one finds for all $k \in \mathbb{N}$ that

$$k^t(k+1)^{-2t} \le k^t(2k)^{-2t} = 2^{-2t}k^{-t}. \tag{25}$$

For the sequence $\{\eta_k\}$, one finds with (6) and since $t < 0$ that, for all $k \in \mathbb{N}$,

$$\begin{aligned}\frac{\frac{1}{2}\mu_1\Delta k^t}{\mu_1 k^t + \frac{1}{2}\kappa_{\nabla f,\mathcal{B},\infty}\Delta} - \theta_0(k+1)^t &\ge \frac{\frac{1}{2}\mu_1\Delta k^t}{\mu_1 k^t + \frac{1}{2}\kappa_{\nabla f,\mathcal{B},\infty}\Delta} - \theta_0 k^t \\ &= \left(\frac{\frac{1}{2}\mu_1\Delta}{\mu_1 k^t + \frac{1}{2}\kappa_{\nabla f,\mathcal{B},\infty}\Delta} - \theta_0\right)k^t \\ &\ge \left(\frac{\frac{1}{2}\mu_1\Delta}{\mu_1 + \frac{1}{2}\kappa_{\nabla f,\mathcal{B},\infty}\Delta} - \theta_0\right)k^t, \end{aligned} \tag{26}$$

where one finds that (6) (i.e., $\theta_0 < \frac{\Delta}{2}$) and

$$\mu_1 > \frac{\frac{1}{2}\theta_0\kappa_{\nabla f,\mathcal{B},\infty}\Delta}{\frac{1}{2}\Delta - \theta_0} \quad \text{imply} \quad \frac{\frac{1}{2}\mu_1\Delta}{\mu_1 + \frac{1}{2}\kappa_{\nabla f,\mathcal{B},\infty}\Delta} > \theta_0.$$

Hence, combining (24), (25), and (26), there exists $c \in \mathbb{R}_{>0}$ such that

$$\gamma_k\alpha_k \ge \underline{r}\min\{\beta_k, \eta_k\} \ge \underline{r}ck^t, \tag{27}$$

and since $t \in [-1, 0)$, one concludes that the infinite series in (17) is unbounded.

Finally, for all $k \in \mathbb{N}$, it follows from the given parameter choices that one finds $\mu_k\theta_{k-1}^{-1} = \mu_1\theta_0^{-1}$ for all $k \in \mathbb{N}$, so $\{\mu_k\theta_{k-1}^{-1}\} \equiv \{\mu_1\theta_0^{-1}\}$ is bounded, as claimed. □

**Remark 3.3.** *Related to Remark 3.2, we observe that the claimed asymptotic convergence guarantee in [12] overlooks a critical issue in terms of the step sizes. The method in [12] employs step sizes that, amongst other considerations, ensure that each iterate remains feasible. The claimed convergence guarantee then assumes that the step sizes are unsummable. However, when the step sizes need to be reduced to maintain feasibility, one cannot presume that the resulting step sizes are unsummable. This issue is overlooked in [12], but—through a careful balance between the barrier-parameter, step-size, and neighborhood parameter sequences in our algorithm—we find in Corollary 3.1 that there exist parameter choices such that* (17) *holds. We also carry this property forward for the stochastic setting, as seen in Section 3.3.*

We conclude this subsection by observing that if $\mathcal{B}$ is bounded, then $\{x_k\}$ has a convergent subsequence and, under the stated conditions in Corollary 3.1, an infinite-cardinality set $\mathcal{K}$ of the type described in the corollary is guaranteed to exist.

## 3.3 Stochastic Setting

We now provide a convergence guarantee for Algorithm 1 in a stochastic setting when, in every run for all $k \in \mathbb{N}$, $q_k$ is computed using an unbiased stochastic gradient estimate $g_k$ with bounded error; see upcoming Assumption 3.2. Formally, we consider the stochastic process defined by the algorithm, namely, $\{(X_k, G_k, Q_k, H_k, D_k, \overline{\Gamma}_k, A_k, \Gamma_k)\}$, where, for all $k \in \mathbb{N}$, the random variables correspond to the iterate $X_k$, stochastic gradient estimator $G_k$, stochastic barrier-augmented function gradient $Q_k$, scaling matrix $H_k$, direction $D_k$, neighborhood enforcement parameter $\overline{\Gamma}_k$, step size $A_k$, and neighborhood enforcement parameter $\Gamma_k$. A realization of this process is $\{(x_k, g_k, q_k, H_k, d_k, \bar{\gamma}_k, \alpha_k, \gamma_k)\}$, as in Algorithm 1. (Here, we have introduced a slight abuse of notation in terms of $H_k$, which acts as both a random variable and its realization. We prefer this slightly abused notation rather than introduce additional notation; it should not lead to confusion since, for our analysis in this subsection—which considers $H_k$ as a random variable for all $k \in \mathbb{N}$—ultimately relies on the fact that the eigenvalues of the elements of $\{H_k\}$ can be bounded by the prescribed bound sequences $\{\lambda_{k,\min}\}$ and $\{\lambda_{k,\max}\}$.) The behavior of any run of the algorithm is determined by the initial conditions (including that $X_1 = x_1$) and the sequence of stochastic gradient estimators $\{G_k\}$. Let $\mathcal{F}_1$ denote the $\sigma$-algebra defined by the initial conditions and, for all $k \in \mathbb{N}$ with $k \geq 2$, let $\mathcal{F}_k$ denote the $\sigma$-algebra defined by the initial conditions and the random variables $\{G_1, \dots, G_{k-1}\}$, a realization of which determines the realizations of $\{X_j\}_{j=1}^{k}$ and $\{(G_j, Q_j, D_j, \overline{\Gamma}_j, A_j, \Gamma_j)\}_{j=1}^{k-1}$. In this manner, the sequence $\{\mathcal{F}_k\}$ is a filtration.

For our analysis in this subsection, we continue to make Assumptions 2.1 and 3.1, where for Assumption 3.1 we assume that the set $\mathcal{X}$ and real number $\chi$ are uniform over all possible realizations of the stochastic process. In terms of the stochastic gradient estimators and scaling matrices, we make the following assumption.

**Assumption 3.2.** *For all $k \in \mathbb{N}$, one has $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$. In addition, there exists $(\sigma_2, \sigma_\infty) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ such that, for all $k \in \mathbb{N}$, one has*

$$\|G_k - \nabla f(X_k)\|_2 \leq \sigma_2 \ \ \text{and} \ \ \|G_k - \nabla f(X_k)\|_\infty \leq \sigma_\infty.$$

*Finally, for all $k \in \mathbb{N}$, the matrix $H_k \in \mathbb{S}^n$ is $\mathcal{F}_k$-measureable.*

In Assumption 3.2, the existence of $\sigma_\infty$ follows from that of $\sigma_2$, and vice versa, but we introduce both of these values for the sake of notational convenience. It follows under Assumption 3.2 that, for all $k \in \mathbb{N}$, one has $\|G_k\|_\infty \leq \kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty$.

In terms of the algorithmic choices that we consider in our present analysis, we state the following rule that can be seen as a slightly modified combination of Parameter Rules 3.1 and 3.2. Overall, unlike in the deterministic setting where one can prove convergence guarantees (see Theorem 3.1 and Corollary 3.1) using $\alpha_{k,\max} \leftarrow \infty$ and $\gamma_{k,\max} \leftarrow 1$ for all $k \in \mathbb{N}$ and without knowledge of $\kappa_{\nabla f,\mathcal{B},\infty}$ appearing in the definition of $\gamma_{k,\min}$ in Lemma 3.6, for the stochastic setting our analysis requires more conservative choices and information. Specifically, we employ the following rule.

**Parameter Rule 3.3.** *With prescribed $\{\alpha_{k,\mathrm{buff}}\} \subset \mathbb{R}_{>0}$ and $\{\gamma_{k,\mathrm{buff}}\} \subset \mathbb{R}_{>0}$ such that $\{\alpha_{k,\mathrm{buff}}\} = \mathcal{O}(k^{2t})$ and $\{\gamma_{k,\mathrm{buff}}\} = \mathcal{O}(k^t)$ for some $t \in [-1, -\frac{1}{2})$, the algorithm employs for all $k \in \mathbb{N}$ the prescribed (i.e., not random) values*

$$\alpha_{k,\min} := \tfrac{\lambda_{k,\min}}{\ell_{\nabla f,\mathcal{B}} + 2\mu_k\theta_k^{-2}}, \qquad \gamma_{k,\min} := \min\left\{1, \frac{\lambda_{k,\min}\left(\frac{\frac{1}{2}\mu_k\Delta}{\mu_k + \frac{1}{2}(\kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty)\Delta} - \theta_k\right)}{\alpha_{k,\max}(\kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty + \mu_k\theta_{k-1}^{-1})}\right\},$$

$$\alpha_{k,\max} := \alpha_{k,\min} + \alpha_{k,\text{buff}}, \qquad \textit{and} \;\; \gamma_{k,\max} := \min\{1, \gamma_{k,\min} + \gamma_{k,\text{buff}}\},$$

*and makes the (run-and-iterate-dependent) choice* $\alpha_k \leftarrow \min\left\{\frac{\lambda_{k,\min}}{\ell_{\nabla f,\mathcal{B},k}}, \alpha_{k,\max}\right\}$.

