Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Exponential Concentration in Stochastic Approximation

Kody J.H. Law
School of Mathematics, University of Manchester, Manchester, M13 9PL, kody.law@manchester.ac.uk
Neil Walton
Durham University Business School, Millhill Lane, Durham, DH1 3LB, UK neil.walton@durham.ac.uk
Shangda Yang
School of Mathematics, University of Manchester, Manchester, M13 9PL, shangda.yang@manchester.ac.uk

We analyze the behavior of stochastic approximation algorithms where iterates, in expectation, progress towards an objective at each step. When progress is proportional to the step size of the algorithm, we prove exponential concentration bounds. These tail-bounds contrast asymptotic normality results, which are more frequently associated with stochastic approximation. The methods that we develop rely on a proof of geometric ergodicity. This extends the result of Markov chains due to Hajek (1982) to stochastic approximation algorithms. We apply our results to several different stochastic approximation algorithms, specifically Projected Stochastic Gradient Descent, Kiefer-Wolfowitz, and Stochastic Frank-Wolfe algorithms. When applicable, our results prove faster $O(1/t)$ and linear convergence rates for Projected Stochastic Gradient Descent with a non-vanishing gradient.

Key words: stochastic approximation, projected stochastic gradient descent, concentration bounds.

1. Introduction

We consider stochastic approximation algorithms where the expected progress toward the optimum is proportional to the algorithm's step size. For instance, a stochastic gradient descent algorithm applied to a convex function will satisfy this property when bounded away from the optimum. However, this property can continue to hold as an algorithm approaches the optimum. For instance, a stochastic gradient descent algorithm applied to a convex function will satisfy this property when bounded away from the optimum. However, this property can continue to hold as an algorithm approaches the optimum. As we will discuss, this is true when the convex objective function is *sharp*. Stated informally, these objectives have a V-shape at the optimum rather than a quadratic

U-shape. The latter case is extensively studied. The asymptotic error has a normal distribution; see Chung (1954), Fabian (1968). However, in the former case, little is known about the limit distribution of the error. In these settings, we will show that the error for algorithms such as Projected Stochastic Gradient Descent, Kiefer-Wolfowitz, and Frank-Wolfe have an exponential concentration and a faster rate of convergence than would be anticipated by standard results for stochastic optimization with a smooth objective. We develop methods which are typically used in probability to analyze random walks or in applied probability to analyze queueing networks. For stochastic approximation, our results establish new exponential concentration bounds.

We now summarize the background and problems where our results apply.

Stochastic Gradient Descent: Standard Asymptotic Results. Due to its applicability in machine learning, there is now a vast literature on stochastic gradient descent (Bottou et al. 2018). The rate of convergence found to the optimal point for a (projected) stochastic gradient descent procedure on a convex objective has order $O(1/\sqrt{t})$ of the optimum after t-iterations of the algorithm (Nemirovski et al. 2009, Moulines and Bach 2011, Bottou et al. 2018). In this paper, we find conditions under which the improved $O(1/t)$ convergence rate holds, and developing on the work of Davis et al. (2019), we also find linear convergence results. Our results apply to optimization problems where the gradient does not vanish as we approach the optimum. A critical feature of our analysis is an exponential concentration bound.

Asymptotic Normality, Exponential Bounds, and Reflected Random Walks. For stochastic approximation, the normal distribution has long been known to characterize the limiting behavior of a stochastic approximation procedure. See Fabian (1968) and Chapter 10 of Kushner and Yin (2003). Such theories are statistically efficient for smooth optimization problems with and without constraints. See Duchi and Ruan (2021), Davis et al. (2023) and Moulines and Bach (2011) respectively, and results are motivated by the asymptotic normality results for maximum likelihood estimators (MLE) of Le Cam (1953) and Hájek (1972). However, one should note that such asymptotics may not always lead to a Gaussian limit. For example, the MLE of a uniform distribution is not asymptotically normal but is instead exponentially distributed. (See Section EC.1 of the E-companion for a proof.) The stochastic optimization algorithms considered in this paper are settings where the normal distribution is not asymptotically optimal.

While asymptotic normality has a long history, the exponential bounds found here are not well-understood and do not appear in stochastic approximation literature. We argue that when the objective's gradient is non-vanishing as we approach the optimum, the normal approximation will not hold, and an exponential concentration bound is more appropriate.

We establish exponential concentration bounds using a geometric Lyapunov bound. These arguments are commonly employed to establish the exponential ergodicity of Markov chains. See

Kendall's Renewal Theorem in Chapter 15 of Meyn and Tweedie (2012). Hajek (1982), in particular, provides a proof that converts a drift condition into an exponential Martingale that establishes fast convergence rates for ergodic Markov chains. A key contribution of this paper is to extend this argument to stochastic approximation.

These bounds are typically applied to queueing networks (Kingman 1964, Bertsimas et al. 2001) because many queueing processes are random walks with constraints and non-zero drift. These conditions lead to exponential distribution bounds (Harrison and Williams 1987). Kushner and Yin (2003) discusses these connections when analyzing the diffusion approximation of stochastic approximation procedures with constraints. Nonetheless, as we will discuss, diffusion analysis does not fully recover the required exponential concentration. The concentration results proven here are, to the best of our knowledge, new in the context of stochastic approximation.

Constrained Stochastic Gradient Descent, Sharp Functions, and Geometric Convergence. Our results are applicable when the gradient of the function does not vanish. In particular, our results can be applied to constrained stochastic approximation when the optimum lies on the boundary. The text Kushner and Clark (1978a) analyses the convergence of stochastic approximation algorithms on constrained regions. Buche and Kushner (2002) prove convergence rates. These authors observe that analysis typically applied to unconstrained stochastic approximation does not readily apply to the constrained case.

Boundary constraints are not a requirement of our analysis. Our results apply under a non-vanishing gradient condition. This closely relates to the property of a function being *sharp*. Davis et al. (2019) presents a variety of machine learning tasks for which the objective is sharp. We show that our non-vanishing gradient condition is equivalent to sharpness for convex functions. Our exponential concentration bounds are tighter than Gaussian concentration bounds. Applying this concentration bound to the work of Davis et al. (2019) leads to an improved linear convergence rate for projected stochastic gradient descent. Recent work by Davis et al. (2023) analyses the asymptotic normality of stochastic gradient descent algorithms, which exhibit sharpness away from a smooth manifold around the optimum.

Further Stochastic Approximation Algorithms. The main result of the paper considers a generic stochastic algorithm with non-vanishing drift and sub-exponential noise (Conditions (C1) and (C2)). For this reason, our results hold for other mainstream stochastic approximation algorithms. We consider the Kiefer-Wolfowitz and the Frank-Wolfe (or conditional gradient algorithm) as examples. See Kiefer and Wolfowitz (1952) and Frank and Wolfe (1956).

Kiefer-Wolfowitz is the primary alternative to the Robbins and Monro (1951) stochastic gradient descent algorithm. Here, gradient estimates are replaced by a finite difference approximation. We

prove that exponential concentration and linear convergence hold for Kiefer-Wolfowitz under a non-vanishing drift condition.

Frank-Wolfe is a popular projection-free alternative to projected gradient descent algorithms. See Jaggi (2013), Hazan and Kale (2012). The stochastic Frank-Wolfe algorithm is proposed and analyzed in Hazan and Luo (2016). We provide conditions analogous to sharpness along with extra critical conditions that ensure exponential concentration for the stochastic Frank-Wolfe algorithm. Linear convergence analogous to Davis et al. (2019) can also occur for these algorithms.

The results as a whole establish a sequence of connections between stochastic modeling bounds used in queueing and stochastic approximation. These exponential tail bounds differ from Gaussian concentration bounds typically analyzed in unconstrained stochastic approximation. Moreover, these results lead to faster convergence rates than standard stochastic approximation results.

1.1. Organization.

This article is structured as follows. Section 2 gives initial notation. (Further notation will be introduced as we present each of our results.) Section 3 presents the paper's main results. Section 3.1 provides intuition on exponential concentration. Section 3.2 presents a generic Lyapunov function result for exponential concentration. Section 3.3 applies our results to Projected Stochastic Gradient Descent (PSGD). In Section 3.4, we provide an exponential concentration bound for the Kiefer-Wolfowitz stochastic approximation algorithm. In Section 3.5, we give an exponential concentration bound for the Stochastic Frank-Wolfe algorithm. A linear convergence result for PSGD is presented in Section 3.6. Proofs for the results are given in Section 4, with later results deferred to the E-companion. Numerical experiments are presented in Section 5. In Section 6, we show that, although exponential concentration holds, the exponential distribution is not, in general, the limiting distribution when gradients do not vanish. We discuss the interplay between the normal approximation and exponential approximation, and we conjecture the asymptotic optimal performance under sharpness.

2. Problem setting and initial assumptions

We provide some basic notation and assumptions that hold throughout this paper. The algorithms and results considered will require some specific assumptions, which will be presented in the sections relevant to those results.

Basic Notation. We apply the convention that $\mathbb{Z}_+ = \{n : n = 0, 1, 2, ...\}$ and $\mathbb{R}_+ = \{x : x \geq 0\}$. Implied multiplication has precedence over division, that is, $2a/3bc = (2 \times a)/(3 \times b \times c)$.

Optimization Notation. Unless explicitly stated otherwise, we let $\mathcal{X}$ denote a nonempty closed bounded convex subset of $\mathbb{R}^d$. For a continuous function $f : \mathcal{X} \to \mathbb{R}$, we consider the minimization

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})\,. \tag{1}$$

We let $\mathcal{X}^\star$ be the set of minimizers of the above optimization problem. We let $\Pi_{\mathcal{X}}(\boldsymbol{x})$ denote the projection of x onto the set $\mathcal{X}$. That is $\Pi_{\mathcal{X}}(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathcal{X}} ||\boldsymbol{x}-\boldsymbol{y}||^2$, where $||\cdot||$ denotes the Euclidean norm. We let $d(\boldsymbol{x},\mathcal{X})$ denote the distance from a point to its projection. That is $d(\boldsymbol{x},\mathcal{X}) := \min_{\boldsymbol{y}\in\mathcal{X}} ||\boldsymbol{x}-\boldsymbol{y}||$. We let relint$(\mathcal{X})$ denote the relative interior of $\mathcal{X}$. We let F be the gap between the maximum and minimum of $f(\boldsymbol{x})$ on $\mathcal{X}$, that is

$$F := \max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) - \min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})\,.$$

Stochastic Iterations. We consider a generic stochastic iterative procedure for solving the optimization problem (1). Consider a random sequence $\{\boldsymbol{x}_t\}_{t=0}^\infty$ adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^\infty$ with $\boldsymbol{x}_t \in \mathcal{X}$ for each $t \in \mathbb{Z}_+$. The sequence $\{\alpha_t\}_{t=0}^\infty$ determines the distance between successive terms. We define $\boldsymbol{c}_t := (\boldsymbol{x}_t - \boldsymbol{x}_{t+1})/\alpha_t$ and thus

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \boldsymbol{c}_t\,. \tag{2}$$

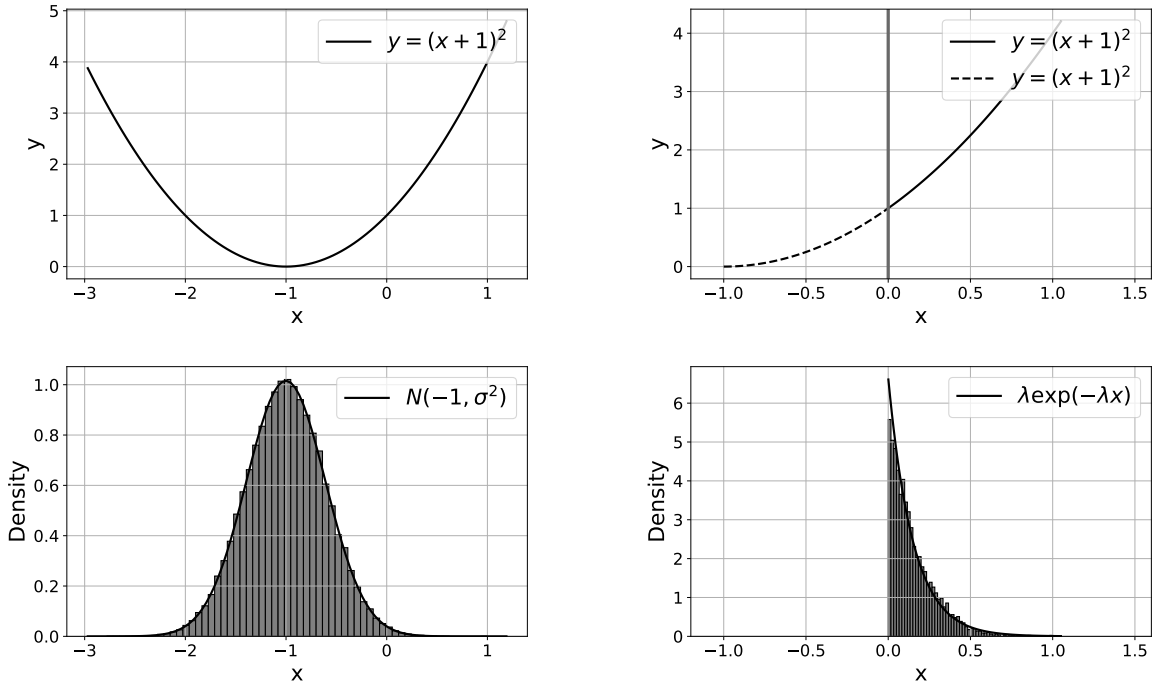
3. Main Results

In this section, we present our main results, as well as intuition and counter-examples.

3.1. Informal description of the main result

The normal distribution is typically associated with the dispersion of a random walk. However, when a random walk is constrained, the exponential distribution is the limiting distribution. So, while the Gaussian concentration is applied in the analysis of smooth stochastic approximation algorithms, exponential concentration we show occurs in non-smooth problems.

With reference to Figure 1, the high-level intuition for this behavior in a stochastic gradient algorithm is as follows. Consider a projected stochastic gradient descent algorithm with a small but fixed learning rate. When the optimum is in the interior of the constraint set and the objective is smooth, the algorithm's progress will slow as the iterates approach the minimizer in a manner that is roughly proportional to the distance to the optimum. In this regime, the process is well approximated by an Ornstein-Uhlenbeck (OU) process, for instance, see Chapter 10 of Kushner and Yin (2003). An OU process is known to have a normal distribution as its limiting stationary distribution. This stationary distribution determines the rate of convergence to the optimum; see Chen et al. (2022). If we consider the same iterates, but instead, these are now projected to belong to a constraint set, then the gradient of iterates need not approach zero as we approach the optimum.



(a) Unconstrained Stochastic Gradient Descent (b) Constrained Stochastic Gradient Descent

Figure 1 The above plots a simulation of a stochastic gradient descent algorithm with a constant step size on the function $f(x) = (x+1)^2$. Figure 1(a): When the objective is unconstrained the density of the location of iterates is well approximated by a normal distribution with variance $\sigma^2 = O(\alpha)$, where α is the step size of the algorithm. The distance to the optimum is $O(\alpha^{1/2})$. Figure 1(b): When the value of x is constrained to the positive orthant the gradient no longer vanishes. The distribution of iterates away from zero now has an exponential decay with rate $\lambda = O(\alpha^{-1})$. So for step size, α, the distance to the optimum is $O(\alpha)$. This paper proves that exponential concentration holds more generally for stochastic approximation procedures with non-vanishing gradients.

See Figure 1b). The resulting process behaves in a manner that is approximated by a reflected Brownian motion. When the gradient is non-zero on the boundary, it is well known that a reflected Brownian motion with negative drift has an exponential distribution as its stationary distribution, see Harrison and Williams (1987). We seek to establish bounds that exhibit this exponential, stationary behavior while allowing for time-dependent step sizes. This provides intuition for the exponential concentration results found in this paper.

To summarize, *strongly* convex functions are approximately quadratic around their optimum; this leads to the normal approximation. Think of such functions being U-shaped. However, for exponential concentration, the function is V-shaped at the optimum. Such functions are *sharp* convex functions. (See Remark 1 for a discussion on sharpness.) Here, the gradient does not approach

zero as we approach the optimum. We show that when a convex function is sharp, a stochastic gradient descent algorithm has an exponential concentration around its optimum, leading to much faster convergence.

3.2. An Exponential Lyapunov Bound

We consider a general stochastic optimization algorithm and show that exponential concentration holds when the expected progress towards its objective is proportional to the learning rate α_t. This leads to the following condition.

ASSUMPTION 1 **(Drift Condition)**. *The sequence* $\{\boldsymbol{x}_t\}_{t=0}^{\infty}$ *satisfies*

$$\mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)|\mathcal{F}_t] \leq -2\alpha_t\kappa \tag{C1}$$

whenever $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \geq \alpha_t B$ *for some* $\kappa > 0$ *and some* $B > 0$.

We also assume that noise is sub-exponential.

ASSUMPTION 2 **(Moment Condition)**. *There exists a constant* $\lambda > 0$ *and a random variable* Y *such that*

$$\big[|f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)|\big|\mathcal{F}_t\big] \leq \alpha_t Y \quad \textit{and} \quad \mathbb{E}[e^{\lambda Y}] < \infty\,. \tag{C2}$$

Condition (C1) states that the stochastic iterates will make progress against its objective when away from the optimum. The Condition (C2) is a mild noise condition. For example, if $f(X)$ is Lipschitz continuous, then it is sufficient that $||\boldsymbol{c}_t||$ has a sub-exponential tail. (See Lemma EC.1 in E-companion EC.2.1.1 for verification of this claim.) Shortly we will establish the Conditions (C1) and (C2) when applying projected stochastic gradient descent. However, for now, we leave (C1) and (C2) as general conditions that can be satisfied by a stochastic approximation algorithm.