Since the barrier parameter sequence $\{\mu_k\}$ and neighborhood parameter sequence $\{\theta_k\}$ are prescribed and Lemmas 3.1, 3.2, and 3.3 hold independently of any algorithm, it follows that the result of Lemma 3.4 holds in the present setting. We formalize this fact in the following lemma, the proof of which is omitted since it would follow the same line of argument as the proof of Lemma 3.4 stated previously.

**Lemma 3.7.** *For all* $k \in \mathbb{N}$*, with* $\ell_{\nabla f,\mathcal{B},\mu_k,X_k,X_{k+1}}$ *defined as in Lemma 3.2 and* $\ell_{\nabla f,\mathcal{B},k}$ *defined as in Parameter Rule 3.1, one finds that*

$$\ell_{\nabla f,\mathcal{B},\mu_k,X_k,X_{k+1}} \leq \ell_{\nabla f,\mathcal{B},k} \leq \ell_{\nabla f,\mathcal{B},\mu_k,\theta_{k-1},\theta_k} \leq \ell_{\nabla f,\mathcal{B}} + 2\mu_k\theta_k^{-2},$$

*from which it follows that Parameter Rule 3.3 guarantees* $A_k \in [\alpha_{k,\min}, \alpha_{k,\max}]$.

We also have that the following result, similar to Lemma 3.6, holds in the present setting. The proof is omitted since it would follow the same line of argument as the proof of Lemma 3.6, the primary differences being that, for all $k \in \mathbb{N}$, one has $A_k \leq \alpha_{k,\max}$ and in place of $|\nabla_i f(x_k)| \leq \kappa_{\nabla f,\mathcal{B},\infty}$ one can employ $|G_{k,i}| \leq \kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty$.

**Lemma 3.8.** *For all* $k \in \mathbb{N}$*, Parameter Rule 3.3 guarantees* $\Gamma_k \in [\gamma_{k,\min}, \gamma_{k,\max}]$.

Our next lemma provides a preliminary upper bound on the expected per-iteration change in the shifted barrier-augmented function.

**Lemma 3.9.** *For all* $k \in \mathbb{N}$*, one finds that*

$$\begin{aligned} &\tilde{\phi}(X_{k+1}, \mu_{k+1}) - \tilde{\phi}(X_k, \mu_k) \\ \leq\; & -\Gamma_k A_k \|\nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 + \Gamma_k A_k \nabla_x \tilde{\phi}(X_k, \mu_k)^T H_k^{-1}(\nabla_x \tilde{\phi}(X_k, \mu_k) - Q_k) \\ & + \tfrac{1}{2}\Gamma_k^2 A_k^2 \lambda_{k,\min}^{-1} \ell_{\nabla f,\mathcal{B},k} \|Q_k\|_{H_k^{-1}}^2. \end{aligned}$$

*Proof.* Similarly as in the proof of Lemma 3.5, for all $k \in \mathbb{N}$, one finds from Lemmas 3.3 and 3.7, line 6 of Algorithm 1, and line 2 of Algorithm 1 that

$$\begin{aligned} &\tilde{\phi}(X_{k+1}, \mu_k) - \tilde{\phi}(X_k, \mu_k) \\ &\leq \nabla_x \tilde{\phi}(X_k, \mu_k)^T (X_{k+1} - X_k) + \tfrac{1}{2}\ell_{\nabla f,\mathcal{B},\mu_k,X_k,X_{k+1}} \|X_{k+1} - X_k\|_2^2 \\ &\leq -\nabla_x \tilde{\phi}(X_k, \mu_k)^T (\Gamma_k A_k H_k^{-1} Q_k) + \tfrac{1}{2}\ell_{\nabla f,\mathcal{B},k} \|\Gamma_k A_k H_k^{-1} Q_k\|_2^2 \\ &\leq -\Gamma_k A_k \nabla_x \tilde{\phi}(X_k, \mu_k)^T H_k^{-1} Q_k + \tfrac{1}{2}\Gamma_k^2 A_k^2 \lambda_{k,\min}^{-1} \ell_{\nabla f,\mathcal{B},k} \|Q_k\|_{H_k^{-1}}^2. \end{aligned}$$

Adding and subtracting $-\Gamma_k A_k \|\nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2$ on the right-hand side and using the fact that Lemma 3.1 and $\mu_{k+1} < \mu_k$ imply that $\tilde{\phi}(X_{k+1}, \mu_{k+1}) < \tilde{\phi}(X_{k+1}, \mu_k)$, one reaches the desired conclusion. □

Our next lemma provides an upper bound on the conditional expectation of the middle term on the right-hand side of the inequality in Lemma 3.9.

**Lemma 3.10.** *For all* $k \in \mathbb{N}$*, one finds that*

$$\begin{aligned} &\mathbb{E}[\Gamma_k A_k \nabla_x \tilde{\phi}(X_k, \mu_k)^T H_k^{-1}(\nabla_x \tilde{\phi}(X_k, \mu_k) - Q_k) | \mathcal{F}_k] \\ \leq\; & (\gamma_{k,\min}\alpha_{k,\text{buff}} + \gamma_{k,\text{buff}}\alpha_{k,\min} + \gamma_{k,\text{buff}}\alpha_{k,\text{buff}}) \lambda_{k,\min}^{-1} (\kappa_{\nabla f,\mathcal{B},2} + 2\sqrt{n}\mu_k\theta_{k-1}^{-1})\sigma_2. \end{aligned}$$

*Proof.* Let $\mathcal{I}_k$ be the event that $P_k := \nabla_x \tilde{\phi}(X_k, \mu_k)^T H_k^{-1}(\nabla_x \tilde{\phi}(X_k, \mu_k) - Q_k) \geq 0$ and let $\mathcal{I}_k^c$ be the complementary event that $P_k < 0$. By Assumption 3.2, Parameter Rule 3.3, the Law of Total Expectation, $\mathbb{E}[P_k|\mathcal{F}_k] = 0$, and Lemmas 3.7 and 3.8,

$$\begin{aligned}
&\mathbb{E}[\Gamma_k A_k P_k|\mathcal{F}_k] \\
&= \mathbb{E}[\Gamma_k A_k P_k|\mathcal{F}_k \wedge \mathcal{I}_k]\mathbb{P}[\mathcal{I}_k|\mathcal{F}_k] + \mathbb{E}[\Gamma_k A_k P_k|\mathcal{F}_k \wedge \mathcal{I}_k^c]\mathbb{P}[\mathcal{I}_k^c|\mathcal{F}_k] \\
&\leq \gamma_{k,\max}\alpha_{k,\max}\mathbb{E}[P_k|\mathcal{F}_k \wedge \mathcal{I}_k]\mathbb{P}[\mathcal{I}_k|\mathcal{F}_k] + \gamma_{k,\min}\alpha_{k,\min}\mathbb{E}[P_k|\mathcal{F}_k \wedge \mathcal{I}_k^c]\mathbb{P}[\mathcal{I}_k^c|\mathcal{F}_k] \\
&\leq \gamma_{k,\min}\alpha_{k,\min}(\mathbb{E}[P_k|\mathcal{F}_k \wedge \mathcal{I}_k]\mathbb{P}[\mathcal{I}_k|\mathcal{F}_k] + \mathbb{E}[P_k|\mathcal{F}_k \wedge \mathcal{I}_k^c]\mathbb{P}[\mathcal{I}_k^c|\mathcal{F}_k]) \\
&\quad + (\gamma_{k,\min}\alpha_{k,\text{buff}} + \gamma_{k,\text{buff}}\alpha_{k,\min} + \gamma_{k,\text{buff}}\alpha_{k,\text{buff}})\mathbb{E}[P_k|\mathcal{F}_k \wedge \mathcal{I}_k]\mathbb{P}[\mathcal{I}_k|\mathcal{F}_k] \\
&= (\gamma_{k,\min}\alpha_{k,\text{buff}} + \gamma_{k,\text{buff}}\alpha_{k,\min} + \gamma_{k,\text{buff}}\alpha_{k,\text{buff}})\mathbb{E}[P_k|\mathcal{F}_k \wedge \mathcal{I}_k]\mathbb{P}[\mathcal{I}_k|\mathcal{F}_k].
\end{aligned}$$

By the Cauchy-Schwarz inequality and Assumptions 2.1 and 3.2, one has

$$\begin{aligned}
&\mathbb{E}[P_k|\mathcal{F}_k \wedge \mathcal{I}_k]\mathbb{P}[\mathcal{I}_k|\mathcal{F}_k] \\
&\leq \mathbb{E}[\|H_k^{-1}\nabla_x \tilde{\phi}(X_k, \mu_k)\|_2\|\nabla_x \tilde{\phi}(X_k, \mu_k) - Q_k\|_2|\mathcal{F}_k \wedge \mathcal{I}_k]\mathbb{P}[\mathcal{I}_k|\mathcal{F}_k] \\
&= \mathbb{E}[\|H_k^{-1}\nabla_x \tilde{\phi}(X_k, \mu_k)\|_2\|G_k - \nabla f(X_k)\|_2|\mathcal{F}_k \wedge \mathcal{I}_k]\mathbb{P}[\mathcal{I}_k|\mathcal{F}_k] \\
&\leq \lambda_{k,\min}^{-1}(\kappa_{\nabla f,\mathcal{B},2} + 2\sqrt{n}\mu_k\theta_{k-1}^{-1})\sigma_2,
\end{aligned}$$

which combined with the result above yields the desired conclusion. □

Our next lemma provides an upper bound on the last term on the right-hand side of the inequality in Lemma 3.9.