The main result of this section is as follows.

THEOREM 1. *For learning rates of the form* $\alpha_t = a/(u+t)^\gamma$ *with* $a, u > 0$ *and* $\gamma \in [0, 1]$, *if Conditions* (C1) *and* (C2) *are satisfied by a stochastic approximation algorithm, then*

$$\mathbb{P}(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star) \geq z) \leq Ie^{-\frac{J}{\alpha_t}z} \tag{3}$$

and

$$\mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star)] \leq K\alpha_t \tag{4}$$

for time independent constants I, J *and* K.

These results show that once the dependence on the system's initial state has diminished, the process $f(\boldsymbol{x}_t)$ has an exponential concentration and will be within a factor of α_t of the optimum. We focus on learning rates of the form $\alpha_t = a/(u+t)^\gamma$; however, a result for more general learning rates is proved in Proposition 1 in Section 4.1.

It is worth remarking that Theorem 1 (and Proposition 1) hold for any algorithm for which the generic Conditions (C1) and (C2) hold. The results are not intended to apply to any particular stochastic optimization, nor do we place specific design restrictions on the algorithm. The result emphasizes that a convergence rate may differ depending on the geometry of the problem at hand, and this convergence may well be faster than anticipated.

3.3. Projected Stochastic Gradient Descent

In Theorem 1, we did not specify the stochastic approximation procedure used nor did we explore settings where Conditions (C1) and (C2) hold. This section provides a standard setting where our results apply. We consider projected stochastic gradient descent on the Lipschitz continuous function $l : X \to \mathbb{R}$. That is we wish to solve the optimization problem:

$$\text{minimize} \qquad l(\boldsymbol{x}) \qquad \text{over} \quad \boldsymbol{x} \in \mathcal{X}\,. \tag{5}$$

We analyze the Project Stochastic Gradient Descent (PSGD) algorithm:

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \alpha_t \boldsymbol{c}_t \tag{6a}$$

$$\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{y}_{t+1}) \tag{6b}$$

where $\mathbb{E}\big[\boldsymbol{c}_t \big| \mathcal{F}_t\big] = \nabla l(\boldsymbol{x}_t)$ and $\alpha_t = a/(u+t)^\gamma$ for $a > 0$, $u \in \mathbb{R}_+$, $\gamma \in [0,1]$. Above $\nabla l(\boldsymbol{x})$ can be either the gradient or a sub-gradient of l. We let $\mathcal{X}^\star = \arg\min_{\boldsymbol{x}\in\mathcal{X}} l(\boldsymbol{x})$ be the set of optimizers of (5).

Previously, we required Conditions (C1) and (C2), which jointly placed assumptions on the iterates and objective. Now that we have specified the iterative procedure, we can decouple to give conditions that only depend on the properties of the objective function.

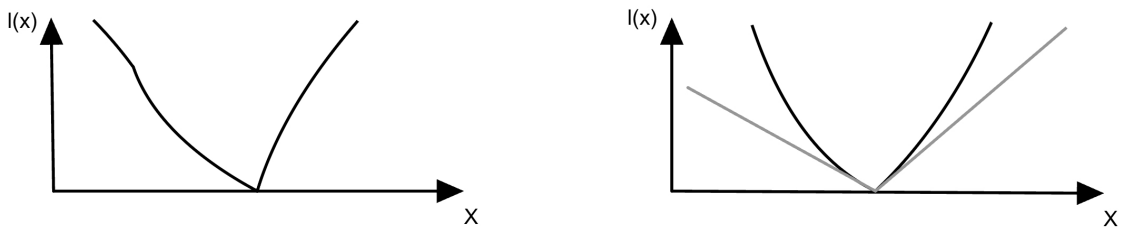
Assumption 3 **(Gradient Condition)**. *There exists a positive constant $\kappa > 0$ such that for all $\boldsymbol{x} \in \mathcal{X}$*

$$\nabla l(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{x}^\star) \geq \kappa \|\boldsymbol{x} - \boldsymbol{x}^\star\|, \tag{D1}$$

where $\boldsymbol{x}^\star = \Pi_{\mathcal{X}^\star}(\boldsymbol{x})$.

Assumption 4 **(Sub-Exponential Noise)**. *There exists a $\lambda > 0$ such that*

$$\sup_{t\in\mathbb{N}} \mathbb{E}[e^{\lambda||\boldsymbol{c}_t||} | \mathcal{F}_t] < \infty\,. \tag{D2}$$



(a) Gradient Condition (D1). (b) Convexity and Sharpness (D1′)

Figure 2 Under the gradient condition (D1), the objective need not be convex nor continuously differentiable. We require the derivative in the direction of the optimum to be non-zero. Under convexity and sharpness (D1′), the envelope of the function is bounded below by a cone. Here condition (D1) is satisfied.

Conditions (D1) and (D2) replace Conditions (C1) and (C2). Let's interpret these new conditions. Firstly, (D1) states that the (unit) directional derivative in the direction from $\boldsymbol{x}$ to $\boldsymbol{x}^\star$ is bounded above by $-\kappa$. I.e. we require strictly negative slope in the direction of the minimum. Note that that this does not require convexity of our objective functionl. (See Figure 2.) Condition (D2) assumes that the tail behavior of the gradient estimates is sub-exponential. (See Lemma EC.1.)

We can prove the following result that holds as a consequence of Theorem 1.

THEOREM 2. *If Condition* (D1) *and* (D2) *hold and* $\alpha_t = a/(u+t)^\gamma$ *for* $a, u > 0$ *then PSGD satisfies*

$$\mathbb{P}\left(\min_{\boldsymbol{x}^\star \in \mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \geq z\right) \leq J e^{-\frac{I}{\alpha_t} z}, \quad \mathbb{E}\left[\min_{\boldsymbol{x}^\star \in \mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|\right] \leq K\alpha_t, \quad \mathbb{E}\left[l(\boldsymbol{x}_{t+1}) - \min_{\boldsymbol{x} \in \mathcal{X}} l(\boldsymbol{x})\right] \leq L\alpha_t$$

where above J, I, K, L *are positive constants.*

For Stochastic Gradient Descent (SGD), the expected distance from $\boldsymbol{x}_t$ to the set of optima $\mathcal{X}^\star$ is known to converge at rate $\Omega(1/\sqrt{t})$. The convergence rate found in Theorem 2 is faster than that typically assumed for SGD. Notice this improved convergence is not due to any change in the algorithm but due to the geometry of the problem. (If we happen to know the problem geometry in advance, we can adjust the algorithm to significantly improve performance; see Section 3.6 and Davis et al. (2019).) The above result holds because of tighter exponential concentration occurs around the optimum in such cases. While the Gaussian concentration around the optimum for smooth convex objectives has been known for around 70 years (Chung (1954), Fabian (1968)), the exponential concentration found here does not appear in prior work on PSGD.

REMARK 1 (CONVEXITY AND SHARPNESS.). Sharpness is a condition which, stated informally, ensures that an objected function has a V-shape around its optimum. This is considered in the paper of Davis et al. (2019). A close relationship exists between our gradient condition (D1) and the sharpness condition. In particular, the two conditions are equivalent for convex optimization problems. (See Figure 2.)

A function $l(\boldsymbol{x})$ is sharp if for all $x \in \mathcal{X}$

$$l(\boldsymbol{x}) - \min_{\boldsymbol{x}^\star \in \mathcal{X}} l(\boldsymbol{x}^\star) \geq \kappa' \min_{\boldsymbol{x}^\star \in \mathcal{X}^\star} \|\boldsymbol{x} - \boldsymbol{x}^\star\|. \tag{D1$'$}$$

The lemma below proves that the gradient condition (D1) implies sharpness and, for convex functions, the two properties are equivalent:

LEMMA 1. *If the function $l(\boldsymbol{x})$ is absolutely continuous then the gradient condition* (D1) *implies the function is sharp,* (D1$'$)*. Moreover, if the function $l(\boldsymbol{x})$ is convex, then the gradient condition* (D1) *is equivalent to the function being sharp* (D1$'$)*.*

We prove Lemma 1 in Section EC.2.1.2 of the E-companion. The immediate consequence of this lemma is that Theorem 2 holds for sharp convex functions. So, there is a tighter exponential concentration for sharp convex objectives when compared with the Gaussian concentration bounds found for smooth convex objectives.

REMARK 2 (SMOOTH FUNCTIONAL CONSTRAINTS). Suppose the optimization (5) takes the form

$$\text{minimize} \quad l(\boldsymbol{x}) \quad \text{subject to} \quad l_i(\boldsymbol{x}) \leq 0, \quad i = 1, ..., m \quad \text{over} \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{7}$$

where $l(\boldsymbol{x})$ and $l_i(\boldsymbol{x})$ are smooth convex functions defining the bounded constraint set $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d : l_i(\boldsymbol{x}) \leq 0, i = 1, ..., m\}$. It is argued that a stochastic approximation algorithm obeys a central limit theorem if there are m_0 active constraints at the optimum with $m_0 < d$. See Kushner and Clark (1978b), Shapiro (1989), Duchi and Ruan (2021), Davis et al. (2023). However, if $m_0 \geq d$, the normal approximation degenerates. In this case, our results can be applied. With the following lemma and Theorem 2, we see that PSGD obeys an exponential concentration bound rather than a normal approximation.

LEMMA 2. *Suppose that at the optimum $-\nabla l(\boldsymbol{x}^\star) \in$ relint $\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star)$, where $\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star) := \{\boldsymbol{v} : \boldsymbol{v}^\top(\boldsymbol{x}^\star - \boldsymbol{y}) \leq 0, \forall \boldsymbol{y} \in \mathcal{X}\}$ and $\nabla l(\boldsymbol{x}^\star) \neq 0$ and that there are at least d active constraints at $\boldsymbol{x}^\star$ (w.l.o.g. $i = 1, ..., d$) and*

$$\{\nabla l_i(\boldsymbol{x}^\star) : i = 1, ..., d\} \text{ are linearly independent} \tag{D1$''$}$$

then the function f is sharp at $\boldsymbol{x}^\star$ and Assumption (D1) *holds.*

A proof of Lemma 2 is given in Section EC.2.1.3 of the E-companion. The premise of the above lemma is taken from Assumption B from Duchi and Ruan (2021). However, rather than a Gaussian approximation as found in Duchi and Ruan (2021) for $m_0 < d$ constraints, a consequence of the above Lemma is that exponential concentration will hold for PSGD if there are $m_0 \geq d$ linearly independent, active constraints at the optimum. Because of the degeneracy indicated in prior

works, the Gaussian approximation is not asymptotically optimal when (D1) holds, i.e. the normal vectors to the set of active constraints has full rank. Essentially, there is insufficient smoothness at the optimum for a central limit theorem to hold. Large deviation effects will likely determine the asymptotically optimal concentration at the optimum. When the Gaussian approximation fails, the theory of asymptotic optimality for stochastic optimization with constraints appears to be open. See Section 6 for further discussion.

Remark 3 (Projection). Although projection is a common requirement for stochastic gradient descent, the projection step (6b) can present computational overhead, so we discuss that here.

There are settings such as sharp objective functions where the optimum belongs to the interior of the constraint set. In this case, a finite number of projections are required. (A proof is given in Proposition EC.1 in the E-companion.)

Some constraints exhibit low complexity projection. It is common to select a set $\mathcal{X}$ that allows simple projection e.g. a box or disk containing $\mathcal{X}^\star$. For convex constraints $\{\boldsymbol{z} : \boldsymbol{g}_j(\boldsymbol{z}) \leq 0, j = 1, .., d\}$, the dual of a constraint set is $\mathbb{R}^d_+ = \{\boldsymbol{x} : \boldsymbol{x} \geq 0\}$ and thus the dual has simple projection. Low complexity projections exist for single constraint problems such as projection onto the probability simplex (Michelot (1986), Duchi et al. (2008)). Chapter 7 of Hazan et al. (2016) gives several examples of fast projection available with conditional gradient algorithms. We will analyze stochastic conditional gradient algorithms shortly. See also Bertsekas (2015) for further examples of low complexity projection.

There are practical general projection algorithms. The cyclic projection algorithm of Bregman (1967) can be used for the intersection of a finite number of convex sets. Here, Bregman also proposes other non-Euclidaen distances that can be used to simplify projection. Mandel (1984) proves Bregman's algorithm converges linearly for polytope constraints. A number of linear convergent and parallelizable projection algorithms are given in Censor and Zenios (1997).

Projection is a standard requirement in the analysis of SGD. However, projection is not a requirement of our general result Theorem 1; instead, we require iterates to be bounded. For instance, we will apply our results to Stochastic Frank Wolfe as a non-projective alternative to PSGD shortly. In general we find the boundedness of $\mathcal{X}$ can be removed when learning rate is constant, $\alpha_t \equiv \alpha$. (See Lemma (EC.6) in the E-companion statement and proof.) From this we see that our linear convergence results apply Stochastic Gradient Descent (without projection). So neither projection nor bounded iterates are required for linear convergence.

3.4. Kiefer-Wolfowitz Algorithm: Exponential Concentration

So far, we have applied the concentration bound in Theorem 1 to Projected Stochastic Gradient Descent (PSGD). In this section, we consider the Kiefer-Wolfowitz algorithm. The Kiefer-Wolfowitz

is a well-known alternative to Robbins and Monro's Stochastic Gradient Descent algorithm. Here, gradient estimates are replaced by noisy finite difference operators.

We consider the Kiefer-Wolfowitz algorithm under the analogous sharpness in noise conditions we thought for a PSGD. Typically, the Kiefer-Wolfowitz algorithm has a worse rate of convergence than PSGD. However, under the non-vanishing gradient condition, we show that the Kiefer-Wolfowitz algorithm with an appropriate finite difference estimator will have the same concentration and asymptotic convergence rate as PSGD.

Consider the optimization:

$$\text{minimize} \qquad l(\boldsymbol{x}) := \mathbb{E}_{\hat{w}}[l(\boldsymbol{x}, \hat{w})] \qquad \text{over} \quad x \in \mathcal{X}\,, \tag{8}$$

where $\hat{w}$ is a random variable. For $\nu \in \mathbb{R}$, we define the vector $\boldsymbol{l}(\boldsymbol{x}+\boldsymbol{\nu}, w) := (l(\boldsymbol{x}+\nu\boldsymbol{e}_i, w) : i = 1, ..., d)$ where $\boldsymbol{e}_i$ is the ith unit vector. The Kiefer–Wolfowitz (KW) algorithm is as follows:

$$\boldsymbol{c}_t = \frac{\boldsymbol{l}(\boldsymbol{x}_t + \boldsymbol{\nu}, \hat{w}_t^+) - \boldsymbol{l}(\boldsymbol{x}_t - \boldsymbol{\nu}, \hat{w}_t^-)}{2\nu} \tag{9a}$$

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \alpha_t \boldsymbol{c}_t \tag{9b}$$

$$\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{y}_{t+1}) \tag{9c}$$

where $\alpha_t = \frac{a}{(u+t)^\gamma}$ for $a > 0$, $u \in \mathbb{R}_+$, $\gamma \in [0, 1]$. Above $\hat{w}_t^+, \hat{w}_t^-$ are IIDRVs equal in distribution to $\hat{w}$. Notice the main change from the PSGD algorithm is that $\boldsymbol{c}_t$ is no longer an unbiased estimator of $\nabla l(\boldsymbol{x}_t)$. However, for sufficiently well-behaved functions, there exists a constant c such that for all $x \in \mathcal{X}$

$$\left\| \nabla l(\boldsymbol{x}) - \frac{\boldsymbol{l}(\boldsymbol{x}+\boldsymbol{\nu}) - \boldsymbol{l}(\boldsymbol{x}-\boldsymbol{\nu})}{2\nu} \right\| \leq c d^{\frac{1}{2}} \nu^2\,. \tag{D3}$$

(Above we include d to emphasize the dependence on the dimension of $\mathcal{X}$.)

We further assume that the random variable $l(\boldsymbol{x}, \hat{w})$ has a uniformly bounded variance over $\boldsymbol{x} \in \mathcal{X}$. This is a standard assumption for the analysis of the KW algorithm. See Fabian (1967). We will assume Conditions (D1) and (D2) hold, with $\boldsymbol{c}_t$ as defined in (9a). Ordinarily, convergence of the KW algorithm requires ν to decrease with time. However, we see that this is not necessary for sharp functions. The parameter ν needs to be below a certain threshold, and once satisfied, exponential concentration results hold.

THEOREM 3. *If Conditions* (D1), (D2), (D3) *hold and if*

$$\nu \leq \left(\frac{\kappa}{3cd^{\frac{1}{2}}} \right)^{\frac{1}{2}}$$

then the Kiefer-Wolfowitz algorithm satisfies

$$\mathbb{P}\left(\min_{\boldsymbol{x} \in \mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\| \geq z \right) \leq \hat{J} e^{-\frac{\hat{I}}{\alpha_t} z}, \quad \mathbb{E}\left[\min_{\boldsymbol{x} \in \mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\| \right] \leq \hat{K} \alpha_t, \quad \mathbb{E}\left[l(\boldsymbol{x}_{t+1}) - \min_{\boldsymbol{x} \in \mathcal{X}} l(\boldsymbol{x}) \right] \leq \hat{L} \alpha_t$$

where above $\hat{J}, \hat{I}, \hat{K}, \hat{L}$ are positive constants.