**Lemma 3.11.** *For all $k \in \mathbb{N}$, one finds that*

$$\begin{aligned}
&\tfrac{1}{2}\Gamma_k^2 A_k^2 \lambda_{k,\min}^{-1}\ell_{\nabla f,\mathcal{B},k}\|Q_k\|_{H_k^{-1}}^2 \\
&\leq \tfrac{3}{4}\Gamma_k^2 A_k^2 \lambda_{k,\min}^{-1}\ell_{\nabla f,\mathcal{B},k}\|\nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 + \tfrac{3}{2}\gamma_{k,\max}^2\alpha_{k,\max}^2\lambda_{k,\min}^{-2}\sigma_2^2.
\end{aligned}$$

*Proof.* Consider arbitrary $k \in \mathbb{N}$. Since for any $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$, the fact that $\|\frac{1}{2}a - b\|_{H_k^{-1}}^2 \geq 0$ implies that $\|a + b\|_{H_k^{-1}}^2 \leq \frac{3}{2}\|a\|_{H_k^{-1}}^2 + 3\|b\|_{H_k^{-1}}^2$, it follows that

$$\begin{aligned}
\tfrac{1}{2}\|Q_k\|_{H_k^{-1}}^2 &= \tfrac{1}{2}\|\nabla_x \tilde{\phi}(X_k, \mu_k) + Q_k - \nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 \\
&\leq \tfrac{3}{4}\|\nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 + \tfrac{3}{2}\|Q_k - \nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 \\
&\leq \tfrac{3}{4}\|\nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 + \tfrac{3}{2}\lambda_{k,\min}^{-1}\sigma_2^2.
\end{aligned}$$

Hence, the conclusion follows by Parameter Rule 3.3 and Lemmas 3.7 and 3.8. □

Combining the prior three lemmas, we obtain the following result. This result is reminiscent of Lemma 3.5, but accounts for the stochastic gradient errors.

**Lemma 3.12.** *For all $k \in \mathbb{N}$, one finds that*

$$\begin{aligned}
&\mathbb{E}[\tilde{\phi}(X_{k+1}, \mu_{k+1})|\mathcal{F}_k] - \tilde{\phi}(X_k, \mu_k) \\
\leq\ &-\tfrac{1}{4}\gamma_{k,\min}\alpha_{k,\min}\|\nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 \\
&+ (\gamma_{k,\min}\alpha_{k,\text{buff}} + \gamma_{k,\text{buff}}\alpha_{k,\min} + \gamma_{k,\text{buff}}\alpha_{k,\text{buff}})\lambda_{k,\min}^{-1}(\kappa_{\nabla f,\mathcal{B},2} + 2\sqrt{n}\mu_k\theta_{k-1}^{-1})\sigma_2 \\
&+ \tfrac{3}{2}\gamma_{k,\max}^2\alpha_{k,\max}^2\lambda_{k,\min}^{-2}\sigma_2^2.
\end{aligned}$$

*Proof.* Consider arbitrary $k \in \mathbb{N}$. Combining Lemmas 3.9, 3.10, and 3.11,

$$\begin{aligned}
&\mathbb{E}[\tilde{\phi}(X_{k+1}, \mu_{k+1})|\mathcal{F}_k] - \tilde{\phi}(X_k, \mu_k) \\
\leq\ & -\mathbb{E}[\Gamma_k A_k (1 - \tfrac{3}{4}\Gamma_k A_k \lambda_{k,\min}^{-1} \ell_{\nabla f,\mathcal{B},k}) \|\nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 |\mathcal{F}_k] \\
&+ (\gamma_{k,\min}\alpha_{k,\text{buff}} + \gamma_{k,\text{buff}}\alpha_{k,\min} + \gamma_{k,\text{buff}}\alpha_{k,\text{buff}})\lambda_{k,\min}^{-1}(\kappa_{\nabla f,\mathcal{B},2} + 2\sqrt{n}\mu_k\theta_{k-1}^{-1})\sigma_2 \\
&+ \tfrac{3}{2}\gamma_{k,\max}^2 \alpha_{k,\max}^2 \lambda_{k,\min}^{-2} \sigma_2^2.
\end{aligned}$$

Now, one finds under Parameter Rule 3.3 that

$$A_k \leq \tfrac{\lambda_{k,\min}}{\ell_{\nabla f,\mathcal{B},k}} \implies 1 - \tfrac{3}{4}\Gamma_k A_k \lambda_{k,\min}^{-1} \ell_{\nabla f,\mathcal{B},k} \geq 1 - \tfrac{3}{4}\Gamma_k \geq \tfrac{1}{4}.$$

Thus, from above, Parameter Rule 3.3, and Lemma 3.7, the conclusion follows. □

We now show that if the parameter sequences are chosen similarly as in Corollary 3.1, then the coefficients in the upper bound proved in Lemma 3.12 satisfy critical properties for proving our ultimate convergence guarantee.

**Lemma 3.13.** *If for $t \in [-1, -\frac{1}{2})$ in Parameter Rule 3.3 and for some $\underline{r} \in \mathbb{R}_{>0}$ and $\mu_1 \in \mathbb{R}_{>0}$ with $\mu_1 > \frac{\frac{1}{2}\theta_0(\kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty)\Delta}{\frac{1}{2}\Delta - \theta_0}$ the algorithm has $\mu_k = \mu_1 k^t$, $\theta_{k-1} = \theta_0 k^t$, and $\underline{r} \leq \lambda_{k,\min} \leq \lambda_{k,\max}$ for all $k \in \mathbb{N}$, then there is $(\tilde{k}, c, C) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ with*

$$\mathbb{E}[\tilde{\phi}(X_{k+1}, \mu_{k+1})|\mathcal{F}_k] - \tilde{\phi}(X_k, \mu_k) \leq -\underline{r}ck^t \|\nabla_x \tilde{\phi}(X_k, \mu_k)\|_{H_k^{-1}}^2 + Ck^{2t}$$

*for all $k \in \mathbb{N}$ with $k \geq \tilde{k}$.*

*Proof.* Under the conditions and Parameter Rule 3.3, one has for all $k \in \mathbb{N}$ that

$$\begin{aligned}
\gamma_{k,\min}\alpha_{k,\min} &\geq \min\left\{\alpha_{k,\min}, \frac{\alpha_{k,\min}\lambda_{k,\min}\left(\frac{\frac{1}{2}\mu_k\Delta}{\mu_k + \frac{1}{2}(\kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty)\Delta} - \theta_k\right)}{(\alpha_{k,\min} + \alpha_{k,\text{buff}})(\kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty + \mu_k\theta_{k-1}^{-1})}\right\} \\
&=: \min\{\alpha_{k,\min}, \hat{\eta}_k\}.
\end{aligned}$$

The proof can now proceed similarly as in the proof of Corollary 3.1. In particular, with the given parameter choices and by (25), one finds for all $k \in \mathbb{N}$ that

$$\alpha_{k,\min} \geq \tfrac{\underline{r}}{\ell_{\nabla f,\mathcal{B}} + 2\mu_1\theta_0^{-2}k^t(k+1)^{-2t}} \geq \tfrac{\underline{r}}{\ell_{\nabla f,\mathcal{B}} + 2\mu_1\theta_0^{-2}2^{-2t}k^{-t}}, \tag{28}$$

so there exists $\hat{c} \in \mathbb{R}_{>0}$ such that $\alpha_{k,\min} \geq \hat{c}k^t$ for all $k \in \mathbb{N}$. Hence, since $\{\alpha_{k,\text{buff}}\} = \mathcal{O}(k^{2t})$ and $t < 0$, it follows that there exists $\hat{k} \in \mathbb{N}$ such that $\alpha_{k,\text{buff}} \leq \alpha_{k,\min}$ for all $k \in \mathbb{N}$ with $k \geq \hat{k}$, which in turn means that

$$\tfrac{\alpha_{k,\min}}{\alpha_{k,\min} + \alpha_{k,\text{buff}}} \geq \tfrac{1}{2} \ \text{ for all } \ k \in \mathbb{N} \ \text{ with } \ k \geq \hat{k}. \tag{29}$$

For the sequence $\{\hat{\eta}_k\}$, one finds with (6), $t < 0$, and a similar derivation as in the proof of Corollary 3.1 that there exists $\tilde{c} \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$,

$$\frac{\frac{1}{2}\mu_k\Delta}{\mu_k + \frac{1}{2}(\kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty)\Delta} - \theta_k \geq \tilde{c}k^t. \tag{30}$$

Combining (28)–(30), there exists $c \in \mathbb{R}_{>0}$ such that $\frac{1}{4}\gamma_{k,\min}\alpha_{k,\min} \geq \underline{r}ck^t$ for all $k \in \mathbb{N}$ with $k \geq \hat{k}$. On the other hand, the conditions of the lemma and Parameter Rule 3.3 imply for all $k \in \mathbb{N}$ that $\lambda_{k,\min}^{-1} \leq \underline{r}^{-1}$, $\mu_k\theta_{k-1}^{-1} = \mu_1\theta_0^{-1}$, and

$$\alpha_{k,\min} = \tfrac{\lambda_{k,\min}}{\ell_{\nabla f,\mathcal{B}} + 2\mu_1\theta_0^{-2}k^t(k+1)^{-2t}} \leq \tfrac{\lambda_{k,\min}}{\ell_{\nabla f,\mathcal{B}} + 2\mu_1\theta_0^{-2}(k+1)^{-t}}. \tag{31}$$

Combining these facts and Parameter Rule 3.3, one finds $\{\gamma_{k,\min}\alpha_{k,\text{buff}}\lambda_{k,\min}^{-1}\} = \mathcal{O}(k^{2t})$, $\{\gamma_{k,\text{buff}}\alpha_{k,\min}\lambda_{k,\min}^{-1}\} = \mathcal{O}(k^{2t})$, $\{\gamma_{k,\text{buff}}\alpha_{k,\text{buff}}\lambda_{k,\min}^{-1}\} = o(k^{2t})$, and finally $\{\frac{3}{2}\gamma_{k,\max}^2\alpha_{k,\max}^2\lambda_{k,\min}^{-2}\} = \mathcal{O}(k^{2t})$, so the desired conclusion follows from Lemma 3.12. □

We now prove our main convergence theorem for the stochastic setting.

**Theorem 3.2.** *Suppose that Assumptions 2.1 and 3.1 and Parameter Rule 3.3 hold, and that the parameter sequences are chosen as in Lemma 3.13. Then,*

$$\liminf_{k\to\infty} \|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2 = 0 \quad \textit{almost surely,}$$

*meaning that if there exists $\overline{r} \in \mathbb{R}_{>0}$ such that $\lambda_{k,\max} \leq \overline{r}$ for all $k \in \mathbb{N}$, then*

$$\liminf_{k\to\infty} \|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_2^2 = 0 \quad \textit{almost surely.}$$

*Consequently, if all of the aforementioned choices of the parameter sequences are made and, in a given run of the algorithm generating a realization of the iterate sequence $\{x_k\}$ there exists an infinite-cardinality set $\mathcal{K} \subseteq \mathbb{N}$ such that $\{\nabla\tilde{\phi}(x_k,\mu_k)\}_{k\in\mathcal{K}} \to 0$ and $\{x_k\}_{k\in\mathcal{K}} \to \overline{x}$ for some $\overline{x} \in \mathcal{B}$, then the limit point $\overline{x}$ is a KKT point for* (1).

*Proof.* By the Law of Total Expectation, it follows from Lemma 3.13 that there exists $(\tilde{k}, c, C) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$ with $k \geq \tilde{k}$, one has

$$\mathbb{E}[\tilde{\phi}(X_{k+1},\mu_{k+1})] - \mathbb{E}[\tilde{\phi}(X_k,\mu_k)] \leq -\underline{r}ck^t\mathbb{E}[\|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2] + Ck^{2t}.$$

Summing for $k \in \{\tilde{k},\dots,\tilde{k}+K\}$, it follows along with Lemma 3.1 that

$$\begin{aligned} f_{\inf} - \mathbb{E}[\tilde{\phi}(x_{\tilde{k}},\mu_{\tilde{k}})] &\leq \mathbb{E}[\tilde{\phi}(X_{k+1},\mu_{k+1})] - \mathbb{E}[\tilde{\phi}(x_{\tilde{k}},\mu_{\tilde{k}})] \\ &\leq -\underline{r}c\sum_{k=\tilde{k}}^{\tilde{k}+K} k^t\mathbb{E}[\|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2] + C\sum_{k=\tilde{k}}^{\tilde{k}+K} k^{2t}, \end{aligned}$$

which after rearrangement yields

$$\sum_{k=\tilde{k}}^{\tilde{k}+K} k^t\mathbb{E}[\|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2] \leq \tfrac{1}{\underline{r}c}(\mathbb{E}[\tilde{\phi}(x_{\tilde{k}},\mu_{\tilde{k}})] - f_{\inf}) + \tfrac{C}{\underline{r}c}\sum_{k=\tilde{k}}^{\tilde{k}+K} k^{2t}.$$

Under Assumptions 2.1 and 3.1 and since $t \in [-1,-\frac{1}{2})$, the right-hand side of this inequality converges to a finite limit as $K \to \infty$. Since $\sum_{k=1}^{\infty} k^t = \infty$, one finds along with the nonnegativity of $\|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2$ and Fatou's lemma that

$$0 = \liminf_{k\to\infty} \mathbb{E}[\|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2] \geq \mathbb{E}[\liminf_{k\to\infty} \|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2] = 0.$$

Consider the random variable $L := \liminf_{k\to\infty} \|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2$. By nonnegativity of $\|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2$ and the Law of Total Expectation, it follows from above that $0 = \mathbb{E}[L] \geq \mathbb{P}[L > 0]\mathbb{E}[L|L > 0]$, so $0 = \mathbb{P}[L > 0] = \mathbb{P}[\liminf_{k\to\infty} \|\nabla_x\tilde{\phi}(X_k,\mu_k)\|_{H_k^{-1}}^2 > 0]$, which is the first desired conclusion. The second desired conclusion follow from the fact that if $\overline{r}$ exists as stated, then $\|\cdot\|_{H_k^{-1}}^2 \geq \overline{r}^{-1}\|\cdot\|_2^2$ for all $k \in \mathbb{N}$. The last desired conclusion follows using the same argument as in the proof of Theorem 3.1. □

Results similar to Theorem 3.2 have been proved for the stochastic gradient method in the unconstrained setting [9]. In fact, for the stochastic gradient method in an unconstrained setting, one can prove under conditions that are similar to ours that the gradient of the objective function convergences to zero almost surely; this is stronger than a $\liminf$ result of the type in Theorem 3.2. However, such results rely on the gradient of the objective function being Lipschitz continuous. In our setting, we have the Lipschitz-continuity-type property in Lemma 3.2, but the gradient of the (shifted) barrier function is not globally Lipschitz over $\mathcal{B}$, meaning that such a result in the unconstrained setting does not carry over to our present setting.

# 4 Obstacles for a Simplified Algorithm

A reader may wonder if convergence guarantees can be proved for a simpler variant of our algorithm, namely, one that simply employs projections onto the inner neighborhoods of the feasible region. After all, in the setting of (strongly) convex optimization, convergence guarantees exist for projected-(stochastic)-gradient methods; see, e.g., [20]. In this section, we argue that the situation is not straightforward, and even though one may extend our algorithm and analysis to consider a less conservative strategy (recall Remark 2.1), one runs into obstacles when trying to provide a convergence guarantee for a variant of our algorithm that simply uses a projection of $x_k + \alpha_k d_k$ for all $k \in \mathbb{N}$.

Let us consider the setting when $f$ is $\psi$-strongly convex for some $\psi \in \mathbb{R}_{>0}$, and let us consider a simplified variant of our algorithm, where, for all $k \in \mathbb{N}$,

$$H_k = I,\ \alpha_k = \tfrac{1}{\ell_{\nabla f,\mathcal{B}} + 2\mu_k\theta_k^{-2}},\ \text{and}\ x_{k+1} \leftarrow \text{Proj}_{\mathcal{N}_{[l,u]}(\theta_k)}(x_k - \alpha_k \nabla_x \phi(x_k,\mu_k)), \tag{32}$$

where $\text{Proj}_{\mathcal{S}}(\cdot)$ denotes the orthogonal projection operator on convex $\mathcal{S} \subseteq \mathbb{R}^n$. For reasons seen in our subsequent analysis, suppose for all $k \in \mathbb{N}$ that

$$\theta_k \leq c\mu_k, \quad \text{where} \quad c := \min_{i \in \mathcal{L}\cup\mathcal{U}} \left\{ \left( \kappa_{\nabla f,\mathcal{B},\infty} + \tfrac{2\mu_1}{u_i - l_i} \right)^{-1} \right\}. \tag{33}$$

Following a standard approach in the convex optimization literature, let us attempt to prove convergence of the algorithm defined by (32)–(33) by showing that the distance to the unique solution of (1), call it $x_* \in \mathbb{R}^n$, vanishes. Since each iteration of the algorithm makes a step toward minimizing $\tilde{\phi}(\cdot,\mu_k)$, it is natural to approach the analysis by considering the distance between the unique minimizer of this function to $x_*$. A typical result in the literature is that, for sufficiently small barrier parameter values, this distance is proportional to $\mu_k$. For concreteness, we state the following result; for further discussion and a proof, see [28].