The proof of Theorem 3 is given in Section EC.2.3 of the E-companion.

If we take $\alpha_t = 1/t$, then we see that the Kiefer-Wolfowitz algorithm has a convergence rate $\mathbb{E}\|x_t - x^\star\| = O(1/t)$. This is faster than the $O(1/t^{1/3})$ rate, which is typically found to be optimal for the KW algorithm. Again, typically, KW iterations follow a normal approximation. For instance, see Ruppert (1982). However, again, an exponential concentration is more appropriate than a normal approximation if there is non-vanishing drift. Interestingly, a constant finite difference approximation can be used to obtain results and the rate is now the same as PSGD.

3.5. Stochastic Frank-Wolfe: a non-Projective Algorithm

We now investigate exponential concentration in stochastic algorithms that do not require projection. Non-vanishing negative drift is the main requirement for exponential concentration, not projection or boundary effects. We emphasize this by proving the exponential concentration for a projection-free algorithm, specifically, the Stochastic Frank-Wolfe algorithm. Here we require further assumptions in addition to sharpness conditions.

The Frank-Wolfe algorithm or Conjugate Gradient algorithm, as it is sometimes called, is proposed as the standard "projection-free" alternative to projected gradient descent algorithms, see Jaggi (2013) and Hazan and Kale (2012). We form an analysis of the Stochastic Frank-Wolfe (SFW) algorithm, as described by Hazan and Luo (2016). This is a standard implementation of Frank-Wolfe with a sample estimate of the gradient. As with the PSGD and KW algorithms, we choose a standard stochastic optimization algorithm. Our aim is not algorithm design but instead to emphasize that exponential (rather than normal) concentration can naturally occur in stochastic approximation depending on the geometry of the problem.

Consider the same setting from Section 3.3. That is we wish to solve the optimization

$$\text{minimize} \qquad l(\boldsymbol{x}) \qquad \text{over} \quad x \in \mathcal{X}\,. \tag{10}$$

For a sequence of positive integers $(m_t \in \mathbb{N} : t \in \mathbb{N})$ and a sequence $(\alpha_t \in (0,1) : t \in \mathbb{N})$, the Stochastic Frank-Wolfe algorithm is defined as follows

$$\boldsymbol{c}_t = \sum_{i=1}^{m_t} \frac{\boldsymbol{c}_t^i}{m_t} \tag{11a}$$

$$\boldsymbol{v}_t \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \boldsymbol{c}_t^\top \boldsymbol{x} \tag{11b}$$

$$\boldsymbol{x}_{t+1} = (1-\alpha_t)\boldsymbol{x}_t + \alpha_t \boldsymbol{v}_t\,. \tag{11c}$$

Above, where $\boldsymbol{c}_t^i$ are random variables, independent (after conditioning on $\boldsymbol{x}_t$) with a uniformly bounded variance such that $\mathbb{E}[\boldsymbol{c}_t^i|\mathcal{F}_t] = \nabla l(\boldsymbol{x}_t)$. We continue to assume that $\mathcal{X}$ is a closed bounded set. In addition to condition (D1) and (D2), we add the following conditions.

ASSUMPTION 5 **(Smooth convex square-error)**. *For the error* $\epsilon(\boldsymbol{x}) := l(\boldsymbol{x}) - l(\boldsymbol{x}^\star)$,

$$\epsilon(\boldsymbol{x})^2 \text{ is a smooth convex function.} \tag{E1}$$

ASSUMPTION 6 **(Interior Optimum)**. *The set of optima belongs to the interior of the set* $\mathcal{X}$. *That is*

$$\mathcal{X}^\star \subset \mathcal{X}^\circ. \tag{E2}$$

We briefly discuss these two conditions. Assumption (E1) is a non-standard cone condition. Analogous to a smoothness convex function having quadratic behavior at the optimum, Condition (E1) requires that the objective function has a behavior that behaves like the distance function to the optimum. We illustrate this with Lemma EC.3 in the E-companion. Here, we show that Condition (E1) is satisfied if we take $l(x)$ to be the distance to the desired set of optimal points:

$$d_{\mathcal{X}^\star}(\boldsymbol{x}) = \min_{x^\star \in \mathcal{X}^\star} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_S .$$

Here S is a positive semi-definite matrix.

Condition (E2) requires that the optimum is in the interior. The rate of convergence found on the boundary is typically slower see Proposition EC.2 in the E-companion. When analyzing projected stochastic gradient descent, one case we discussed is when the optimum is on the boundary of the constraint set. Interestingly, the case of the Stochastic Frank-Wolfe algorithm is different: the algorithm will converge faster when the optimum is *not* on the boundary. (If it is known in advance that the optimum is not in the interior we would recommend running PSGD due to its faster convergence, see Section EC.2.2 in the E-companion.) Nonetheless, Frank-Wolfe is a non-projective optimization algorithm, the results of this section emphasize that exponential concentration is not the property of projection on a specific boundary type but is really about non-vanishing drift at the optimum.

The main result of this subsection is given below. It gives sufficient conditions for exponential concentration for the Frank-Wolfe algorithm.

THEOREM 4. *For learning rates of the form* $\alpha_t = a/(u+t)^\gamma$ *with* $a, u > 0$ *and* $\gamma \in [0,1]$, *if Conditions* (D1), (D2), (E1) *and* (E2) *hold and if* $m_t \geq (3\sigma/\kappa\alpha_t)^2$ *then the stochastic Frank-Wolfe algorithm satisfies*

$$\mathbb{P}\left(l(\boldsymbol{x}_{t+1}) - \min_{\boldsymbol{x}\in\mathcal{X}} l(\boldsymbol{x}) \geq z\right) \leq Ie^{-\frac{J}{\alpha_t}z},$$

for constants I,J.

The proof of Theorem 4 can be found in Section EC.2.4 of the E-companion.

3.6. Linear Convergence under Exponential Concentration

We have seen that, without adjusting the algorithm, the convergence of the stochastic approximation procedure improves under exponential concentration. However, if we know exponential concentration holds, then we can further adjust the algorithm to give even faster convergence.

The linear convergence of Projected Stochastic Gradient Descent (PSGD) on convex objectives is established in Theorem 3.2 of Davis et al. (2019). This result relies on a normal approximation concentration result (Davis and Drusvyatskiy 2019, Theorem 4.1), which is not as tight as the exponential concentration bound in Theorem 2 above. Thus, below in Theorem 5, we provide an improvement to Theorem 3.2 of Davis et al. (2019). We give a general version of the linear convergence that can be proven under the conditions of Theorem 1. The result does not require the set $\mathcal{X}$ to be bounded. From this, extensions of these linear convergence results hold for the Projected Stochastic Gradient Descent, the Kiefer-Wolfowitz algorithm, and the stochastic Frank-Wolfe algorithm. Here we present the result for PSGD. The futher for the Kiefer-Wolfowitz algorithm, and the stochastic Frank-Wolfe algorithm results are stated in Section EC.2.6 of the E-companion.

We wish to solve the optimization problem (5). We consider a Stochastic Approximation algorithm (2) implemented over several stages, $s = 1, ..., S$. We let t_s be the number of iterations in the sth stage. The idea is that within each stage s, the learning rate $\hat{\alpha}_s$ is fixed and is chosen so that the error of the stochastic approximation algorithm should be halved by the end of each stage. Specifically, we let $\hat{\boldsymbol{x}}_s$ be the state at the end of stage s. We define $T_s = \sum_{s'=1}^{s} t_{s'}$ and

$$\hat{\boldsymbol{x}}_s = \boldsymbol{x}_{T_s}, \qquad \text{and} \qquad \alpha_t = \hat{\alpha}_s, \qquad \text{for} \quad T_{s-1} \le t < T_s, \text{ and } s = 1, ..., S\,. \tag{12}$$

The following theorem gives choices for $\hat{\alpha}_s$ and t_s to ensure a linear rate of convergence.

THEOREM 5. *We assume that $\mathcal{X}$ is a convex set that may be unbounded. Assume Conditions* (C1) *and* (C2) *hold for a stochastic approximation procedure with rates given in* (12)*:*
a) If, for $\hat{\epsilon} > 0$ and $\hat{\delta} \in (0,1)$, we set

$$S = \log\left(\frac{F}{\hat{\epsilon}}\right), \quad \hat{\alpha}_s = \frac{2^{-s} F \kappa}{E \log\left(\frac{RS}{\hat{\delta}}\right)}, \quad \textit{and} \quad t_s = \left\lceil \frac{2}{\kappa^2} \log\left(\frac{RS}{\hat{\delta}}\right) \right\rceil$$

then with probability greater than $1 - \hat{\delta}$ it holds that $\min_{\boldsymbol{x} \in \mathcal{X}^\star} \|\hat{\boldsymbol{x}}_S - \boldsymbol{x}\| \le \hat{\epsilon}$. Moreover, the number of iterations (12) *required to achieve this bound is*

$$\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right) \right\rceil \left\lceil \frac{2}{\kappa^2} \log\left(\frac{R}{\hat{\delta}}\right) + \log\left(\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right) \right\rceil\right) \right\rceil .$$

(Above $F = \min_{x^\star \in \mathcal{X}^\star} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|$ and R and E are time-independent constants that depend on the constants given in Conditions (C1) *and* (C2)*.)*

b) For $\hat{\alpha}_s = \frac{a}{2^s \log(s+1)}$ *and* $t_s = \log^2(s+1)$*, there exists positive constants* A *and* M *such that* $\forall \hat{\delta} \in (0,1)$ *if* $a \geq A/\hat{\delta}$ *then*

$$\mathbb{P}\left(\min_{\boldsymbol{x} \in \mathcal{X}^\star} \|\hat{\boldsymbol{x}}_s - \boldsymbol{x}\| \leq 2^{-s} M, \quad \forall s \in \mathbb{N}\right) \geq 1 - \hat{\delta}\,.$$

The proof of Theorem 5 is given in Section EC.2.6 of the E-companion. We apply an improved exponential concentration bound and make some adjustments; however, other than that, the argument largely follows that of (Davis et al. 2019, Theorem 3.2.), so we refer the reader to Davis et al. (2019) also.

For PSGD, (Davis et al. 2019, Theorem 3.2) proves a sample complexity bound of order $O(\hat{\delta}^{-2}[\log \hat{\epsilon}^{-1}]^3)$. The bound in Theorem 5a) above improves this and has an order $O(\log \hat{\epsilon}^{-1}\big[\log\log \hat{\epsilon}^{-1} + \log \hat{\delta}^{-1}\big])$. The above bound is the same order as the best bound found in Davis et al. (2019), namely Theorem 3.8, which holds for an ensemble method consisting of three adaptively regularized gradient descent algorithms. The sample complexity is improved for the PSGD case because the exponential concentration bound is tighter that Gaussian bound Theorem 3.2 Davis et al. (2019). Other than this our proof follows the ideas laid out in Davis et al. (2019). We find that similar results hold for Kiefer-Wolfowitz, and Frank-Wolfe.

Unlike Theorem 2, Theorem 5a) suggests that we require a refined understanding to calibrate parameters to improve convergence. The implementation of Theorem 5b) only requires one parameter, a, which needs to be chosen sufficiently large. (For instance, experiments could increase the parameter a until convergence is observed.) So, although there is some cost to the algorithm's complexity, we do not require detailed knowledge of the problem at hand to implement a geometrically convergent algorithm. Further, Theorem 5b) holds for a stronger mode of convergence, in that the geometric convergence holds for all time with arbitrarily high probability.

We note that the above bound applies when the function f and constraint sets $\mathcal{X}$ are unbounded. This is because we apply Lemma 4 under constant step sizes. For this result, the bounded constraint assumption is not required.

4. Proofs

This section proves our main result Theorem 1. We then apply this to Projection Stochastic Gradient Descent to prove Theorem 2. The proofs of the remaining results are contained in the E-companion.

4.1. Proof of Theorem 1

Several constants are introduced in the proof of Theorem 1. For later reference, these are listed in Section EC.3.1 of the E-companion.

The proof of Theorem 1 relies on Proposition 1, which is a somewhat more general yet more abstract version of Theorem 1. This result establishes that once the sum $\sum_{s=1}^{t} \alpha_s$ is sufficiently large, the error of stochastic iterations is of the order of α_t. Much like Theorem 1 the key observation is that $\alpha_t^{-1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star))$ has an exponential concentration rather than the normal concentration of $\alpha_t^{-1/2}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star))$.

For the stochastic iterates described in Section 2, we will assume that $\{\alpha_t\}_{t\geq 0}$ is a deterministic non-increasing sequence such that

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \qquad \liminf_{t\to\infty} \frac{\alpha_{2t}}{\alpha_t} > 0 \qquad \text{and} \qquad \lim_{t\to\infty} \frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = 0\,. \tag{13}$$

We do not need to assume that $\alpha_t \to 0$. Later we consider small but constant step sizes. This condition is satisfied by any sequence of the form $\alpha_t = a/(u+t)^\gamma$ for $a, u > 0$ and $\gamma \in [0,1]$.

Proposition 1. *When Conditions* (C1) *and* (C2) *are satisfied, there exist positive constants* E, G, H, T_0 *independent of* t *such that*

$$\mathbb{P}(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star) \geq z) \leq e^{-\frac{\kappa G^n}{2E\alpha_t}(z - \alpha_t B - F - \alpha_0 B + \sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2})} + He^{-\frac{\kappa G^n}{2E\alpha_t}(z - \alpha_t B)}\,. \tag{14}$$

for any n *with* $t/2^n > T_0$. *Further, for any* t *such that* $\sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \kappa \geq 2(F + \alpha_0 B)$ *there exists a constant* C *such that*

$$\mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star)] \leq C\alpha_t\,. \tag{15}$$

4.1.1. Proof of Proposition 1. Here, we briefly outline the proof of Proposition 1. The proof uses Lemma 3, Lemma 4, Lemma 5, and Proposition 2, which are stated below. Lemma 3, although not critical to our analysis, simplifies the drift Condition (C1) by eliminating some terms and boundary effects. Lemma 4, on the other hand, is an important component of our proof. It converts the drift Condition (C1) into an exponential bound, which we then iteratively expand. The lemma extends Theorem 2.3 from Hajek (1982) by allowing for adaptive time-dependent step sizes. Proposition 2 applies standard moment generating function inequalities to the results found in Lemma 4. Lemma 5 is a technical lemma used in the proof of Proposition 2. After Proposition 2 is proven, the proof of Proposition 1 follows.

We now proceed with the steps outlined above. We let

$$L_t := f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) - \alpha_t B \tag{16}$$

where α_t satisfies (13). First, we simplify the above Conditions (C1) and (C2) to give the Lyapunov conditions (17) and (18) stated below. The following is a technical lemma.

LEMMA 3. *Given Conditions* (C1) *and* (C2) *hold, there exists a deterministic constant* T_0 *such that the sequence of random variables* $(L_t : t \geq T_0)$ *satisfies*

$$\mathbb{E}\big[L_{t+1} - L_t \big| \mathcal{F}_t\big]\mathbb{I}[L_t \geq 0] < -\alpha_t \kappa, \tag{17}$$

and

$$[|L_{t+1} - L_t||\mathcal{F}_t] \leq \alpha_t Z \quad \textit{where} \quad D := \mathbb{E}[e^{\lambda Z}] < \infty\,. \tag{18}$$

The proof can be found in the E-companion, Section EC.3.2.

Given Lemma 3, we will now assume (17) and (18) hold in place of (C1) and (C2). We will convert the drift condition (17) into an exponential bound and then iterate to give the bound below.

LEMMA 4. *For any* t *and* $\hat{t}$ *with* $t \geq \hat{t} \geq T_0$ *and for any* $\eta > 0$ *such that* $\alpha_{\hat{t}}\eta \leq \lambda$ *then*

$$\mathbb{E}[e^{\eta L_{t+1}}|\mathcal{F}_{\hat{t}}] \leq \mathbb{E}[e^{\eta L_{T_1}}|\mathcal{F}_{\hat{t}}]\prod_{k=\hat{t}}^{t}\rho_t + D\sum_{\tau=\hat{t}+1}^{t+1}\prod_{k=\tau}^{t}\rho_k\,,$$

where $\rho_t = e^{-\alpha_t\eta\kappa+\alpha_t^2\eta^2 E}$, *and* $E := \mathbb{E}\left[(e^{\lambda Z} - 1 - \lambda Z)/\lambda^2\right] < \infty$.

Proof of Lemma 4. Let $Z_t = (L_{t+1} - L_t)/\alpha_t$. From (18), we have $[|Z_t||\mathcal{F}_t] \leq Z$ where $\mathbb{E}[e^{\lambda Z}] < \infty$. From (17), we have $\mathbb{E}[Z_t|\mathcal{F}_t] \leq -\kappa$ on the event $\{L_t \geq 0\}$. Thus, on the event $\{L_t \geq 0\}$ the following holds:

$$\begin{aligned}
\mathbb{E}[e^{\eta(L_{t+1}-L_t)}|\mathcal{F}_t] = \mathbb{E}[e^{\alpha_t\eta Z_t}|\mathcal{F}_t] &= 1 + \alpha_t\eta\mathbb{E}[Z_t|\mathcal{F}_t] + \alpha_t^2\eta^2\mathbb{E}\left[\frac{e^{\alpha_t\eta Z_t} - 1 - \alpha_t\eta Z_t}{\alpha_t^2\eta^2}\Big|\mathcal{F}_t\right] \\
&\leq 1 + \alpha_t\eta\mathbb{E}[Z_t|\mathcal{F}_t] + \alpha_t^2\eta^2\sum_{k=2}^{\infty}\frac{1}{k!}\mathbb{E}[|Z_t|^k|\mathcal{F}_t]\eta^{k-2}\alpha_t^{k-2} \\
&\leq 1 - \alpha_t\eta\kappa + \alpha_t^2\eta^2\sum_{k=2}^{\infty}\frac{1}{k!}\mathbb{E}[Z^k]\lambda^{k-2} \\
&= 1 - \alpha_t\eta\kappa + \alpha_t^2\eta^2\mathbb{E}\left[\frac{e^{\lambda Z} - 1 - \lambda Z}{\lambda^2}\right] \qquad (19) \\
&= 1 - \alpha_t\eta\kappa + \alpha_t^2\eta^2 E \\
&\leq e^{-\alpha_t\eta\kappa+\alpha_t^2\eta^2 E} =: \rho_t\,. \qquad (20)
\end{aligned}$$

We apply a Taylor expansion and the (conditional) Monotone Convergence Theorem in the first inequality above, see (Williams 1991, 9.7e)). In the second inequality, we apply (17) and (18) above, and also recall that α_t is decreasing. In the final inequality, we applied the standard bound $1 + x \leq e^x$. We note that ρ_t as define above satisfies $\rho_t < 1$ whenever $\alpha_t < \kappa/\eta E$. We note that E is finite since by assumption $\mathbb{E}[e^{\lambda Z}] < \infty$. Also from the expansion given in (19) (which holds by the Monotone Convergence Theorem), it is clear that E is positive.

The bound (20) holds on the event $\{L_t \geq 0\}$. Now notice

$$\begin{aligned}\mathbb{E}[e^{\eta L_{t+1}}|\mathcal{F}_t] &= \mathbb{E}[e^{\eta(L_{t+1}-L_t)}|\mathcal{F}_t]e^{\eta L_t}\mathbb{I}[L_t \geq 0] + \mathbb{E}[e^{\eta(L_{t+1}-L_t)}|\mathcal{F}_t]e^{\eta L_t}\mathbb{I}[L_t < 0] \\ &\leq \rho_t e^{\eta L_t}\mathbb{I}[L_t \geq 0] + \mathbb{E}[e^{\eta\alpha_t Z}]e^{\eta L_t}\mathbb{I}[L_t < 0] \\ &\leq \rho_t e^{\eta L_t}\mathbb{I}[L_t \geq 0] + D\mathbb{I}[L_t < 0] \\ &\leq \rho_t e^{\eta L_t} + D\,.\end{aligned}$$

The first inequality applies the above bound (20) and the second inequality applies the boundedness condition (18). Taking expectations above gives

$$\mathbb{E}[e^{\eta L_{t+1}}|\mathcal{F}_{\hat{t}}] \leq \rho_t \mathbb{E}[e^{\eta L_t}|\mathcal{F}_{\hat{t}}] + D.$$

By induction, we have

$$\mathbb{E}[e^{\eta L_{t+1}}|\mathcal{F}_{\hat{t}}] \leq \mathbb{E}[e^{\eta L_{T_1}}|\mathcal{F}_{\hat{t}}]\prod_{k=\hat{t}}^{t}\rho_t + D\sum_{\tau=\hat{t}+1}^{t+1}\prod_{k=\tau}^{t}\rho_k\,,$$

as required. □

Note that the above lemma does not require the set of values $\mathcal{X}$ to be bounded (or convex). This is a point that we will later utilize in the proof of Theorem 5. The following is a technical lemma.

LEMMA 5. *If α_t, $t \in \mathbb{Z}_+$, is a decreasing positive sequence, then*

$$\min_{s=\hat{t},\ldots,t}\left\{\frac{\sum_{k=s}^{t}\alpha_k}{\sum_{k=s}^{t}\alpha_k^2}\right\} = \frac{\sum_{k=\hat{t}}^{t}\alpha_k}{\sum_{k=\hat{t}}^{t}\alpha_k^2}\,. \tag{21}$$

Moreover, if α_t, $t \in \mathbb{Z}_+$ satisfies the learning rate condition (13) *then*

$$\frac{1}{\alpha_{\lfloor t/2^n \rfloor}} \geq \frac{G^n}{\alpha_t} \qquad \textit{and} \qquad \min_{s=\lfloor t/2^n \rfloor,\ldots,t}\left\{\frac{\sum_{k=s}^{t}\alpha_k}{\sum_{k=s}^{t}\alpha_k^2}\right\} \geq \frac{G^n}{\alpha_t} \tag{22}$$

for some constant $G \in (0,1]$ and for $n \in \mathbb{N}$ such that $t/2^n > 1$.

A proof is given in Section EC.3.2 of the E-companion. Looking ahead to the proof of Theorem 1, for step sizes of the form $\alpha_t = a/(u+t)^\gamma$, we have $G = 1/4^\gamma$ and we will take $n = 1$ for $\gamma < 1$. For $\gamma = 1$, we need to have to be more careful choosing n, which will be a constant depending on a, u, B and F.

With the moment generating function bound in Lemma 4 and the bound in Lemma 5, we can bound the tail probabilities and expectation of L_t.

PROPOSITION 2. *For any sequence satisfying* (13)*, there exists a constants H and Q such that*

$$\mathbb{P}(L_{t+1} \geq z) \leq e^{-\frac{QG^n}{\alpha_t}(z - F - \alpha_0 B + \sum_{s=\lfloor t/2^n \rfloor}^{t}\alpha_s\frac{\kappa}{2})} + He^{-\frac{QG^n}{\alpha_t}z} \tag{23}$$

for $z \geq 0$ and for $n \in \mathbb{N}$ such that $t/2^n > T_0$. Further, for t is such that $\sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2} \geq F + \alpha_0 B$, then

$$\mathbb{E}[L_{t+1}] \leq \frac{(1+H)}{QG^n}\alpha_t \,.$$

Proof of Proposition 2. By Lemma 5, we see that

$$\frac{\lambda}{\alpha_{\lfloor t/2^n \rfloor}} \geq \frac{\lambda G^n}{\alpha_t} \qquad \text{and} \qquad \min_{\lfloor t/2^n \rfloor \leq s \leq t} \left\{ \frac{\sum_{k=s}^{t} \alpha_k \kappa}{2\sum_{k=s}^{t} \alpha_k^2 E} \right\} \geq \frac{\kappa G^n}{2\alpha_t E}$$

for a constant $G > 0$. So that η lower bounds the above two expressions, we take

$$\eta = Q\frac{G^n}{\alpha_t} \qquad \text{where} \quad Q = \lambda \wedge (\kappa/2E)\,.$$

We apply Lemma 4 which gives

$$\begin{aligned}
\mathbb{P}(L_{t+1} \geq z) &\leq e^{-\eta z}\mathbb{E}[e^{\eta L_{t+1}}] \\
&\leq e^{-\eta z}\mathbb{E}[e^{\eta L_{\lfloor t/2^n \rfloor}}] \prod_{k=\lfloor t/2^n \rfloor}^{t} \rho_t + e^{-\eta z} D \sum_{\tau=\lfloor t/2^n \rfloor+1}^{t+1} \prod_{k=\tau}^{t} \rho_k \\
&= e^{-\eta z}\mathbb{E}[e^{\eta L_{\lfloor t/2^n \rfloor}}] e^{\sum_{k=\lfloor t/2^n \rfloor}^{t} -\alpha_k \eta\kappa + \alpha_k^2 \eta^2 E} + e^{-\eta z} D \sum_{\tau=\lfloor t/2^n \rfloor+1}^{t+1} e^{\sum_{k=\tau}^{t} -\alpha_t \eta\kappa + \alpha_t^2 \eta^2 E}\,. \qquad (24)
\end{aligned}$$

Notice, for η as defined above, it holds that

$$\sum_{k=\tau}^{t} -\alpha_k \eta\kappa + \alpha_k^2 \eta^2 E \leq -\frac{1}{2}\sum_{k=\tau}^{t} \alpha_k \eta\kappa \leq -\frac{1}{2}(t-\tau)\alpha_t \eta\kappa, \qquad \forall \tau = \lfloor t/2^n \rfloor, ..., t\,.$$

Applying this to (24) gives

$$\begin{aligned}
\mathbb{P}(L_{t+1} \geq z) &\leq e^{-\eta z}\mathbb{E}[e^{\eta L_{\lfloor t/2^n \rfloor}}] e^{\sum_{k=\lfloor t/2^n \rfloor}^{t} -\alpha_k \eta \frac{\kappa}{2}} + e^{-\eta z} D \sum_{\tau=\lfloor t/2^n \rfloor+1}^{t+1} e^{-(t-\tau)\alpha_t \eta \frac{\kappa}{2}} \\
&\leq e^{-\eta z}\mathbb{E}[e^{\eta L_{\lfloor t/2^n \rfloor}}] e^{\sum_{k=\lfloor t/2^n \rfloor}^{t} -\alpha_k \eta \frac{\kappa}{2}} + e^{-\eta z} D \frac{e^{\alpha_t \eta \frac{\kappa}{2}}}{1 - e^{-\alpha_t \eta \frac{\kappa}{2}}} \qquad (25)
\end{aligned}$$

In the 1st inequality above we note that $\alpha_k \geq \alpha_t$ for all $k \leq t$. In the 2nd inequality, we note that the summation over τ are terms from a geometric series, so we upper bound this by the appropriate infinite sum.

Thus, the bound (25) becomes

$$\mathbb{P}(L_{t+1} \geq z) \leq \mathbb{E}\Big[e^{\frac{QG^n L_{\lfloor t/2^n \rfloor}}{\alpha_t}}\Big] e^{-\frac{QG^n}{\alpha_t}(z + \sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2})} + e^{-\frac{QG^n}{\alpha_t} z} D \frac{e^{\frac{\kappa QG^n}{2}}}{1 - e^{-\frac{\kappa QG^n}{2}}}\,.$$

Noting that $L_{\lfloor t/2^n \rfloor} \leq \max_{x \in \mathcal{X}} f(x) - \min_{x \in \mathcal{X}} f(x) + \alpha_0 B = F + \alpha_0 B$, by the definition of F. We simplify the above expression as follows

$$\begin{aligned}
\mathbb{P}(L_{t+1} \geq z) &\leq e^{\frac{QG^n}{\alpha_t}[F + \alpha_0 B - z - \sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2}]} + e^{-\frac{QG^n}{\alpha_t} z} D \frac{e^{\frac{\kappa QG^n}{2}}}{1 - e^{-\frac{\kappa QG^n}{2}}} \\
&\leq e^{-\frac{QG^n}{\alpha_t}(z + \sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2} - F - \alpha_0 B)} + H e^{-\frac{QG^n}{\alpha_t} z}\,. \qquad (26)
\end{aligned}$$

Above we define $H := De^{\frac{\kappa QG^n}{2}}/(1-e^{-\frac{\kappa QG^n}{2}})$. This gives (23).

Notice if t is such that $F+\alpha_0 B - \sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2} \leq 0$, then the above inequality (26) can be bounded by

$$\mathbb{P}(L_{t+1} \geq z) \leq (1+H)\, e^{-\frac{QG^n}{\alpha_t} z}.$$

Thus

$$\mathbb{E}[L_{t+1}] \leq \mathbb{E}[L_{t+1} \vee 0] = \int_0^\infty \mathbb{P}(L_{t+1} \geq z) dz \leq (1+H) \int_0^\infty e^{-\frac{QG^n}{\alpha_t} z}\, dz = (1+H)\, \frac{\alpha_t}{QG^n}\ ,$$

as required. □

With Proposition 2 in place we can prove Proposition 1.

Proof of Proposition 1. From Proposition 2

$$\mathbb{P}(L_{t+1} \geq z') \leq e^{-\frac{QG^n}{\alpha_t}(z' - F - \alpha_0 B + \sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2})} + He^{-\frac{QG^n}{\alpha_t} z'}$$

for $z' \geq 0$ where $f(\boldsymbol{x}_{t+1}) = L_{t+1} + \alpha_{t+1} B + f(\boldsymbol{x}^\star)$. Taking $z' = z - \alpha_{t+1} B$, gives

$$\begin{aligned}\mathbb{P}(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star) \geq z) &= \mathbb{P}(L_{t+1} \geq z - \alpha_{t+1} B) \\ &\leq e^{-\frac{QG^n}{\alpha_t}(z - \alpha_t B - F - \alpha_0 B + \sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2})} + He^{-\frac{QG^n}{\alpha_t}(z - \alpha_t B)}\,,\end{aligned}$$

which gives (14) as required. Also by Proposition 2, we thus taking $C = [(1+H)/2QG^n + B]$, the required bound (15) holds. □

4.1.2. Proof of Theorem 1. We can now prove Theorem 1.

Proof of Theorem 1. We notice that the bound $\sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2} \geq \alpha_0 B + F$ can be achieved for all $t \geq T_1$ for fixed constants T_1 and n. This holds since $\sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \to \infty$ as $t \to \infty$. (See Lemma (EC.7) in the E-companion for verification of this and a concrete choice of T_1 and n.) Thus applying bound (14) from Proposition 1 with $T_2 = \max\{T_0, T_1\}$, we see that

$$\begin{aligned}\mathbb{P}(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star) \geq z) &\leq e^{-\frac{QG^n}{\alpha_t}(z - \alpha_t B - F - \alpha_0 B + \sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \frac{\kappa}{2})} + He^{-\frac{QG^n}{\alpha_t}(z - \alpha_t B)} \\ &\leq (1+H) e^{-\frac{QG^n}{\alpha_t}(z - \alpha_t B)} && \text{for } t \geq T_2 \\ &\leq Ie^{-\frac{J}{\alpha_t} z} && \text{for } t \geq 0\,.\end{aligned}$$

Thus we see that (3) holds for $t \geq 0$ with suitable choice of I and J (e.g. $I = (1+H)e^{QG/(F/\alpha_{T_2} - B)}$ and $J = QG^n$). Integrating the bound (3) then gives

$$\mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star)] \leq \int_0^\infty Ie^{-(J/\alpha_t)z} dz = \frac{I}{J}\alpha_t\,.$$

□

4.2. Proof of Theorem 2

Typical Stochastic Gradient Descent proofs use $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2$ as a Lyapunov function. However, we want to use $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$ instead. We start with the standard SGD drift argument and then take a square root to gain our drift condition for $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$. We can then apply Theorem 1, which goes through the mechanics of converting a linear Lyapunov drift condition into an exponential Lyapunov function. The idea of converting linear drift into an exponential Lyapunov function is reasonably well-known in the Markov chains analysis but currently not for SGD. Our proof shows how to adapt and apply these ideas. The basic mechanics to apply Theorem 1 are the same as for other SA procedures, for example, Kiefer-Wolfowitz and Stochastic Frank-Wolfe.

Proof of Theorem 2. In this proof, we will apply Proposition 1 with the choice $f(\boldsymbol{x}) := \min_{\boldsymbol{x}^\star \in \mathcal{X}^\star} \|\boldsymbol{x} - \boldsymbol{x}^\star\|$. We also define $\boldsymbol{x}_t^\star := \arg\min_{\boldsymbol{x} \in \mathcal{X}^\star} \|\boldsymbol{x}_t - \boldsymbol{x}\|$. Now observe that

$$\begin{aligned} f(\boldsymbol{x}_{t+1})^2 = \left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\right\|^2 \le \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t^\star\|^2 &= \|\Pi_{\mathcal{X}}(\boldsymbol{x}_t - \alpha_t \boldsymbol{c}_t) - \Pi_{\mathcal{X}}(\boldsymbol{x}_t^\star)\|^2 \\ &\le \|\boldsymbol{x}_t - \alpha_t \boldsymbol{c}_t - \boldsymbol{x}_t^\star\|^2 = \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|^2 - 2\alpha_t \boldsymbol{c}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_t^\star) + \alpha_t^2 \|\boldsymbol{c}_t\|^2 . \quad (27) \end{aligned}$$

Condition (D2) implies all moments of $\|\boldsymbol{c}_t\|$ are uniformly bounded. In particular, suppose σ is such that $\mathbb{E}[\|\boldsymbol{c}_t\|^2 | \mathcal{F}_t] < \sigma^2$ for all t. On the event where $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \ge \alpha_t \frac{\sigma^2}{\kappa} > 0$ then (27) gives

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\le \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| \sqrt{1 - 2\alpha_t \boldsymbol{c}_t^\top \frac{(\boldsymbol{x}_t - \boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|^2} + \alpha_t^2 \frac{\|\boldsymbol{c}_t\|^2}{\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|^2}} \\ &\le \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| \left(1 - \alpha_t \boldsymbol{c}_t^\top \frac{(\boldsymbol{x}_t - \boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|^2} + \frac{\alpha_t^2}{2} \frac{\|\boldsymbol{c}_t\|^2}{\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|^2}\right) . \end{aligned}$$

Above the first inequality follows from (27) and in the second inequality we note that $\sqrt{1+x} \le 1 + \frac{x}{2}$. Taking expectations on both sides shows that, on the event $\{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \ge \alpha_t \frac{\sigma^2}{\kappa}\}$, it holds that

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\right\| | \mathcal{F}_t\right] \le \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| - \alpha_t \frac{\nabla l(\boldsymbol{x}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|} + \frac{\alpha_t^2}{2} \frac{\mathbb{E}[\|\boldsymbol{c}_t\|^2 | \mathcal{F}_t]}{\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|} \le \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| - \alpha_t \kappa + \alpha_t \frac{\kappa}{2}$$

Or in other words $\mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) | \mathcal{F}_t] \le -\alpha_t \frac{\kappa}{2}$ whenever $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \ge \alpha_t \frac{\sigma^2}{\kappa}$. Thus we see that Condition (C1) holds.