**Proposition 4.1.** *Suppose that Assumptions 2.1 and 3.1 hold and the objective $f$ is $\psi$-strongly convex. Let $x_* \in \mathbb{R}^n$ be the unique point such that* (2) *holds for some $(y_*, z_*) \in \mathbb{R}^n \times \mathbb{R}^n$ and let $\bar{x}_k \in \mathbb{R}^n$ denote the unique minimizer of $\tilde{\phi}(\cdot,\mu_k) : \mathbb{R}^n \to \mathbb{R}$ for all $k \in \mathbb{N}$. Then, for all sufficiently large $k \in \mathbb{N}$, it holds that $\|\bar{x}_k - (x_* + \mu_k\zeta)\|_2 = \mathcal{O}(\mu_k^2)$, where the vector $\zeta \in \mathbb{R}^n$ is defined independently from $\{\mu_k\}$.*

Let us now show the expected result that with the step-size choice in (33), the iterate update in (32) corresponds to a step toward the unique minimizer of $\tilde{\phi}(\cdot,\mu_k)$.

**Proposition 4.2.** *Suppose that Assumptions 2.1 and 3.1 hold, the objective $f$ is $\psi$-strongly convex, and the algorithm employs the update in* (32)*–*(33)*. Let $\bar{x}_k \in \mathbb{R}^n$ denote the unique minimizer of $\tilde{\phi}(\cdot,\mu_k) : \mathbb{R}^n \to \mathbb{R}$ for all $k \in \mathbb{N}$. Then, for all $k \in \mathbb{N}$, one finds that $x_k \in \mathcal{N}_{[l,u]}(\theta_{k-1}) \subset \mathcal{N}_{[l,u]}(\theta_k)$, $x_{k+1} \in \mathcal{N}_{[l,u]}(\theta_k)$, and $\bar{x}_k \in \mathcal{N}_{[l,u]}(c\mu_k)$, where $c \in \mathbb{R}_{>0}$ is defined as in* (33)*. Consequently, it follows for all $k \in \mathbb{N}$ that $(x_k, x_{k+1}, \bar{x}_k) \in \mathcal{N}_{[l,u]}(\theta_k) \times \mathcal{N}_{[l,u]}(\theta_k) \times \mathcal{N}_{[l,u]}(\theta_k)$ and*

$$\|x_{k+1} - \bar{x}_k\|_2^2 \leq (1 - \alpha_k\psi)\|x_k - \bar{x}_k\|_2^2.$$

*Proof.* That $x_k \in \mathcal{N}_{[l,u]}(\theta_{k-1}) \subset \mathcal{N}_{[l,u]}(\theta_k)$ and $x_{k+1} \in \mathcal{N}_{[l,u]}(\theta_k)$ hold for all $k \in \mathbb{N}$ follows from (32) and the fact that $\{\theta_k\} \searrow 0$. Now consider arbitrary $k \in \mathbb{N}$ and observe from the definition of $\bar{x}_k$ and Lemma 3.1

that it is the unique vector such that (4) holds with $x = \bar{x}_k$ and $\mu = \mu_k$. Let us prove the desired conclusion that $\bar{x}_k \in \mathcal{N}_{[l,u]}(c\mu_k)$ by proving that $c\mu_k$ is a lower bound for $\bar{s}_{k,i} := \min\{\bar{x}_{k,i} - l_i, u_i - \bar{x}_{k,i}\}$ for all $i \in [n]$. Consider arbitrary $i \in [n]$. If $i \notin \mathcal{L} \cup \mathcal{U}$, then $\bar{s}_{k,i} = \infty \geq c\mu_k$, as desired. If $i \in \mathcal{L} \setminus \mathcal{U}$, then (4) and Assumption 2.1 imply that $\bar{x}_{k,i} - l_i = \mu_k / \nabla_i f(\bar{x}_k) \geq \mu_k / \kappa_{\nabla f, \mathcal{B}, \infty}$, which again yields that $\bar{s}_{k,i} \geq c\mu_k$. Using a similar argument, the bound also holds for $i \in \mathcal{U} \setminus \mathcal{L}$. Finally, if $i \in \mathcal{L} \cap \mathcal{U}$, then assuming without loss of generality that $\bar{x}_{k,i} \leq (u_i + l_i)/2$, it follows that $1/(u_i - \bar{x}_{k,i}) \leq 2/(u_i - l_i)$, so (4) yields

$$\tfrac{1}{\bar{x}_{k,i} - l_i} = \tfrac{\nabla_i f(\bar{x}_k)}{\mu_k} + \tfrac{1}{u_i - \bar{x}_{k,i}} \leq \tfrac{\nabla_i f(\bar{x}_k)}{\mu_k} + \tfrac{2}{u_i - l_i} = \tfrac{\nabla_i f(\bar{x}_k)(u_i - l_i) + 2\mu_k}{\mu_k (u_i - l_i)},$$

which along with Assumption 2.1 and $\bar{s}_{k,i} = \bar{x}_{k,i} - l_i$ yields the desired conclusion. (If $\bar{x}_{k,i} \geq (u_i + l_i)/2$, the conclusion follows using a similar argument with $\bar{s}_{k,i} = u_i - \bar{x}_{k,i}$.)

It has been shown that $(x_k, x_{k+1}, \bar{x}_k) \in \mathcal{N}_{[l,u]}(\theta_k) \times \mathcal{N}_{[l,u]}(\theta_k) \times \mathcal{N}_{[l,u]}(\theta_k)$, as desired. Consequently, using an argument similar to the proof of Lemma 3.2, $\nabla \tilde{\phi}(\cdot, \mu_k)$ is Lipschitz over $\mathcal{N}_{[l,u]}(\theta_k)$ with constant $\ell_{\nabla f, \mathcal{B}} + 2\mu_k \theta_k^{-2}$, which with the choice of $\alpha_k$, the fact that $\tilde{\phi}(\cdot, \mu_k)$ is $\psi$-strongly convex for all $k \in \mathbb{N}$, and [10, Eq. (3.14)] yields

$$\begin{aligned} 0 &\leq \tilde{\phi}(x_{k+1}, \mu_k) - \tilde{\phi}(\bar{x}_k, \mu_k) \\ &\leq \tfrac{1}{\alpha_k}(x_k - x_{k+1})^T (x_k - \bar{x}_k) - \tfrac{1}{2\alpha_k}\|x_k - x_{k+1}\|_2^2 - \tfrac{\psi}{2}\|x_k - \bar{x}_k\|_2^2. \end{aligned}$$

Therefore, it follows that

$$\begin{aligned} \|x_{k+1} - \bar{x}_k\|_2^2 &= \|x_{k+1} - x_k + x_k - \bar{x}_k\|_2^2 \\ &= \|x_{k+1} - x_k\|_2^2 + 2(x_{k+1} - x_k)^T (x_k - \bar{x}_k) + \|x_k - \bar{x}_k\|_2^2 \\ &\leq \|x_{k+1} - x_k\|_2^2 - \|x_k - x_{k+1}\|_2^2 - \alpha_k \psi \|x_k - \bar{x}_k\|_2^2 + \|x_k - \bar{x}_k\|_2^2 \\ &= (1 - \alpha_k \psi)\|x_k - \bar{x}_k\|_2^2, \end{aligned}$$

which is the final desired conclusion. □

Let us now use the prior two results to show a relationship between consecutive distances from an iterate to the solution of (1) that holds for all $k \in \mathbb{N}$.

**Proposition 4.3.** *Suppose that Assumptions 2.1 and 3.1 hold, the objective $f$ is $\psi$-strongly convex, and the algorithm employs the updates in* (32)*–*(33)*. Then, for all sufficiently large $k \in \mathbb{N}$ and $\zeta \in \mathbb{R}^n$ defined as in Proposition 4.1, one has*

$$\|x_{k+1} - x_*\|_2 \leq \sqrt{1 - \alpha_k \psi}\|x_k - x_*\|_2 + 2\mu_k \|\zeta\|_2 + \mathcal{O}(\mu_k^2). \tag{34}$$

*Proof.* Combining Propositions 4.1 and 4.2 and the triangle inequality, one has

$$\begin{aligned} \|x_{k+1} - x_*\|_2 &\leq \|x_{k+1} - \bar{x}_k\|_2 + \|\bar{x}_k - x_*\|_2 \\ &\leq \sqrt{1 - \alpha_k \psi}\|x_k - \bar{x}_k\|_2 + \|\bar{x}_k - x_*\|_2 \\ &\leq \sqrt{1 - \alpha_k \psi}(\|x_k - x_*\|_2 + \|\bar{x}_k - x_*\|_2) + \|\bar{x}_k - x_*\|_2 \\ &\leq \sqrt{1 - \alpha_k \psi}\|x_k - x_*\|_2 + 2\mu_k \|\zeta\|_2 + \mathcal{O}(\mu_k^2), \end{aligned}$$

as desired. □

At first glance, the result of Proposition 4.3 might appear to be useful since $\{\mu_k\} \searrow 0$ also implies that $\{\alpha_k\} \searrow 0$. Unfortunately, however, the last two terms on the right-hand side of (34) obstruct an ability to prove that $\{x_k\} \to x_*$, even in this strongly convex and deterministic setting. To see this, note that the recurrence defined by (34) has the form of sequences $\{u_k\}$, $\{v_k\}$, and $\{e_k\}$ such that

$$u_{k+1} \leq v_k u_k + e_k, \quad u_k \in \mathbb{R}_{\geq 0}, \quad v_k \in [0, 1), \quad \text{and} \quad e_k \in \mathbb{R}_{>0} \quad \text{for all} \quad k \in \mathbb{N}.$$