We now verify Condition (C2). Projections reduced distances, specifically, if $\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| \le \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\|$ then

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) = \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\| - \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| \le \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t^\star\| - \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| \le \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = \alpha_t \|\boldsymbol{c}_t\|$$

(The analogous argument follows if $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\| \le \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|$.) As discussed in Section 3.3, the MGF condition on $\|\boldsymbol{c}_t\|$ now implies (C2). (See Lemma EC.1 for a proof.)

We can now apply Theorem 1 which gives:

$$\mathbb{P}\Big(\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}\|\geq z\Big)\leq \hat{I}e^{-\frac{\hat{J}}{\alpha_t}z}\quad\text{and}\quad\mathbb{E}\left[\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}\|\right]\leq\hat{K}\alpha_t$$

for constants $\hat{I}$, $\hat{J}$ and $\hat{K}$. Since we also assume in addition that $l:\mathcal{X}\to\mathbb{R}$ is Lispchitz continuous (with Lipschitz constant $\hat{L}/\hat{K}$) we have, as required,

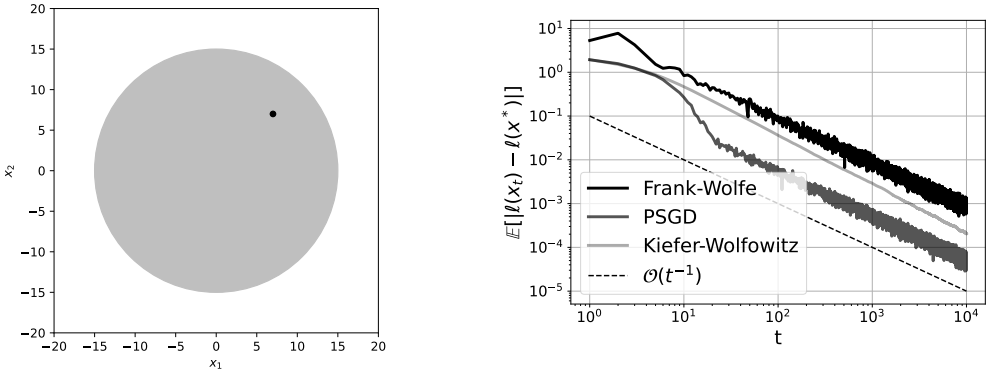
$$\mathbb{E}\left[l(\boldsymbol{x}_{t+1})-\min_{\boldsymbol{x}\in\mathcal{X}}l(\boldsymbol{x})\right]\leq\frac{\hat{L}}{\hat{K}}\mathbb{E}\left[\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}\|\right]\leq\hat{L}\alpha_t\,.$$

□

5. Applications and Numerical Examples

We present several numerical results on the phenomena proven above. (In addition to supplementary files, associated code can be accessed on GitHub; see Yang (2025).)

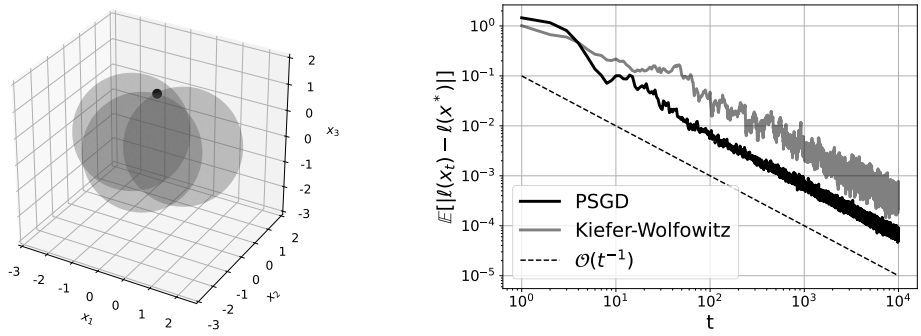
A Circle Constraint. We consider minimizing an objective function $l(x)=||x-x^*||$ for $x^*=(7,7)$ over a large circle with center $(0,0)$ and radius 15. The Frank-Wolfe, PSGD and Kiefer-Wolfowitz algorithms are applied. For the Kiefer-Wolfowitz, we assume the objective function is observed with noise following $N(0,0.01)$. The optimum is in the interior. No projection is required in our simulation. Figure 3 shows the constraint set and the rate $O(1/t)$ with $\gamma=1$ for the three algorithms.



(a) Circle Constraint (b) Convergence of Algorithms

Figure 3 Convergence of Frank-Wolfe, PSGD, and Kiefer-Wolfowitz algorithms on the circle constraint example. Figure 3(a): the black dot is the optimal solution $(7,7)$. Figure 3(b): The expectation is computed over 20 realizations. The stochastic gradients for Frank-Wolfe and PSGD are computed with batch size $B=10$. The parameter $v=0.8$ is chosen for Kiefer-Wolfowitz. The parameters of step size are chosen as $a=0.9, u=1$ and $\gamma=1$ such that $\alpha_t=1/(1+t)$. The fitted slope is -1.00, -1.00 and -1.10 for Frank-Wolfe, PSGD and Kiefer-Wolfowitz.

Three Spherical Constraints. Davis et al. (2023) consider a normal approximation on PSGD with two spherical constraints. We add a third spherical constraint. Specifically, we minimize the objective $l(x) = -x_1 + \hat{w}_i x_i$, $\hat{w}_i \sim N(0,1)$ for $i = 1,2,3$ over the intersection of spheres with center $(1,0,0), (-1,0,0)$ and $(0,1,0)$, and radius 2 using PSGD and Kiefer-Wolfowitz algorithm. See Figure 4(a) for the constraint set. The optimal solution is taken at $(0,0,\sqrt{3})$. Bregman's cyclic algorithm is applied for projection. Rather than $O(1/\sqrt{t})$, Figure 4(b) shows the rate $O(1/t)$ with $\gamma = 1$ for both the PSGD and the Kiefer-Wolfowitz algorithm. Further, we study PSGD and Kiefer-Wolfowitz convergence when halving the step size every $T = 20$ iterations. As suggested in Section 3.6, linear convergence occurs. See Figure 5.



(a) Three Spherical Constraints (b) Convergence of Algorithms

Figure 4 Convergence of PSGD and Kiefer-Wolfowitz algorithms on the three spherical constraints problems. Figure 4(a): the black dot is the optimal solution $(0,0,\sqrt{3})$. Figure 4(b): The expectation is computed over 20 realizations. The stochastic gradients for PSGD are computed with $B = 10$. The parameter $v = 1$ is chosen for Kiefer-Wolfowitz. The parameters of step size are chosen as $a = 1, u = 1$ and $\gamma = 1$ such that $\alpha_t = 1/(1+t)$. The fitted slope is -1.01 and -1.00 for PSGD and Kiefer-Wolfowitz.

Non-Negative Ridge Regression. Duchi and Ruan (2021) consider the normal approximation with constraints for a non-negative least square problem and for ridge regression. We consider non-negative ridge regression and find that the normal approximation is no longer valid. We apply $l(x) = \frac{1}{2}|a^T x - b|^2$ as the objective with a constraint set $\{x \in \mathbb{R}^2_+ : \|x\| \leq \sqrt{0.9}\}$, where a and b are observed with $b_i \sim a_i x_+ + \xi_i$ for $x_+ = (1,-1)$, $a_i \sim N(0, \mathbb{I}_2)$ and $\xi_i \sim N(0,1)$. The optimal solution is taken at $(\sqrt{0.9}, 0)$. Figure 6 shows the rate $O(1/t)$ with $\gamma = 1$ for the PSGD and Kiefer-Wolfowitz.

Linear Programs and Markov Decision Processes. All the theoretical results and simulations so far consider non-linear constraints and objectives. We limit all discussion of linear programs to this paragraph and the E-companion. First, all linear programs are sharp (see Lemma EC.8 in the E-companion for proof). Second, linearly convergent (and parallelizable) algorithms can calculate

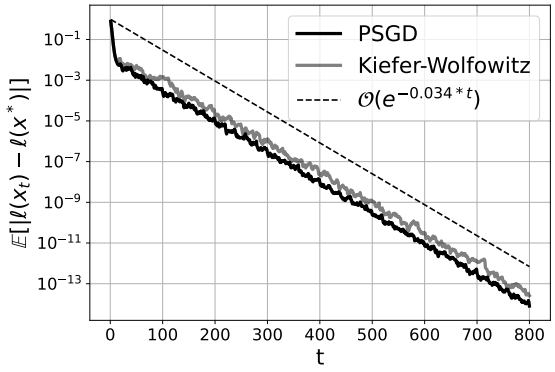


Figure 5 Linear convergence of PSGD and Kiefer-Wolfowitz for the three spherical constraints problem. The expectation is computed over 20 realizations. The stochastic gradients are computed with $B = 10$. The parameter $v = 10$ is chosen for Kiefer-Wolfowitz. The simulations are conducted with learning rates divided by 2 every 20 steps.

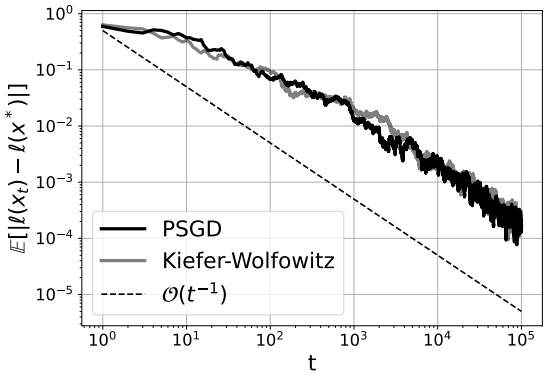
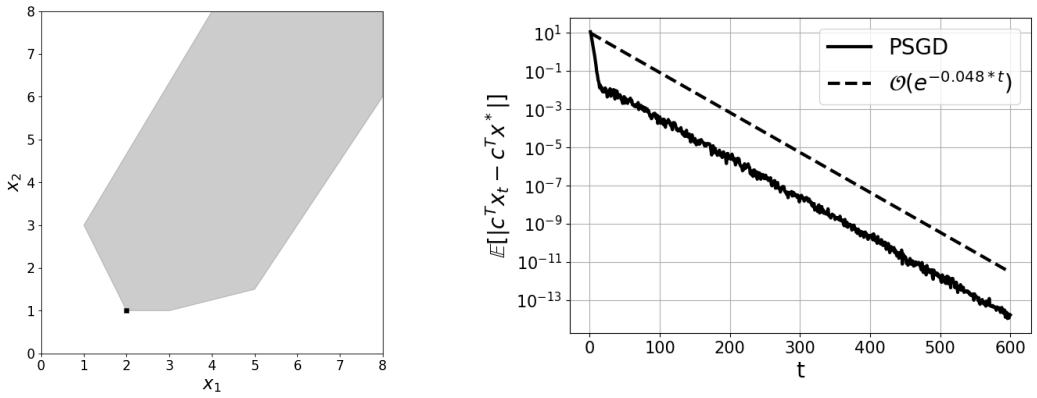


Figure 6 Non-negative least square example. The expectation is computation over 20 realizations. The parameter $v = 1$ is chosen for Kiefer-Wolfowitz. The parameters of step size are chosen as $a = 1$, $u = 1$ and $\gamma = 1$ such that $\alpha_t = 1/(1+t)$. The fitted slope is -1.00 and -1.00 for PSGD and Kiefer-Wolfowitz.

projections onto linear constraints. Hildreths's Projection algorithm is a first-order algorithm that converges linearity, see Iusem and De Pierro (1990). Further, Dos Santos (1987) shows the algorithm can be parallelized and adapted to non-linear constraints. Proved in the results of Section 3.6, Figure 7 demonstrates linear convergence of PSGD on a linear program with unbounded constraints. Markov Decision Processes can be expressed as linear programs, with PSGD on the dual corresponding to a simple policy gradient algorithm. Given this, we solve a general three-state,

two-action MDP, and Blackjack (See, Sutton and Barto (2018)). More detailed discussions can be found in Section EC.4 of the E-companion.



(a) Linear Constraints

(b) Linear convergence of PSGD

Figure 7 Linear convergence of PSGD. Costs are normally distributed with mean $(4, 6)$ and covariance $(25, 0; 0, 25)$. The simulation is conducted with learning rates divided by 1.1 every 2 steps.

6. Further Discussions

We have established exponential concentration in stochastic optimization algorithms. Via examples, counter-examples, and heuristics, we present several different directions for future exploration.

Firstly, we discuss convergence in distribution when exponential concentration occurs. Similar to the classical Gaussian approximation, a natural question arises: is the exponential distribution the limit family distributions? From a counter-example, we argue that there is no simple parametric family characterizing limit behaviour. Second, we discuss the impact of exponential concentration when combined with the Gaussian approximation. Again, through an example, we show that both the exponential concentration and Gaussian limit distributions impact the value of the objective function in a stochastic gradient descent algorithm. Third, we discuss lower bounds that impact the convergence rate of sample average approximation under sharpness. Since the Fisher-Information might not fully describe the optimality of sharp objectives, the aim is to establish some characteristics of asymptotic optimality for constrained stochastic optimization.

The full resolution of these issues is certainly beyond the scope of the present work. However, these do represent promising research directions arising from non-Gaussian behavior in stochastic approximation and offer directions for further advancements in the field.

6.1. The Exponential Approximation

The limit distribution of stochastic approximation is a normal distribution when the shape of the objective around the optimum is approximately quadratic, e.g. like $\|\boldsymbol{x}\|^2$. When the curvature

behaves as $\|\boldsymbol{x}\|$, our analysis proves exponential tail behavior. A natural question is whether the limiting distribution of iterates away from the optimum is exponentially distributed under the constant drift condition? That is do we have convergence in distribution:

$$\frac{x_t - x^\star}{\alpha_t} \xrightarrow[t\to\infty]{\mathcal{D}} X,$$

where X is exponentially distributed? The short answer is no. In general, the limit distribution is *not* exponential. For a simple counter-example, consider projected stochastic gradient descent where $\mathcal{X} = \mathbb{R}_+$ and c_t are i.i.d. random variables with $c_t = -1$ with probability p and $c_t = 1$ with probability $1-p$. If we fix α, we can already see that an exponential distribution limit is not possible since the process x_t/α belongs to the set $\mathbb{Z}_+$. If $p > 1/2$, then the limit distribution is geometrically distributed, not exponential. For general distributions of c_t, the limit distribution is given by an integral equation. See Lindley's Integral Equation (Asmussen 2003, Corollary 6.6). However, the resulting distributions all exhibit exponential tail bounds. (See Kingman's Bound c.f. Kingman (1964))

Convergence in distribution likely holds. However, the limit X is unlikely to be an exponential distribution, and it is unlikely to have a simple form. We cannot aggregate fluctuations in the same manner as found in the normal approximation. We cannot expect a simple statistic like the Fisher Information to determine the directions of statistical error because the stochastic approximation process is much more concentrated. A theory of asymptotic optimality is likely to be characterized in terms of exponents rather than distributions. Sharpness is a natural condition for a convex function in much the same way as smoothness is. However, different techniques are required here because errors and stepsizes are of the same order of magnitude, so exponential tail bounds are not seen in prior literature on stochastic approximation. However, as this article shows, we can understand convergence behavior by constructing these exponential concentration bounds.

6.2. Exponential and Gaussian Bounds

Exponential concentration occurs for locally linear objectives, which are, thus, informally stated, V-shaped, whereas Gaussian concentration occurs in locally quadratic objectives, or U-shaped. We briefly discuss, by example, what we expect to happen when the objective is both V- and U-shaped. We then discuss issues that might occur in a more general theory.

Consider the optimization

$$\min \quad l(x,y) := x^2 + y \quad \text{over} \quad x \in \mathbb{R} \quad y \in \mathbb{R}_+.$$

Suppose that we apply stochastic gradient descent and that the noise for both the x and y components is Gaussian with mean 0 and variance 1. This applies to the following update:

$$x_{t+1} = x_t - \alpha(2x_t + G_t^x)\,, \qquad y_{t+1} = \max\left\{y_t - \alpha(1 + G_t^y), 0\right\}\,.$$

These can be seen as the Euler-Maruyama approximations to the following stochastic differential equations:

$$\mathrm{d}X(t) = -2X(t)\mathrm{d}t + \sqrt{\alpha}\mathrm{d}B^x(t)\,, \qquad \mathrm{d}Y(t) = -\mathrm{d}t + \sqrt{\alpha}\mathrm{d}B^y(t) + \mathrm{d}L(t)\,.$$

Here $B^x(t)$ and $B^y(t)$ are independent standard Brownian motions and $L(t)$ is the local-time of the process $Y(t)$ at zero. See the text of Stroock and Varadhan (1997) for a proof of convergence for the standard diffusion case and see Stroock and Varadhan (1971) for the case with reflection. Notice that the $X(t)$ above is an Ornstein-Uhlenbeck process, and $Y(t)$ is a reflected Brownian motion. Also, X and Y are independent. The stationary distributions are Gaussian $X \sim \mathcal{N}(0,\alpha)$ and exponentially distributed $Y \sim \exp(2/\alpha)$, respectively. Thus, if we consider the convergence of these stationary distributions, then

$$\alpha^{-1/2}(X_\alpha, Y_\alpha) \xrightarrow[\alpha\to 0]{\mathcal{D}} (X, 0) \qquad \text{where} \qquad X \sim \mathcal{N}(0,1)\,. \tag{28}$$

This result is consistent with prior results on the asymptotic optimality of stochastic approximation. However, if we examine the limit of our objective function, we begin to see significant differences. Notice

$$\alpha^{-1} l(X,Y) \xrightarrow[\alpha\to 0]{\mathcal{D}} X^2 + Y \qquad \text{where} \qquad X^2 \sim \chi^2(1), \quad \text{and} \quad Y \sim \exp(2)\,. \tag{29}$$

Notice that from (28), we see the distance to the optimum not affected by the exponential concentration in Y. However, we see in (29) that there is an impact on the objective function from both Gaussian and Exponential terms. So, while an SGD algorithm can have an asymptotically optimal distance to the optimum, it may be that the performance concerning the optimization objective is not optimal due to the impacts of boundary constraints.