Such a recurrence yields $\{u_k\} \to 0$ if $\sum_{k=1}^{\infty}(1 - v_k) = \infty$ and $\lim_{k\to\infty} e_k/(1 - v_k) = 0$; see, e.g., [22, pg. 45]. However, with $v_k := \sqrt{1 - \alpha_k\psi}$ and $e_k := C\mu_k$ for some $C \in \mathbb{R}_{>0}$ (for simplicity), and for example supposing that $\theta_k = c\mu_k$ for all $k \in \mathbb{N}$ (see (33)), one indeed finds $\sum_{k=1}^{\infty}(1 - v_k) = \infty$, but

$$\frac{C\mu_k}{1-\sqrt{1-\alpha_k\psi}} = \frac{C\mu_k}{1-\sqrt{1-\frac{\psi}{\ell_{\nabla f,\mathcal{B}}+2c^{-2}\mu_k^{-1}}}} \xrightarrow{k\to\infty} \frac{C4c^{-2}}{\psi} > 0.$$

Consequently, Proposition 4.3 does not readily lead to a convergence guarantee for the simplified algorithm stated in (32)–(33). One might modify the algorithm and/or analysis in this section to reach such a guarantee in the deterministic setting upon which one might build a convergence theory for a stochastic algorithm, but we contend that the ultimate conclusions would only be comparable to those for the algorithm analyzed in Section 3, perhaps with the extensions mentioned in Remark 2.1.

# 5 Numerical Results

Our numerical experiments serve two main purposes: (1) we demonstrate that our (stochastic)-interior-point method, which we refer to as SIPM, is reliable over a well known set of test problems, and (2) we compare the performance of SIPM with a projected-(stochastic)-gradient method, which we refer to as PSGM. We implemented a set of test problems and the algorithms in Matlab. The experiments were conducted on the High Performance Computing cluster at Lehigh University with Matlab R2021b using the Deep Learning Toolbox.

## 5.1 Test problems

We tested the algorithms by training prediction models for binary classification using data from LIBSVM [13]. From LIBSVM, we selected the 43 binary classification datasets with training data file size at most 8 GB; for these, the numbers of features are in the range $[2, 47263]$ and the numbers of data points are in the range $[44, 5000000]$. We provide results pertaining to training data, and for those datasets with corresponding testing data, we provide results pertaining to that data as well. Each dataset from LIBSVM consists of $A \in \mathbb{R}^{m\times n_f}$ and $b \in \{-1, 1\}^m$, where $m$ is the number of data points and $n_f$ is the number of features.

To cover both a convex and a nonconvex objective, we consider two models: (1) logistic regression and (2) a neural network with one hidden layer and a cross-entropy loss function. For training a (convex) logistic regression model, the number of optimization variables is the number of features plus one for the bias term, i.e., $n = n_f + 1$. The (nonconvex) neural network model consists of a fully connected hidden layer with $h$ neurons and tanh activation and a fully connected output layer with sigmoid activation. The number of optimization variables is the number of weights plus bias terms at each node in the hidden and output layers, so $n = (n_f + 2)h + 1$, where $h := \max\{2, \min\{\lceil \frac{n_f}{2} \rceil, 100\}\}$. For both models, we set $l = -1 \times \mathbb{1}$ and $u = 1 \times \mathbb{1}$, which causes many bounds to be active at a solution. Table 1 (pg. 24) shows the number of variables for each problem, i.e., objective and dataset pair.

## 5.2 Implementation details

We generated $x_1$ for each problem with elements drawn from a uniform distribution over $[-0.01, 0.01]$. This point was fixed for all runs.

SIPM requires the problem-dependent parameters $\kappa_{\nabla f,\mathcal{B},\infty}$, $\ell_{\nabla f,\mathcal{B}}$, and $\sigma_\infty$. For consistency across our experiments in both the deterministic and stochastic settings, we employed estimates $\overline{\kappa_{\nabla f,\mathcal{B},\infty}}$ and $\overline{\ell_{\nabla f,\mathcal{B}}}$, which were set by: (1) temporarily setting these values to 1; (2) running 500 iterations of SIPM using true gradients, these temporary values, and the remaining parameters set as in the next paragraph; and (3) setting, at termination, the values as $\overline{\kappa_{\nabla f,\mathcal{B},\infty}} \leftarrow \max_{k\in[500]}\{\|\nabla f(x_k)\|_\infty\}$ and $\overline{\ell_{\nabla f,\mathcal{B}}} \leftarrow \max_{k\in[500]\setminus\{1\}}\{\|\nabla f(x_{k-1}) - \nabla f(x_k)\|_2/\|x_{k-1} - x_k\|_2\}$. For the deterministic setting, we set $\overline{\sigma_\infty} \leftarrow 0$, whereas for the stochastic setting we employed an estimate $\overline{\sigma_\infty}$ for each dataset, which was set by: (1) generating 100 stochastic gradients at $x_1$ with a mini-batch size of $\lceil 0.01m \rceil$ (which was also the mini-batch size used in all of our experiments for the stochastic setting) and (2) setting $\overline{\sigma_\infty}$ as the maximum $\infty$-norm difference between each of these stochastic gradients and $\nabla f(x_1)$. Once all of these values were computed—see Table 1—they were fixed for all of our experiments.

Since the performance of an interior-point method is affected by the initial and final values of the barrier parameter, we recommend choosing $\{\mu_k\}$ and $\{\theta_k\}$ based on the computational budget. Hence, let us define `maxiter` as an iteration limit for the deterministic setting and `maxiter` = (number of epochs)/0.01 for the stochastic setting, where for the former our experiments consider `maxiter` $\in \{100, 1000\}$ and for the latter our experiments consider the number of epochs in $\{1, 1000\}$. (Note that setting `maxiter` as above for the stochastic setting is consistent with the mini-batch size of $\lceil 0.01m \rceil$.) Using this value, and letting $g(x_1)$ denote the gradient (estimate) at the initial point in a run, the parameters for SIPM were set as

$$\begin{aligned}
&\bar{\Delta} \leftarrow 100, \quad \Delta \leftarrow \min\{\bar{\Delta}, \min_{i\in[n]}\{u_i - l_i\}\}, \\
&\mu_1 \leftarrow \max\left\{10^{-5}, \min\left\{\tfrac{10^{-3}\|g(x_1)\|_2}{\|\Psi(u-x_1)^{-1}-\Psi(x_1-l)^{-1}\|_2}, 1\right\}\right\}, \quad \text{and} \\
&\theta_0 \leftarrow \min\left\{\min_{i\in[n]}\{x_i - l_i\}, \min_{i\in[n]}\{u_i - x_i\}, \bar{\theta}_0\right\}, \quad \text{where } \bar{\theta}_0 \leftarrow \frac{1}{\frac{2}{\Delta} + \frac{\overline{\kappa_{\nabla f,\mathcal{B},\infty}} + \overline{\sigma_\infty}}{\mu_1}},
\end{aligned}$$

as well as $\mu_k \leftarrow \mu_1 s_k$ and $\theta_k \leftarrow \theta_0 s_k$ for all $k \in [\texttt{maxiter}]$, where $\{s_k\}$ is composed of equal-length repetitions of the elements in $\{1, 0.1, \cdots, 10^{-8}/\mu_1\}$, i.e., $\{s_k\} = \{1, \dots, 1, 0.1, \dots, 0.1, \dots, 10^{-8}/\mu_1, \dots, 10^{-8}/\mu_1\}$. In this manner, $\mu_1$ ensures that the initial search direction is not dominated by the log-barrier term, whereas $\mu_{\texttt{maxiter}} = 10^{-8}$ ensures that SIPM terminates with a prescribed small barrier parameter. For the remaining parameters, the implementation used for all $k \in \mathbb{N}$: $H_k \leftarrow \overline{\ell_{\nabla f,\mathcal{B}}} I + \mu_k(\Psi(x_k - l))^{-2} + \mu_k(\Psi(u - x_k))^{-2}$, $\lambda_{k,\min} \geq \overline{\ell_{\nabla f,\mathcal{B}}}$ as the smallest eigenvalue of $H_k$, $\alpha_{k,\text{buff}} \leftarrow (\texttt{maxiter}/k)^{1.1} \geq 1$, $\gamma_{k,\text{buff}} \leftarrow (\texttt{maxiter}/k)^{0.55} \geq 1$, and $(\alpha_k, \gamma_{k,\max})$ as in Assumption 3.2. By choosing $\alpha_{k,\text{buff}}$ and $\gamma_{k,\text{buff}}$ in this manner, SIPM employs $\alpha_k = \lambda_{k,\min}/\ell_{\nabla f,\mathcal{B},k}$ and $\gamma_{k,\max} = 1$ for all $k \in [\texttt{maxiter}]$. Other formulas for $\alpha_{k,\text{buff}}$ and $\gamma_{k,\text{buff}}$ were tested; the values above worked best for our experiments.