So, both the normal approximation and exponential concentration provide insight into the performance of PSGD algorithms. It should be noted. However, it appears we can't only consider normal and exponential tail behavior. There are attributes reached from a stochastic approximation algorithm that are neither normal nor exponential. For instance, a variety of stationary dynamics can be reached for objectives of the form $|x|^\alpha$ or when there are multiple optima Harrison and Reiman (1981). So, a complete characterization of the limiting distribution set for constrained stochastic gradient descent is undoubtedly challenging and will require appropriate assumptions for a general theory. However, we can conclude from this discussion that the behavior of SGD on smooth convex objective function with smooth convex constraints is meaningfully impacted by non-Gaussian asymptotic phenomena.

6.3. Asymptotically optimal rates of convergence

There is a well-developed theory of asymptotic optimality under the normal approximation. For example, the inverse Fisher Information gives the best form of asymptotic variance that can be achieved by a statistical procedure. In the case of exponential concentration, it is reasonable to also consider the best concentration rate. Here, we briefly indicate the form of exponential concentration under the best possible policy and provide characteristics of a theoretical result. A complete theory of asymptotic optimality under sharpness is beyond the scope of the present work.

Consider the following optimization and its sample average approximation:

$$\boldsymbol{x}^\star \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}[l(\boldsymbol{x};\xi)]\,, \qquad \text{and} \qquad \hat{\boldsymbol{x}}_t \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{t}\sum_{s=1}^{t} l(\boldsymbol{x};\xi_t)$$

where ξ is a random variable and ξ_t are independent identically distributed random variables. Also $\boldsymbol{x}^\star$ and $\hat{\boldsymbol{x}}_t$ respectively solve the optimizations

$$\boldsymbol{x}^\star \in \arg\min_{x\in\mathcal{X}} -\boldsymbol{x}^\top \mathbb{E}[\nabla l(\boldsymbol{x},\xi)]\,, \qquad \text{and} \qquad \hat{\boldsymbol{x}}_t \in \arg\min_{x\in\mathcal{X}} -\boldsymbol{x}^\top \frac{1}{t}\sum_{s=1}^{t} \nabla l(\boldsymbol{x},\xi_t)\,.$$

Given the above two optimizations, we ask what is the likelihood of a perturbation in the gradients $\hat{\nabla}_t := -\frac{1}{t}\sum_{s=1}^{t}\nabla l(\boldsymbol{x}^\star,\xi)$, which differs sufficiently from their mean of the random variable $\nabla^\star := -\nabla l(\boldsymbol{x}^\star,\xi)$, so that leads $\hat{\boldsymbol{x}}_t$ is not equal to $\boldsymbol{x}^\star$.

Recall from Lemma 2, we define $\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star) := \{\boldsymbol{v} : \boldsymbol{v}^\top(\boldsymbol{x}^\star - \boldsymbol{y}) \le 0\,, \forall \boldsymbol{y}\in\mathcal{X}\}$ and, given that our objective is sharp, we can assume $\mathbb{E}[\nabla^\star] \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star)^\circ$ and that $\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star)$ is closed. Then $\hat{\boldsymbol{x}}_t$ will be different from $\boldsymbol{x}^\star$ on the event $\{\hat{\nabla}_t \notin \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star)\}$. We can characterize the exponential concentration of this event, specifically, by Cramer's Theorem [See (Dembo and Zeitouni 2009, Theorem 2.2.30)]:

$$\liminf_{t\to\infty} \frac{1}{t}\log \mathbb{P}(\hat{\boldsymbol{x}}_t \neq \boldsymbol{x}^\star) \ge \liminf_{t\to\infty} \frac{1}{t}\log \mathbb{P}(\hat{\nabla}_t \notin \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star)) \ge -\inf_{\nabla\notin\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star)} D(\nabla;\nabla^\star)$$

where

$$D(\nabla;\nabla^\star) = \sup_{\phi\in\mathbb{R}^d}\left\{\phi^\top\nabla - \log(\mathbb{E}[e^{\phi^\top\nabla^\star}])\right\}.$$

It would seem that for sharp functions that sample average approximations cannot achieve an exponential concentration of rate larger than $\inf_{\nabla\notin\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^\star)} D(\nabla;\nabla^\star)$. A natural question to ask is if this provides a tight bound for asymptotic optimality for constrained stochastic approximation under sharpness.

7. Conclusions

Motivated by results on the exponential distributions found for queueing networks, we have established convergence rates for constrained stochastic approximation algorithms. Our results extend

the findings on Markov chains by Hajek (1982) to stochastic approximation. To the best of our knowledge, these techniques from the theory of Markov chains have not been applied in stochastic approximation.

Asymptotic normality is classical in stochastic approximation, whereas exponential concentration is poorly understood. This paper identifies situations where the asymptotically optimal solution is not Gaussian and provides methods to establish bounds when exponential concentration holds. Our results prove that faster convergence is a potential benefit of exponential concentration.

Acknowledgement: We thank the reviewers for their help in improving this paper. We are particularly grateful to a referee for the reference, Davis et al. (2019), which led to Theorem 5. Also, to Stavros Zenios for advice on projection algorithms and Matthias Troffaes for help with pycddlib. This research was supported by the EPSRC INFORMED-AI project EP/Y028732/1.

References

Asmussen S (2003) *Applied probability and queues*, volume 2 (Springer).

Bertsekas D (2015) *Convex optimization algorithms* (Belmont, MA: Athena Scientific).

Bertsimas D, Gamarnik D, Tsitsiklis JN (2001) Performance of multiclass markovian queueing networks via piecewise linear lyapunov functions. *Annals of Applied Probability* 1384–1428.

Birge JR, Louveaux F (2011) *Introduction to stochastic programming* (New York: Springer Science & Business Media).

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *Siam Review* 60(2):223–311.

Bregman LM (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics* 7(3):200–217.

Broadie M, Cicek D, Zeevi A (2011) General bounds and finite-time improvement for the kiefer-wolfowitz stochastic approximation algorithm. *Operations Research* 59(5):1211–1224.

Buche R, Kushner HJ (2002) Rate of convergence for constrained stochastic approximation algorithms. *SIAM journal on control and optimization* 40(4):1011–1041.

Censor Y, Zenios S (1997) *Parallel Optimization: Theory, Algorithms, and Applications*. Numerical mathematics and scientific computation (Oxford University Press), ISBN 9780195100624, URL `https://books.google.co.uk/books?id=ByG5zQBYC3UC`.

Chen Z, Mou S, Maguluri ST (2022) Stationary behavior of constant stepsize SGD type algorithms: An asymptotic characterization. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6(1):1–24.

Chung KL (1954) On a Stochastic Approximation Method. *The Annals of Mathematical Statistics* 25(3):463 – 483, URL http://dx.doi.org/10.1214/aoms/1177728716.

Davis D, Drusvyatskiy D (2019) Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization* 29(1):207–239, URL http://dx.doi.org/10.1137/18M1178244.

Davis D, Drusvyatskiy D, Charisopoulos V (2019) Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547* .

Davis D, Drusvyatskiy D, Jiang L (2023) Asymptotic normality and optimality in nonsmooth stochastic approximation. *arXiv preprint arXiv:2301.06632* .

Dembo A, Zeitouni O (2009) *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability (Springer Berlin Heidelberg), ISBN 9783642033117, URL https://books.google.co.uk/books?id=iT9JRlGPx5gC.

Dos Santos LT (1987) A parallel subgradient projections method for the convex feasibility problem. *Journal of Computational and Applied Mathematics* 18(3):307–320.

Duchi J, Shalev-Shwartz S, Singer Y, Chandra T (2008) Efficient projections onto the l 1-ball for learning in high dimensions. *Proceedings of the 25th international conference on Machine learning*, 272–279.

Duchi JC, Ruan F (2021) Asymptotic optimality in stochastic optimization. *The Annals of Statistics* 49:21–48.

Fabian V (1967) Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics* 191–200.

Fabian V (1968) On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics* 1327–1332.

Frank M, Wolfe P (1956) An algorithm for quadratic programming. *Naval research logistics quarterly* 3(1-2):95–110.

Hajek B (1982) Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability* 14(3):502–525.

Hájek J (1972) Local asymptotic minimax and admissibility in estimation. *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, 175–194.

Harrison JM, Reiman MI (1981) On the distribution of multidimensional reflected brownian motion. *SIAM Journal on Applied Mathematics* 41(2):345–361, ISSN 00361399, URL http://www.jstor.org/stable/2101076.

Harrison JM, Williams RJ (1987) Multidimensional reflected brownian motions having exponential stationary distributions. *The Annals of Probability* 115–137.

Hazan E, Kale S (2012) Projection-free online learning. *29th International Conference on Machine Learning, ICML 2012*, 521–528.

Hazan E, Luo H (2016) Variance-reduced and projection-free stochastic optimization. *International Conference on Machine Learning*, 1263–1271 (PMLR).

Hazan E, et al. (2016) Introduction to online convex optimization. *Foundations and Trends® in Optimization* 2(3-4):157–325.

Iusem AN, De Pierro AR (1990) On the convergence properties of hildreth's quadratic programming algorithm. *Mathematical programming* 47(1-3):37–51.

Jaggi M (2013) Revisiting frank-wolfe: Projection-free sparse convex optimization. *International conference on machine learning*, 427–435 (PMLR).

Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics* 462–466.

Kingman JFC (1964) A martingale inequality in the theory of queues. *Mathematical Proceedings of the Cambridge Philosophical Society* 60(2):359361, URL `http://dx.doi.org/10.1017/S0305004100037841`.

Kushner H, Clark D (1978a) *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Applied Mathematical Sciences (New York: Springer New York), ISBN 9781468493528.

Kushner H, Clark D (1978b) *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Applied Mathematical Sciences (Springer), ISBN 9783540903413, URL `https://books.google.co.uk/books?id=h4qmAAAAIAAJ`.

Kushner H, Yin G (2003) *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability (New York: Springer New York), ISBN 9780387008943.

Le Cam L (1953) On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *Univ. Calif. Publ. in Statist.* 1:277–330.

Mandel J (1984) Convergence of the cyclical relaxation method for linear inequalities. *Mathematical programming* 30:218–228.

Meyn SP, Tweedie RL (2012) *Markov chains and stochastic stability* (London: Springer Science & Business Media).

Michelot C (1986) A finite algorithm for finding the projection of a point onto the canonical simplex of $\setminus^n$. *Journal of Optimization Theory and Applications* 50:195–200.

Moulines E, Bach F (2011) Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems* 24.

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* 19(4):1574–1609.

Robbins H, Monro S (1951) A stochastic approximation method. *The annals of mathematical statistics* 400–407.

Ruppert D (1982) Almost sure approximations to the robbins-monro and kiefer-wolfowitz processes with dependent noise. *The Annals of Probability* 10(1):178–187.

Schweitzer PJ, Seidmann A (1985) Generalized polynomial approximations in markovian decision processes. *Journal of mathematical analysis and applications* 110(2):568–582.

Shapiro A (1989) Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics* 17(2):841–858.

Shapiro A, Dentcheva D, Ruszczynski A (2021) *Lectures on stochastic programming: modeling and theory* (Philadelphia, PA: SIAM).

Stroock D, Varadhan S (1971) Diffusion processes with boundary conditions. *Communications on Pure and Applied Mathematics* 24(2):147–225, ISSN 0010-3640, URL `http://dx.doi.org/10.1002/cpa.3160240206`.

Stroock DW, Varadhan SS (1997) *Multidimensional diffusion processes*, volume 233 (Springer Science & Business Media).

Sutton RS, Barto AG (2018) *Reinforcement learning: An introduction* (Cambridge, Massachusetts, London, England: MIT press).

Williams D (1991) *Probability with martingales* (Cambridge university press).

Yang S (2025) Code for exponential concentration in stochastic approximation. `https://github.com/Shangda-Yang/PSGD`, accessed: 2025-01-09.

E-companion

EC.1. Appendix to Introduction: Exponential Limit for Uniform MLE

For $X_1, ..., X_n \sim U[0,\theta]$, $\theta > 0$, the joint density at $(x_1, ...x_n)$ is

$$\prod_{i=1}^{n} \frac{1}{\theta} \mathbb{I}\Big[0 \le x_i \le \theta\Big] = \frac{1}{\theta^n} \mathbb{I}\Big[0 \le \min_{i=1,...,n} x_i \le \max_{i=1,...,n} x_i \le \theta\Big]$$

From the above, we see that the maximum likelihood is given by $\hat{\theta}_n = \max_{i=1,...,n} X_i$. For a normal approximation, we would typically analyse $\sqrt{n}(\hat{\theta}_n - \theta)$. However, instead we analyse a more concentrated asymptotic $n(\theta - \hat{\theta}_n)$. For this observe

$$\mathbb{P}(n(\theta - \hat{\theta}_n) \ge z) = \mathbb{P}\Big(\max_{i=1,...,n} X_i \le \theta - \frac{z}{n}\Big) = \Big(1 - \frac{z}{n\theta}\Big)^n \xrightarrow[n\to\infty]{} e^{-\frac{z}{\theta}}$$

From the above, we see that

$$n(\theta - \hat{\theta}_n) \xrightarrow[n\to\infty]{\mathcal{D}} \exp(\theta^{-1})$$

So the limit here is not normally distributed under a $\sqrt{n}$ normalization but is order n and is exponentially distributed.

EC.2. Appendix to Section 3: Main Results

EC.2.1. Appendix to Section 3.3

EC.2.1.1. Sub-exponential noise (D2) implies Condition (C1)

LEMMA EC.1. *For a Lispchitz continuous function f, if Condition* (D2) *holds, that is*

$$\sup_{t\ge 0} \mathbb{E}\big[e^{\lambda \|c_t\|} | \mathcal{F}_t\big] < \infty$$

then Condition (C2) *holds that is*

$$\Big[|f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)| \Big| \mathcal{F}_t\Big] \le \alpha_t Y\,, \quad \textit{with} \quad \mathbb{E}[e^{\eta Y}] < \infty \tag{EC.1}$$

for some $\eta > 0$.

Proof. Since $f(x)$ is Lipschitz

$$|f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)| \le K\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| \le \alpha_t K ||\boldsymbol{c}_t||\,.$$

Since Condition (D2) holds, we take $M \ge \sup_t \mathbb{E}\big[e^{\lambda\|\boldsymbol{c}_t\|}\big|\mathcal{F}_t\big]$. We let Y be the random variable with CCDF: $\mathbb{P}(Y \ge y) = 1 \wedge (Me^{-\frac{\lambda}{K}y})$. Thus for $y \in \mathbb{R}_+$

$$\begin{aligned}\mathbb{P}\left(|f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)| \ge \alpha_t y \Big| \mathcal{F}_t\right) &\le \mathbb{P}\left(\|\boldsymbol{c}_t\| \ge y/K \Big| \mathcal{F}_t\right) \\ &\le \min\big\{1, e^{-(\lambda/K)y}\mathbb{E}\big[e^{\lambda\|\boldsymbol{c}_t\|}\big|\mathcal{F}_t\big]\big\} \le \mathbb{P}(Y \ge y)\end{aligned}$$

Above, we apply a Chernoff bound. From this inequality above, we see that Condition (C2) follows from Condition (D2). □

EC.2.1.2. Proof of Lemma 1: sharpness is equivalent to non-vanishing gradient for convex functions. We now prove Lemma 1.

LEMMA 1. *If the function $l(\boldsymbol{x})$ is absolutely continuous then the gradient condition* (D1) *implies the function is sharp,* (D1′)*. Moreover, if the function $l(\boldsymbol{x})$ is convex, then the gradient condition* (D1) *is equivalent to the function being sharp* (D1′)*.*

Proof. First let's assume condition (D1) holds. Let $\boldsymbol{x}(t) = \boldsymbol{x}^\star + (1-t)(\boldsymbol{x} - \boldsymbol{x}^\star)$. Thus we have

$$\begin{aligned} l(\boldsymbol{x}) - l(\boldsymbol{x}^\star) &= \int_0^1 \frac{dl(\boldsymbol{x}(t))}{dt} dt \\ &= \int_0^1 \nabla l(\boldsymbol{x}(t))(\boldsymbol{x}(t) - \boldsymbol{x}^\star) dt \\ &\geq \int_0^1 \kappa \|\boldsymbol{x}(t) - \boldsymbol{x}^\star\| dt \\ &= \int_0^1 (1-t)\kappa \|\boldsymbol{x} - \boldsymbol{x}^\star\| dt \\ &= \frac{\kappa}{2} \|\boldsymbol{x} - \boldsymbol{x}^\star\| \end{aligned}$$

The first equality follows since absolute continuity implies the fundamental theorem of calculus holds. The second equality holds by the chain rule. The third equality follows by the gradient condition (D1). We then apply the definition of $\boldsymbol{x}(t)$ and integrate. Thus as required, we see that condition (D1) implies (D1′).

If we also suppose that the function $l(\boldsymbol{x})$ is convex and that (D1′) holds then

$$l(\boldsymbol{x}^\star) - l(\boldsymbol{x}) \geq \nabla l(\boldsymbol{x})(\boldsymbol{x}^\star - \boldsymbol{x}) \,.$$

So

$$\nabla l(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{x}^\star) \geq l(\boldsymbol{x}) - l(\boldsymbol{x}^\star) \geq \kappa' \|\boldsymbol{x} - \boldsymbol{x}^\star\|$$

The first inequality rearranges the convexity definition above. The second inequality applies the Sharp Condition (D1′). So we see, as required, for a convex function, the Sharp Condition (D1′) implies the gradient condition (D1). □

EC.2.1.3. Proof of Lemma 2: with d or more active constraints, SGD is not normally distributed. Duchi and Ruan (2021) is designed for smooth problems with fewer active constraints than the problem dimension. Once the number of active constraints exceeds the dimension of the problem, then the normal approximation no longer holds. With the Lemma below, we can say that the limiting distribution has an exponential concentration for PSGD. Calculations can speculate the form of the asymptotic optimality; however, the general theory of asymptotic optimality of stochastic optimization is incomplete, particularly in settings where the normal approximation is invalid. As we indicate, it requires a better understanding of asymptotic optimality in the presence of Sharpness.

LEMMA 2. *Suppose that at the optimum* $-\nabla l(\boldsymbol{x}^\star) \in$ *relint* $\mathcal{N}_\mathcal{X}(\boldsymbol{x}^\star)$*, where* $\mathcal{N}_\mathcal{X}(\boldsymbol{x}^\star) := \{\boldsymbol{v} : \boldsymbol{v}^\top(\boldsymbol{x}^\star - \boldsymbol{y}) \leq 0, \forall \boldsymbol{y} \in \mathcal{X}\}$ *and* $\nabla l(\boldsymbol{x}^\star) \neq 0$ *and that there are at least* d *active constraints at* $\boldsymbol{x}^\star$ *(w.l.o.g.* $i = 1, ..., d$*) and*

$$\{\nabla l_i(\boldsymbol{x}^\star) : i = 1, ..., d\} \text{ are linearly independent} \tag{D1''}$$

then the function f *is sharp at* $\boldsymbol{x}^\star$ *and Assumption* (D1) *holds.*

Proof. Let x_∞ be such that $\|\boldsymbol{x}\|_\infty < x_\infty$ for all $x \in \mathcal{X}$. Let $\boldsymbol{c} = \nabla l(\boldsymbol{x}^\star)$. Let $\boldsymbol{c}_i = \nabla l_i(\boldsymbol{x}^\star)$ and $b_i = \boldsymbol{c}_i^\top \boldsymbol{x}^\star$. Let $\mathcal{P} = \{\boldsymbol{x} : \boldsymbol{c}_i^\top \boldsymbol{x} \geq b_i,\ i = 1, ..., d, \|\boldsymbol{x}\|_\infty \leq x_\infty\}$. Notice that by convexity

$$\mathcal{X} \subseteq \mathcal{P}. \tag{EC.2}$$

Notice by linear independence $\boldsymbol{x}^\star$ is the unique point such that $\boldsymbol{c}_i^\top \boldsymbol{x}^\star = b_i,\ i = 1, ..., d$. That is $\boldsymbol{x}^\star$ is an extreme point of the polytope $\mathcal{P}$. Since $-\nabla l(\boldsymbol{x}^\star) \in \text{relint } \mathcal{N}_\mathcal{X}(\boldsymbol{x}^\star)$, $\boldsymbol{x}^\star$ miniminizes $\boldsymbol{c}^\top \boldsymbol{x}$ over $\boldsymbol{x} \in \mathcal{P}$. Thus by Lemma EC.8

$$\boldsymbol{c}^\top(\boldsymbol{x} - \boldsymbol{x}^\star) \geq K\|\boldsymbol{c}\|\|\boldsymbol{x} - \boldsymbol{x}^\star\|, \qquad \forall \boldsymbol{x} \in \mathcal{P}. \tag{EC.3}$$

By convexity

$$l(\boldsymbol{x}) - l(\boldsymbol{x}^\star) \geq \boldsymbol{c}^\top(\boldsymbol{x} - \boldsymbol{x}^\star) \tag{EC.4}$$

combining (EC.2), (EC.3) and (EC.4) we see that

$$l(\boldsymbol{x}) - l(\boldsymbol{x}^\star) \geq K\|\boldsymbol{c}\|\|\boldsymbol{x} - \boldsymbol{x}^\star\|, \qquad \forall \boldsymbol{x} \in \mathcal{X}.$$

Thus we see that the function $l(\boldsymbol{x})$ is sharp on $\mathcal{X}$. Condition (D1) then follows by Lemma 1 since the function $l(\boldsymbol{x})$ is convex. □

EC.2.2. Finite number of Projections for Interior Optimum.

We say a projection step is trivial if $x \in \mathcal{X}$ and thus $\Pi_\mathcal{X}(x) = x$. Otherwise, we say the projection at x is non-trivial. We can show that in instances where the optimum is in the interior only a finite number of non-trivial projections are required.

PROPOSITION EC.1. *Under the assumptions of Theorem 2, if* $\mathcal{X}^\star$ *belongs to the interior of* $\mathcal{X}$*, then the number of (non-trivial) projection steps required by Projected Stochastic Gradient Descent is finite and bounded in expectation.*

Proof. Let the random variable N denote the number of non-trivial projections. By Theorem 2, we have that

$$\mathbb{P}\left(\min_{\boldsymbol{x}^\star \in \mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \geq z\right) \leq Je^{-It^\gamma z}.$$

We let $\tilde{z}$ be the distance from the set of optima to the boundary of $\mathcal{X}$, that is,

$$\tilde{z} = \min_{\boldsymbol{x}\in\mathcal{X}^\star, \boldsymbol{y}\notin\mathcal{X}} \|\boldsymbol{x}^\star - \boldsymbol{y}\|\,.$$

Since $\mathcal{X}^\star$ belongs to the interior of $\mathcal{X}$, we have that $\tilde{z} > 0$. Note that if a projection is non-trivial then $\min_{\boldsymbol{x}^\star\in\mathcal{X}^\star} ||\boldsymbol{x} - \boldsymbol{x}^\star|| \geq \tilde{z}$. Thus

$$N \leq \sum_{t=0}^{\infty} \mathbb{I}\Big[\min_{\boldsymbol{x}^\star\in\mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \geq \tilde{z}\Big]$$

and so, as required,

$$\mathbb{E}[N] \leq \sum_{t=0}^{\infty} \mathbb{P}\left(\min_{\boldsymbol{x}^\star\in\mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \geq \tilde{z}\right) \leq \sum_{t=0}^{\infty} Je^{-It^\gamma z} < \infty\,.$$

□

EC.2.3. Kiefer-Wolfowitz: Proof of Theorem 3

We now restate and prove Theorem 3.

THEOREM 3. *If Conditions* (D1), (D2), (D3) *hold and if*

$$\nu \leq \left(\frac{\kappa}{3cd^{\frac{1}{2}}}\right)^{\frac{1}{2}}$$

then the Kiefer-Wolfowitz algorithm satisfies

$$\mathbb{P}\left(\min_{\boldsymbol{x}\in\mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\| \geq z\right) \leq \hat{J}e^{-\frac{\hat{I}}{\alpha_t}z}, \quad \mathbb{E}\left[\min_{\boldsymbol{x}\in\mathcal{X}^\star} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|\right] \leq \hat{K}\alpha_t, \quad \mathbb{E}\left[l(\boldsymbol{x}_{t+1}) - \min_{\boldsymbol{x}\in\mathcal{X}} l(\boldsymbol{x})\right] \leq \hat{L}\alpha_t$$

where above $\hat{J}, \hat{I}, \hat{K}, \hat{L}$ are positive constants.

Proof. The proof here combines the proof ideas for Kiefer-Wolfowitz, see Fabian (1967) (or, more recently, Broadie et al. (2011)), with the proof in Theorem 2. As with the proof of Theorem 2 our goal is to verify Conditions C1 and C2, so that we can apply Theorem 1.

We can write the KW recursion as

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \alpha_t \nabla l(\boldsymbol{x}_t) + \alpha_t \boldsymbol{\delta}_t + \alpha_t \boldsymbol{\epsilon}_t \tag{EC.5}$$

$$\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{y}_{t+1}) \tag{EC.6}$$

where

$$\boldsymbol{\delta}_t = \nabla l(\boldsymbol{x}) - \frac{\boldsymbol{l}(\boldsymbol{x}_t + \boldsymbol{\nu}_t) - \boldsymbol{l}(\boldsymbol{x}_t - \boldsymbol{\nu}_t)}{2\nu_t}$$

$$\boldsymbol{\epsilon}_t = \frac{\boldsymbol{l}(\boldsymbol{x}_t + \boldsymbol{\nu}_t) - \boldsymbol{l}(\boldsymbol{x}_t - \boldsymbol{\nu}_t)}{2\nu_t} - \frac{\boldsymbol{l}(\boldsymbol{x}_t + \boldsymbol{\nu}_t, \hat{w}_t^+) - \boldsymbol{l}(\boldsymbol{x}_t - \boldsymbol{\nu}_t, \hat{w}_t^-)}{2\nu_t}\,.$$

Letting $\boldsymbol{x}_t^\star$ be the projection of $\boldsymbol{x}_t$ onto $\mathcal{X}^\star$, then

$$\begin{aligned}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1}^\star\|^2 &\le \|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t^\star\|^2\\ &\le \|\boldsymbol{y}_{t+1}-\boldsymbol{x}_t^\star\|^2 = \|\boldsymbol{y}_{t+1}-\boldsymbol{x}_t+\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|^2\\ &= \|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|^2 - 2\alpha_t\nabla l(\boldsymbol{x}_t)^\top(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)+2\alpha_t\boldsymbol{\delta}_t^\top(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)+2\alpha_t\boldsymbol{\epsilon}_t^\top(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)+\|\boldsymbol{y}_{t+1}-\boldsymbol{x}_t\|^2.\end{aligned}$$

The first inequality above follows since $\boldsymbol{x}_{t+1}^\star$ is a projection. The second follows since $\boldsymbol{x}_{t+1}$ is a projection. We then expand.

We let E_t be the positive number defined below in (EC.16). On the event $\{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|\ge E_t\}$, we have

$$\begin{aligned}&\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1}^\star\|\\ \le&\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|\sqrt{1-2\alpha_t\nabla l(\boldsymbol{x}_t)^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|^2}+2\alpha_t\boldsymbol{\delta}_t^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|^2}+2\alpha_t\boldsymbol{\epsilon}_t^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|^2}+\frac{\|\boldsymbol{y}_{t+1}-\boldsymbol{x}_t\|^2}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|^2}}\\ \le&\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|\\ &-\alpha_t\nabla l(\boldsymbol{x}_t)^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|} &\text{(EC.7)}\\ &+\alpha_t\boldsymbol{\delta}_t^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|} &\text{(EC.8)}\\ &+\alpha_t\boldsymbol{\epsilon}_t^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|} &\text{(EC.9)}\\ &+\frac{\|\boldsymbol{y}_{t+1}-\boldsymbol{x}_t\|^2}{2\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|}. &\text{(EC.10)}\end{aligned}$$

We now analyse the conditional expectation of the four terms above. Term (EC.7) is bounded using to the sharpness condition (D1)

$$-\nabla l(\boldsymbol{x}_t)^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}-\boldsymbol{x}^\star\|}\le-\kappa. \tag{EC.11}$$

Term (EC.8) is bounded by the Taylor approximation condition (D3). Specifically

$$\boldsymbol{\delta}_t^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|}\le\|\boldsymbol{\delta}_t\|=\left\|\nabla l(\boldsymbol{x})-\frac{\boldsymbol{l}(\boldsymbol{x}_t+\boldsymbol{\nu})-\boldsymbol{l}(\boldsymbol{x}_t-\boldsymbol{\nu})}{2\nu}\right\|\le cd^{\frac{1}{2}}\nu^2. \tag{EC.12}$$

Term (EC.9) has zero mean

$$\mathbb{E}\left[\boldsymbol{\epsilon}_t^\top\frac{(\boldsymbol{x}_t-\boldsymbol{x}_t^\star)}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|}\Big|\mathcal{F}_t\right]=0. \tag{EC.13}$$

For Term (EC.10), $\boldsymbol{y}_{t+1}=\boldsymbol{x}_t-\alpha_t\boldsymbol{c}_t$

$$\mathbb{E}\left[\frac{\|\boldsymbol{y}_{t+1}-\boldsymbol{x}_t\|^2}{\|\boldsymbol{x}_t-\boldsymbol{x}_t^\star\|}\Big|\mathcal{F}_t\right]\le\frac{\alpha_t^2}{E_t}\mathbb{E}\left[||\boldsymbol{c}_t||^2\big|\mathcal{F}_t\right]\le\frac{\alpha_t^2\sigma_l^2}{2E_t\nu^2} \tag{EC.14}$$

Since the variance of $l(\boldsymbol{x}, \hat{w})$ is bounded (by σ_l^2), the variance of $||\boldsymbol{c}_t||$ is bounded. Above, we let σ_l^2/ν^2 define this upper bound. Applying bounds (EC.11), (EC.12), (EC.13) and (EC.14) respectively to the terms (EC.7), (EC.8), (EC.9) and (EC.10) gives

$$\mathbb{E}\Big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\|\Big|\mathcal{F}_t\Big] \leq \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| - \alpha_t\kappa + \alpha_t c\nu^2 + \frac{\alpha_t^2\sigma_l^2}{2E_t\nu^2}\,. \tag{EC.15}$$

Notice if we choose

$$\nu \leq \sqrt{\frac{\kappa}{3cd^{\frac{1}{2}}}} \qquad \text{and} \qquad E_t = \frac{3\sigma_l^2}{4\nu^2\kappa}\alpha_t\,, \tag{EC.16}$$

then application to (EC.15) gives

$$\mathbb{E}\Big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\|\Big|\mathcal{F}_t\Big] \leq \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| - \alpha_t\frac{\kappa}{3} \qquad \text{on the event} \qquad \Big\{\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| \geq \frac{3\sigma_l^2}{\nu^4\kappa}\alpha_t\Big\}\,.$$

This verifies that Condition (C1) of Theorem 1 holds.

We must also verify Condition (C2). (The argument that follows is more-or-less identical to the verification of (C2) in Theorem 2.) For this notice that

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\| \leq \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t^\star\| \leq \|\boldsymbol{y}_{t+1} - \boldsymbol{x}_t^\star\| \leq \|\boldsymbol{y}_{t+1} - \boldsymbol{x}_t\| + \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| = \alpha_t\|\boldsymbol{c}_t\| + \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|\,.$$

and

$$\begin{aligned}\|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\| \leq \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}^\star\| &\leq \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| + \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\| \\ &\leq \|\boldsymbol{y}_{t+1} - \boldsymbol{x}_t\| + \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\| = \alpha_t\|\boldsymbol{c}_t\| + \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\|\,.\end{aligned}$$

Thus

$$\Big|\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^\star\| - \|\boldsymbol{x}_t - \boldsymbol{x}_t^\star\|\Big| \leq \alpha_t\|\boldsymbol{c}_t\| \tag{EC.17}$$

Since Condition (D2) holds, we take $M \geq \sup_t \mathbb{E}\big[e^{\lambda\|\boldsymbol{c}_t\|}\big|\mathcal{F}_t\big]$. We let Y be the random variable with CCDF: $\mathbb{P}(Y \geq y) = 1 \wedge (Me^{-\lambda y})$. Thus for $y \in \mathbb{R}_+$

$$\begin{aligned}\mathbb{P}\left(|f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)| \geq \alpha_t y\Big|\mathcal{F}_t\right) &\leq \mathbb{P}\left(\|\boldsymbol{c}_t\| \geq y\Big|\mathcal{F}_t\right) \\ &\leq \min\left\{1, e^{-\lambda y}\mathbb{E}\big[e^{\lambda\|\boldsymbol{c}_t\|}\big|\mathcal{F}_t\big]\right\} \leq \mathbb{P}(Y \geq y)\end{aligned}$$

Above, we apply (EC.17) and a Chernoff bound. From this inequality above, we see that Condition (C2) follows from Condition (D2).

We can now apply Theorem 1 which gives:

$$\mathbb{P}\Big(\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\| \geq z\Big) \leq \hat{I}e^{-\frac{\hat{J}}{\alpha_t}z} \quad \text{and} \quad \mathbb{E}\left[\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|\right] \leq \hat{K}\alpha_t$$

for constants $\hat{I}$, $\hat{J}$ and $\hat{K}$. Since we also assume in addition that $l : \mathcal{X} \to \mathbb{R}$ is Lispchitz continuous (with Lipschitz constant $\hat{L}/\hat{K}$) we have, as required,

$$\mathbb{E}\Big[l(\boldsymbol{x}_{t+1}) - \min_{\boldsymbol{x}\in\mathcal{X}} l(\boldsymbol{x})\Big] \leq \frac{\hat{L}}{\hat{K}}\mathbb{E}\left[\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|\right] \leq \hat{L}\alpha_t\,.$$

□

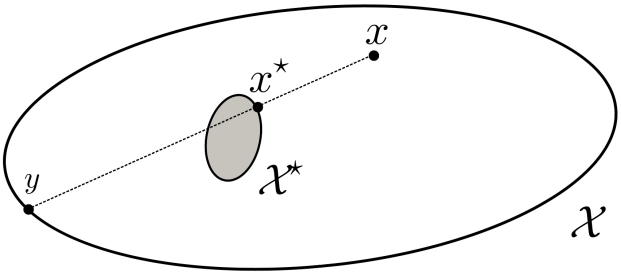


Figure EC.1 Here we give the $\boldsymbol{x}$, $\boldsymbol{x}^\star$ and $\boldsymbol{y}$ terms from Lemma EC.2. Here we take $\boldsymbol{x}$ project onto $\mathcal{X}^\star$ to give $\boldsymbol{x}^\star$, then $\boldsymbol{y}$ is the boundary value on the line passing from $\boldsymbol{x}$ to $\boldsymbol{x}^\star$

EC.2.4. Stochastic Frank-Wolfe: Proof of Theorem 4

The main aim of this section is to prove Theorem 4. We also show that the distance function satisfies the conditions of our main theorem. This suggests that if the objective function behaves linearly rather than quadratically near the optimum, we should anticipate faster convergence. We also discuss how linear convergence can hold for Stochastic Frank-Wolfe in the same manner that we proved for Projected Stochastic Gradient Descent.