For PSGM, in order to have a direct comparison with SIPM, the step sizes were also set using the sequence $\{s_k\}$ defined in the previous paragraph in such a manner that the initial and final step sizes for PSGM were the same as those used by SIPM. We remark in passing that PSGM has convergence(-to-neighborhood) guarantees when a fixed step size is used, but we did not experiment with such a choice since our aim is to compare with SIPM, which is only defined for diminishing step sizes.

## 5.3 Comparison of SIPM and PSGM

All runs of both algorithms terminated when the iteration limit was reached. To compare performance, we considered two measures at the final iterate: the objective value $f(x_{\texttt{maxiter}})$ (computed over the training set and testing set, when available) and the norm of a projected gradient $\|\operatorname{Proj}_{\mathcal{N}_{[l,u]}(0)}(x_{\texttt{maxiter}} - \nabla f(x_{\texttt{maxiter}})) - x_{\texttt{maxiter}}\|_\infty$; see, e.g., [8]. (Even for the stochastic setting, the projected gradient at the final iterate was computed using the true gradient for the purpose of our comparison.) In particular, for all runs and each measure, we computed a relative performance measure; e.g., in terms of $f$, we use

$$r_p := \tfrac{f(x_{\texttt{maxiter}}^{\text{SIPM}}) - f(x_{\texttt{maxiter}}^{\text{PSGM}})}{\max\left\{f(x_{\texttt{maxiter}}^{\text{SIPM}}), f(x_{\texttt{maxiter}}^{\text{PSGM}}), 1\right\}} \in [-1, 1] \quad \text{for } p \in \text{set of problems},$$

and likewise for the norm of the projected gradient. The values are within $[-1, 1]$ since the final objective values and projected-gradient norms are nonnegative.

Figure 1 provides relative performance measures for runs for solving (convex) logistic regression problems. The bar plot in the first column is for final objective values with respect to the training data when `maxiter` = 100 and the plot in the middle column is for projected-gradient norms with respect to the training data when `maxiter` = 100. These show that, within a relatively small iteration limit, SIPM can outperform PSGM.

Table 1: Problem sizes and algorithmic parameters.

| dataset | logistic regression | | | | neural network + cross-entropy loss | | | |
|---|---|---|---|---|---|---|---|---|
| | $n$ | $\overline{\ell_{\nabla f,\mathcal{B}}}$ | $\overline{\kappa_{\nabla f,\mathcal{B},\infty}}$ | $\overline{\sigma_\infty}$ | $n$ | $\overline{\ell_{\nabla f,\mathcal{B}}}$ | $\overline{\kappa_{\nabla f,\mathcal{B},\infty}}$ | $\overline{\sigma_\infty}$ |
| a1a | 124 | 1.36 | 0.14 | 0.35 | 7751 | 3.12 | 0.25 | 0.34 |
| a2a | 124 | 1.38 | 0.14 | 0.22 | 7751 | 3.05 | 0.24 | 0.21 |
| a3a | 124 | 1.36 | 0.13 | 0.33 | 7751 | 3.01 | 0.20 | 0.29 |
| a4a | 124 | 1.37 | 0.14 | 0.16 | 7751 | 2.95 | 0.22 | 0.15 |
| a5a | 124 | 1.37 | 0.14 | 0.16 | 7751 | 2.99 | 0.19 | 0.15 |
| a6a | 124 | 1.36 | 0.13 | 0.11 | 7751 | 3.00 | 0.21 | 0.11 |
| a7a | 124 | 1.36 | 0.14 | 0.11 | 7751 | 2.97 | 0.20 | 0.09 |
| a8a | 124 | 1.36 | 0.13 | 0.09 | 7751 | 2.98 | 0.22 | 0.08 |
| a9a | 124 | 1.36 | 0.13 | 0.07 | 7751 | 3.00 | 0.23 | 0.07 |
| australian | 15 | 0.22 | 0.14 | 0.44 | 113 | 0.25 | 0.07 | 0.44 |
| breast-cancer | 11 | 0.22 | 0.15 | 0.51 | 61 | 0.28 | 0.11 | 0.51 |
| cod-rna | 9 | 0.88 | 0.02 | 0.05 | 41 | 0.25 | 0.12 | 0.05 |
| colon-cancer | 2001 | 1.66 | 0.10 | 0.68 | 200201 | 2.09 | 0.11 | 0.65 |
| covtype | 55 | 0.34 | 0.03 | 0.02 | 1513 | 4.00 | 0.45 | 0.02 |
| diabetes | 9 | 0.55 | 0.08 | 0.40 | 41 | 0.25 | 0.12 | 0.40 |
| duke | 7130 | 4.48 | 0.14 | 0.65 | 713101 | 21.69 | 0.75 | 0.53 |
| fourclass | 3 | 0.37 | 0.09 | 0.42 | 9 | 0.25 | 0.11 | 0.42 |
| german | 25 | 1.63 | 0.14 | 0.50 | 313 | 3.90 | 0.25 | 0.50 |
| gisette_scale | 5001 | 6.27 | 0.50 | 0.26 | 500201 | 25.25 | 0.50 | 0.18 |
| heart | 14 | 0.15 | 0.08 | 0.56 | 106 | 0.25 | 0.06 | 0.56 |
| ijcnn1 | 23 | 0.32 | 0.28 | 0.03 | 265 | 0.25 | 0.30 | 0.03 |
| ionosphere_scale | 35 | 1.22 | 0.11 | 0.64 | 613 | 0.66 | 0.10 | 0.64 |
| leu | 7130 | 1.08 | 0.06 | 0.76 | 713101 | 6.69 | 0.89 | 0.72 |
| liver-disorders | 6 | 0.58 | 0.06 | 0.62 | 22 | 0.25 | 0.09 | 0.62 |
| madelon | 501 | 74.76 | 0.29 | 0.25 | 50201 | 9.08 | 0.50 | 0.25 |
| mushrooms | 113 | 1.30 | 0.13 | 0.17 | 6385 | 0.60 | 0.07 | 0.16 |
| phishing | 69 | 3.34 | 0.48 | 0.14 | 2381 | 4.52 | 0.54 | 0.13 |
| rcv1_train | 47237 | 0.05 | 0.02 | 0.12 | 4.72e+6 | 0.23 | 0.02 | 0.12 |
| real-sim | 20959 | 0.25 | 0.14 | 0.04 | 2.10e+6 | 0.25 | 0.14 | 0.04 |
| skin_nonskin | 4 | 0.44 | 0.18 | 0.02 | 11 | 0.25 | 0.22 | 0.02 |
| sonar_scale | 61 | 2.80 | 0.41 | 0.54 | 1861 | 4.17 | 0.41 | 0.53 |
| splice | 61 | 5.57 | 0.52 | 0.52 | 1861 | 6.46 | 0.52 | 0.52 |
| SUSY | 19 | 0.34 | 0.03 | 0.01 | 181 | 0.25 | 0.03 | 0.01 |
| svmguide1 | 5 | 0.35 | 0.11 | 0.26 | 13 | 0.25 | 0.11 | 0.26 |
| svmguide3 | 23 | 0.61 | 0.11 | 0.38 | 265 | 0.25 | 0.20 | 0.38 |
| w1a | 301 | 0.59 | 0.24 | 0.14 | 30201 | 0.34 | 0.35 | 0.13 |
| w2a | 301 | 0.60 | 0.23 | 0.12 | 30201 | 0.34 | 0.35 | 0.11 |
| w3a | 301 | 0.60 | 0.24 | 0.10 | 30201 | 0.34 | 0.35 | 0.09 |
| w4a | 301 | 0.60 | 0.24 | 0.10 | 30201 | 0.34 | 0.35 | 0.05 |
| w5a | 301 | 0.60 | 0.24 | 0.08 | 30201 | 0.34 | 0.35 | 0.08 |
| w6a | 301 | 0.61 | 0.23 | 0.05 | 30201 | 0.34 | 0.35 | 0.03 |
| w7a | 301 | 0.61 | 0.23 | 0.04 | 30201 | 0.34 | 0.35 | 0.03 |
| w8a | 301 | 0.61 | 0.23 | 0.03 | 30201 | 0.34 | 0.35 | 0.02 |

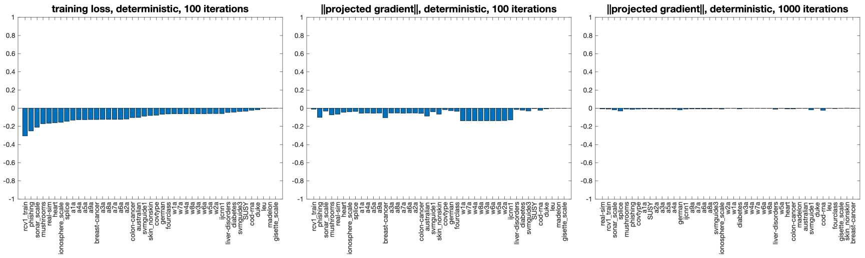

Figure 1: Relative performance of SIPM and PSGM in the deterministic setting when solving logistic regression problems.

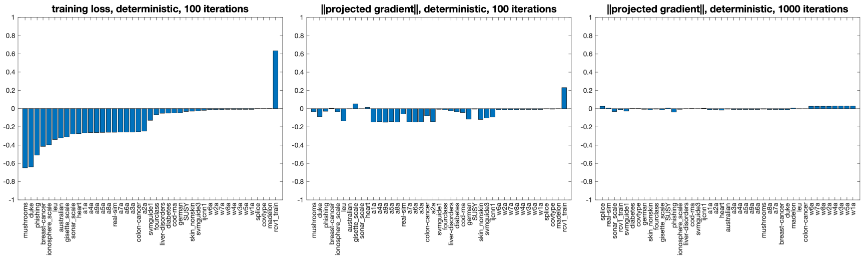

Figure 2: Relative performance of SIPM and PSGM in the deterministic setting when training neural network models (with one hidden layer) with cross-entropy loss.

The bar plot in the third column is for projected-gradient norms with respect to the training data when `maxiter` = 1000. This plot shows that, with a more substantial budget, both algorithms reach points that are nearly stationary, which shows in the deterministic setting that SIPM is as reliable as PSGM. Figure 2 provides similar results for the (nonconvex) neural-network-training problems.

Figures 3 and 4 provide results for the stochastic setting in the form of box plots when each algorithm is employed to solve each problem 10 times. The first rows in each figure consider runs over 1 epoch while the second rows consider runs over 1000 epochs. The first columns are for training loss, the middle columns are for projected-gradient norms over the training data, and the third columns are for testing loss. Corresponding to the goals of our experiments, these results show that SIPM performs well compared to PSGM when the budget is relatively small, and is as reliable as PSGM when the budget is large.

# 6 Conclusion

We have proposed, analyzed, and provided the results of numerical experiments with a stochastic interior-point method for solving continuous bound-constrained optimization problems. The algorithm is unique in various aspects (see Section 1.1). In future work, it will be interesting to pair the algorithmic strategies proposed in this paper with stochastic approximation strategies for solving equality-constrained problems toward the complete design of stochastic interior-point methods for solving generally constrained optimization problems.

# Acknowledgments

This work is supported by the U.S. National Science Foundation (NSF) award CCF-2139735, U.S. Office of Naval Research award N00014-21-1-2532, and OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765. Portions were conducted on Lehigh University's Research Computing infrastructure partially supported by the NSF award 2019035.

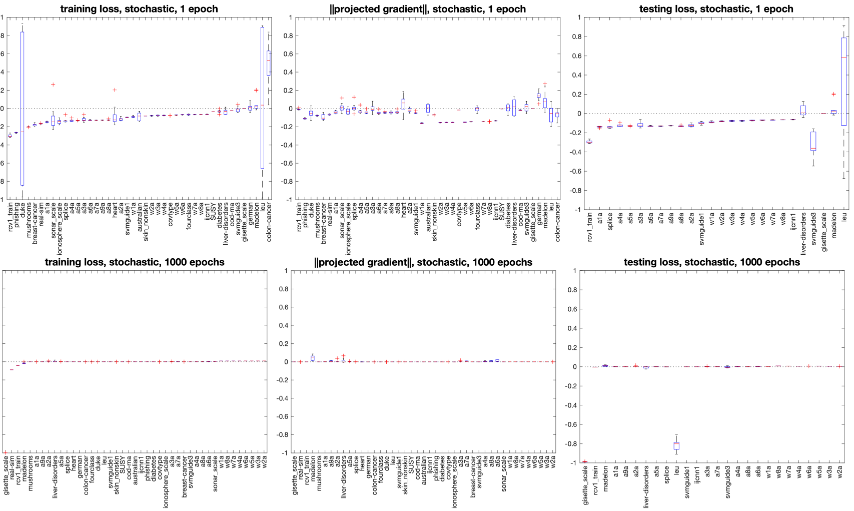

Figure 3: Relative performance of SIPM and PSGM in the stochastic setting (over 10 runs for each problem) when solving logistic regression problems. Among the 43 datasets considered for our test problems, there are 26 with corresponding testing datasets (see last column)

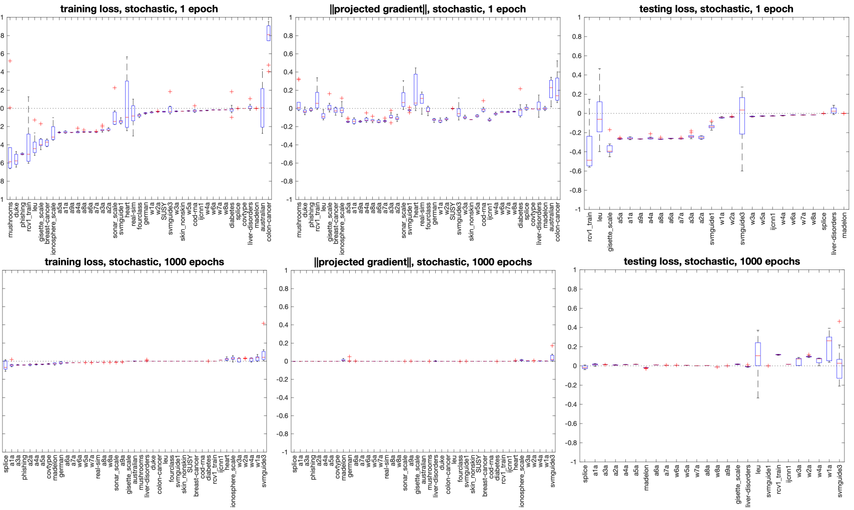

Figure 4: Relative performance of SIPM and PSGM in the stochastic setting (over 10 runs for each problem) when training neural network models (with one hidden layer) with cross-entropy loss.

# References

[1] MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.*, 2019.

[2] Riley Badenbroek and Etienne de Klerk. Complexity analysis of a sampling-based interior point method for convex optimization. *Math. of Operations Research*, 47(1):779–811, 2022.

[3] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, 2021.

[4] Albert S Berahas, Raghu Bollapragada, and Baoyu Zhou. An adaptive sampling sequential quadratic programming method for equality constrained stochastic optimization. *arXiv 2206.00712*, 2022.

[5] Albert S Berahas, Frank E Curtis, Michael J O'Neill, and Daniel P Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient jacobians. *arXiv:2106.13015*, 2021.

[6] Albert S Berahas, Jiahao Shi, Zihong Yi, and Baoyu Zhou. Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction. *arXiv 2204.04161*, 2022.

[7] W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-lipschitz and nonconvex minimization. *Math. Prog.*, 149:301–327, 2015.

[8] Ernesto G Birgin, José Mario Martínez, and Marcos Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211, 2000.

[9] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, 2018.

[10] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[11] Richard H Byrd, Mary E Hribar, and Jorge Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.

[12] Peter Carbonetto, Mark Schmidt, and Nando Freitas. An interior-point stochastic approximation method and an l1-regularized delta rule. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Neural Information Processing Systems*. Curran Assoc., Inc., 2008.

[13] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[14] F. E. Curtis, D. P. Robinson, and B. Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv 2107.03512*, 2021.

[15] Frank E Curtis, Michael J O'Neill, and Daniel P Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *arXiv 2112.14799*, 2021.

[16] Yuchen Fang, Sen Na, Michael W. Mahoney, and Mladen Kolar. Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *arXiv 2211.15943*, 2022.

[17] Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Math. Prog.*, pages 1–71, 2022.

[18] Sen Na and Michael W Mahoney. Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. *arXiv 2205.13687*, 2022.

[19] Hariharan Narayanan. Randomized interior point methods for sampling and optimization. *The Annals of Applied Probability*, 26(1):597–641, 2016.

[20] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Jour. on Opt.*, 19(4):1574–1609, 2009.

[21] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.

[22] Boris T. Polyak. *Introduction to Optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, New York, 1987.

[23] Songqiang Qiu and Vyacheslav Kungurtsev. A sequential quadratic programming method for optimization with stochastic objective functions, deterministic inequality constraints and robust subproblems. *arXiv 2302.07947*, 2023.

[24] Jos F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(1-4):625–653, 1999.

[25] K.C. Toh, M. J. Todd, and R.H. Tutuncu. Sdpt3 — a matlab software package for semidefinite programming, version 1.3. *Optimization Methods and Software*, 11(1-4):545–581, 1999.

[26] R. J. Vanderbei and D. F. Shanno. An Interior-Point Algorithm for Nonconvex Nonlinear Programming. *Computational Optimization and Applications*, 13:231–252, 1999.

[27] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Prog.*, 106(1):25–57, 2006.

[28] Stephen J Wright and Dominique Orban. Properties of the log-barrier function on degenerate nonlinear programs. *Mathematics of Operations Research*, 27(3):585–613, 2002.