Before proceeding with the proof of Theorem 4 we require a couple of lemmas. Lemma EC.2 is used to show that there is sufficient negative drift in the Frank-Wolfe algorithm.

LEMMA EC.2. *If Condition* (D1) *and Condition* (E2) *hold then there exists a* $\hat{\kappa} > 0$ *such that for all* $\boldsymbol{x} \in \mathcal{X} \backslash \mathcal{X}^\star$ *there exists* $\boldsymbol{y} \in \mathcal{X}$ *such that*

$$(\boldsymbol{y} - \boldsymbol{x})^\top \nabla l(\boldsymbol{x}) \leq -\hat{\kappa}\,.$$

Proof. The idea of the proof is as follows. The derivative from $\boldsymbol{x}$ and $\boldsymbol{x}^\star$ at $\boldsymbol{x}$ bounded above by $-\kappa$, by Assumption (D1). Since $\boldsymbol{x}^\star$ is in the interior by (E2), we can increase the directional derivative further by replacing $\boldsymbol{x}^\star$ with $\boldsymbol{y}$, where $\boldsymbol{y}$ is the point on the boundary of $\mathcal{X}$ on the line between $\boldsymbol{x}$ and $\boldsymbol{x}^\star$. See Figure EC.1. We now proceed with the formal argument.

By Condition (E2), there exists a constant $d > 0$ such that

$$\min_{\boldsymbol{x}^\star \in \mathcal{X}^\star, \boldsymbol{y} \in \partial\mathcal{X}} \|\boldsymbol{y} - \boldsymbol{x}^\star\| \geq d\,. \tag{EC.18}$$

(Here $\partial\mathcal{X} := \bar{\mathcal{X}} \backslash \mathcal{X}^\circ$ is the boundary of $\mathcal{X}$.) By Condition (D1), for all $\boldsymbol{x} \notin \mathcal{X}^\star$ there exists $\boldsymbol{x}^\star \in \mathcal{X}^\star$

$$\frac{(\boldsymbol{x}^\star - \boldsymbol{x})^\top}{\|\boldsymbol{x}^\star - \boldsymbol{x}\|} \nabla l(\boldsymbol{x}) \leq -\kappa\,. \tag{EC.19}$$

We let $\boldsymbol{y}(t) = \boldsymbol{x} + t(\boldsymbol{x}^\star - \boldsymbol{x})$ for $t \in \mathbb{R}$. Notice that

$$\frac{(\boldsymbol{y}(t) - \boldsymbol{x})^\top}{\|\boldsymbol{y}(t) - \boldsymbol{x}\|} = \frac{(\boldsymbol{x}^\star - \boldsymbol{x})^\top}{\|\boldsymbol{x}^\star - \boldsymbol{x}\|} \tag{EC.20}$$

Letting $t^\star = \max\{t : \boldsymbol{y}(t) \in \mathcal{X}\}$, we see that

$$\boldsymbol{y} := \boldsymbol{y}(t^\star) \in \delta\mathcal{X} \tag{EC.21}$$

Combining (EC.18-EC.21), we see that

$$(\boldsymbol{y}-\boldsymbol{x})^\top \nabla l(\boldsymbol{x}) = \|\boldsymbol{y}-\boldsymbol{x}\| \frac{(\boldsymbol{x}^\star-\boldsymbol{x})^\top}{\|\boldsymbol{x}^\star-\boldsymbol{x}\|}\nabla l(\boldsymbol{x}) \leq -d\kappa =: -\hat{\kappa}$$

as required. □

We now restate and prove Theorem 4

THEOREM 4. *For learning rates of the form $\alpha_t = a/(u+t)^\gamma$ with $a, u > 0$ and $\gamma \in [0,1]$, if Conditions* (D1), (D2), (E1) *and* (E2) *hold and if $m_t \geq (3\sigma/\kappa\alpha_t)^2$ then the stochastic Frank-Wolfe algorithm satisfies*

$$\mathbb{P}\left(l(\boldsymbol{x}_{t+1}) - \min_{\boldsymbol{x}\in\mathcal{X}} l(\boldsymbol{x}) \geq z\right) \leq Ie^{-\frac{J}{\alpha_t}z},$$

for constants I,J.

Proof of Theorem 4. The proof here combines ideas from Theorem 2 with the adjustments for stochastic effects for the Frank-Wolfe algorithm given in Theorem 3 from Hazan and Luo (2016).

In the proof we define $D := \max_{\boldsymbol{x},\boldsymbol{v}} \|\boldsymbol{x}-\boldsymbol{v}\|$ and we let $\epsilon(\boldsymbol{x}_t) = l(\boldsymbol{x}_t) - l(\boldsymbol{x}^\star)$ and we define σ such that

$$\mathbb{E}\left[\|\nabla l(\boldsymbol{x}_t) - \boldsymbol{c}_t^i\|^2\right] \leq \sigma^2 \qquad\qquad \forall i, t.$$

(Note that σ is finite by the moment generating function condition (D2))

By Condition (E1)

$$\begin{aligned}\frac{\epsilon(\boldsymbol{x}_{t+1})^2}{2} - \frac{\epsilon(\boldsymbol{x}_t)^2}{2} &\leq \epsilon(\boldsymbol{x}_t)\big(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\big)^\top \nabla\epsilon(\boldsymbol{x}_t) + \frac{K}{2}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2 \\ &= \alpha_t\epsilon(\boldsymbol{x}_t)(\boldsymbol{v}_t-\boldsymbol{x}_t)^\top\boldsymbol{c}_t + \alpha_t\epsilon(\boldsymbol{x}_t)(\boldsymbol{v}_t-\boldsymbol{x}_t)^\top\left[\nabla\epsilon(\boldsymbol{x}_t)-\boldsymbol{c}_t\right] + \frac{K}{2}\alpha_t^2\|\boldsymbol{v}_t-\boldsymbol{x}_t\|^2.\end{aligned} \tag{EC.22}$$

We now consider the event where the following bound holds

$$\left\{\epsilon(\boldsymbol{x}_t) \geq \frac{3\alpha_t K D^2}{2\kappa}\right\}. \tag{EC.23}$$

Thus

$$\begin{aligned}&\epsilon(\boldsymbol{x}_{t+1}) \\ &\leq \sqrt{\epsilon(\boldsymbol{x}_t)^2 + 2\alpha_t\epsilon(\boldsymbol{x}_t)(\boldsymbol{v}_t-\boldsymbol{x}_t)^\top\boldsymbol{c}_t + 2\alpha_t\epsilon(\boldsymbol{x}_t)(\boldsymbol{v}_t-\boldsymbol{x}_t)^\top\left[\nabla\epsilon(\boldsymbol{x}_t)-\boldsymbol{c}_t\right] + K\alpha_t^2\|\boldsymbol{v}_t-\boldsymbol{x}_t\|^2} \\ &= \epsilon(\boldsymbol{x}_t)\sqrt{1 + 2\frac{\alpha_t}{\epsilon(\boldsymbol{x}_t)}(\boldsymbol{v}_t-\boldsymbol{x}_t)^\top\boldsymbol{c}_t + 2\frac{\alpha_t}{\epsilon(\boldsymbol{x}_t)}(\boldsymbol{v}_t-\boldsymbol{x}_t)^\top\left[\nabla\epsilon(\boldsymbol{x}_t)-\boldsymbol{c}_t\right] + K\frac{\alpha_t^2}{\epsilon(\boldsymbol{x}_t)^2}\|\boldsymbol{v}_t-\boldsymbol{x}_t\|^2} \\ &\leq \epsilon(\boldsymbol{x}_t) + \alpha_t(\boldsymbol{v}_t-\boldsymbol{x}_t)^\top\boldsymbol{c}_t + \alpha_t(\boldsymbol{v}_t-\boldsymbol{x}_t)^\top\left[\nabla\epsilon(\boldsymbol{x}_t)-\boldsymbol{c}_t\right] + \frac{K}{2}\frac{\alpha_t^2}{\epsilon(\boldsymbol{x}_t)}\|\boldsymbol{v}_t-\boldsymbol{x}_t\|^2 \\ &\leq \epsilon(\boldsymbol{x}_t) + \alpha_t(\boldsymbol{y}_t-\boldsymbol{x}_t)^\top\boldsymbol{c}_t + \alpha_t\|\boldsymbol{v}_t-\boldsymbol{x}_t\|\|\nabla\epsilon(\boldsymbol{x}_t)-\boldsymbol{c}_t\| + \frac{\alpha_t\kappa}{3} \\ &\leq \epsilon(\boldsymbol{x}_t) + \alpha_t(\boldsymbol{y}_t-\boldsymbol{x}_t)^\top\boldsymbol{c}_t + \alpha_t D\|\nabla\epsilon(\boldsymbol{x}_t)-\boldsymbol{c}_t\| + \frac{\alpha_t\kappa}{3}\end{aligned} \tag{EC.24}$$

In the first inequality, above we rearrange the expression (EC.22). In the second inequality, we apply the inequality $\sqrt{1+z} \leq 1 + z/2$. In the third equality, we note that $\boldsymbol{v}_t^\top \boldsymbol{c}_t \leq \boldsymbol{y}_t^\top \boldsymbol{c}_t$, by the definition of $\boldsymbol{v}_t$ (11b). Here we let $\boldsymbol{y}_t \in \mathcal{X}$ be as defined in Lemma EC.2. Also, we apply the Cauchy-Schwarz Inequality and the bound (EC.23). In the final inequality we note that $\|\boldsymbol{v}_t - \boldsymbol{x}_t\| \geq D$.

Taking the conditional expectation of (EC.24), we see that, on the event (EC.23), the following holds

$$\begin{aligned}\mathbb{E}\left[l(\boldsymbol{x}_{t+1}) - l(\boldsymbol{x}_t)|\mathcal{F}_t\right] &= \mathbb{E}\left[\epsilon(\boldsymbol{x}_{t+1}) - \epsilon(\boldsymbol{x}_t)|\mathcal{F}_t\right] \\ &\leq \alpha_t(\boldsymbol{y}_t - \boldsymbol{x}_t)^\top \mathbb{E}[\boldsymbol{c}_t|\mathcal{F}_t] + \alpha_t D\mathbb{E}\left[\|\nabla\epsilon(\boldsymbol{x}_t) - \boldsymbol{c}_t\||\mathcal{F}_t\right] + \frac{\alpha_t\kappa}{3} \\ &\leq -\alpha_t\hat{\kappa} + \alpha_t D\mathbb{E}\left[\|\nabla l(\boldsymbol{x}_t) - \boldsymbol{c}_t\||\mathcal{F}_t\right] + \frac{\alpha_t\kappa}{3}. \end{aligned} \tag{EC.25}$$

Notice that, since $m_t \geq (3\sigma D/\hat{\kappa}\alpha_t)^2$,

$$\mathbb{E}\left[\|\nabla l(\boldsymbol{x}_t) - \boldsymbol{c}_t\|\,|\mathcal{F}_t\right] \leq \sqrt{\mathbb{E}\left[\|\nabla l(\boldsymbol{x}_t) - \boldsymbol{c}_t\|^2\,|\mathcal{F}_t\right]} \leq \frac{\sigma}{\sqrt{m_t}} \leq \frac{\hat{\kappa}}{3D}.$$

Now applying this inequality to (EC.25) gives

$$\mathbb{E}\left[l(\boldsymbol{x}_{t+1}) - l(\boldsymbol{x}_t)|\mathcal{F}_t\right] \leq -\alpha_t\frac{\hat{\kappa}}{3}$$

on the event $l(\boldsymbol{x}_t) - l(\boldsymbol{x}^\star) \geq 3KD/\alpha\hat{\kappa}$. Thus Condition (C1) is met.

For Condition (C2), since l is Lipschitz continuous and the set $\mathcal{X}$ is bounded we have

$$\|l(\boldsymbol{x}_{t+1}) - l(\boldsymbol{x}_t)\| \leq L\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| \leq \alpha_t L\|\boldsymbol{v}_t - \boldsymbol{x}_t\| \leq 2\alpha_t L\max_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|.$$

Thus we see that Condition (C2) holds with a constant upperbound $Y = 2L\max_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|$.

Here we see that the conditions of Theorem 1 are met, and thus we have that

$$\mathbb{P}\left(l(\boldsymbol{x}_{t+1}) - l(\boldsymbol{x}_t) \geq z\right) \leq Ie^{-\frac{J}{\alpha_t}z},$$

as required. □

EC.2.4.1. Cones satisfy Condition (E1) Below we recall that we define the matrix norm $\|\cdot\|_S$ for a positive semi-definite matrix S by

$$\|\boldsymbol{x}\|_S := \sqrt{\boldsymbol{x}^\top S\boldsymbol{x}}$$

Lemma EC.3. *For a symmetric positive definite matrix S, the distance function*

$$d_{\mathcal{X}^\star}(\boldsymbol{x}) = \min_{x^\star\in\mathcal{X}^\star}\|\boldsymbol{x} - \boldsymbol{x}^\star\|_S$$

satisfies Condition (E1).

Proof. We must show that the function

$$d_{\mathcal{X}^\star}(\boldsymbol{x})^2 = \min_{x^\star \in \mathcal{X}^\star} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_S^2$$

is strongly convex.

Given $\boldsymbol{x}$, we let $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}^\star \in \mathcal{X}^\star} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_S$. By the Envelope Theorem,

$$\nabla d_{\mathcal{X}}(\boldsymbol{x})^2 = 2S(\boldsymbol{x} - \boldsymbol{x}^\star) \tag{EC.26}$$

Also

$$\|\boldsymbol{x} - \boldsymbol{y}\|_S^2 \leq \lambda_{\max}(S)\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{EC.27}$$

where $\lambda_{\max}(S)$ is the maximum eigenvalue of S.

Now for any $\boldsymbol{y}$ and $\boldsymbol{x}$,

$$\begin{aligned} d_{\mathcal{X}^\star}(\boldsymbol{y})^2 = \min_{\boldsymbol{y}^\star \in \mathcal{X}^\star} \|\boldsymbol{y} - \boldsymbol{y}^\star\|_S^2 &\leq \|\boldsymbol{y} - \boldsymbol{x}^\star\|^2 \\ &= \|\boldsymbol{y} - \boldsymbol{x} + \boldsymbol{x} - \boldsymbol{x}^\star\|_S^2 \\ &= \|\boldsymbol{x} - \boldsymbol{x}^\star\|_S^2 + 2(\boldsymbol{y} - \boldsymbol{x})^\top S(\boldsymbol{x} - \boldsymbol{x}^\star) + \|\boldsymbol{y} - \boldsymbol{x}\|_S^2 \\ &\leq d_{\mathcal{X}^\star}(\boldsymbol{x})^2 + (\boldsymbol{y} - \boldsymbol{x})\nabla d_{\mathcal{X}^\star}^2(\boldsymbol{x}) + \lambda_{\max}(S)\|\boldsymbol{x} - \boldsymbol{y}\|^2 \end{aligned}$$

In the first inequality, we apply the sub-optimality of $\boldsymbol{x}^\star$ with respect to the point $\boldsymbol{y}$. In the second inequality, we apply (EC.26) and (EC.27). Thus from the above inequality we see that $d_{\mathcal{X}^\star}(x)^2$ is a $\lambda_{\max}(S)$–smoothly convex function, as required. □

EC.2.5. Stochastic Frank-Wolfe Boundary case

Proposition EC.2. *For learning rates of the form $\alpha_t = a/(u+t)^\gamma$ with $a, u > 0$ and $\gamma \in [0,1)$, if Conditions* (D1), (D2), (E1) *and* (E2) *hold and if $m_t \geq (2\sigma E/KD\alpha_t)^2$ then the stochastic Frank-Wolfe algorithm satisfies*

$$\limsup_{t\to\infty} \frac{1}{\sqrt{\alpha_t}} \mathbb{E}\left[l(\boldsymbol{x}_{t+1}) - l(\boldsymbol{x}^\star)\right] < \infty,$$

Proof. The proof here combines ideas from Theorem 2 with the adjustments for stochastic effects for the Frank-Wolfe algorithm given in Theorem 3 from Hazan and Luo (2016).

In the proof we let $\epsilon(\boldsymbol{x}_t) = l(\boldsymbol{x}_t) - l(\boldsymbol{x}^\star)$ and we define σ such that

$$\mathbb{E}\left[\|\nabla l(\boldsymbol{x}_t) - \boldsymbol{c}_t^i\|^2\right] \leq \sigma^2 \qquad \forall i, t.$$

(Note that σ is finite by the moment generating function condition (D2)). We define $D := \max_{\boldsymbol{x},\boldsymbol{v}} \|\boldsymbol{x} - \boldsymbol{v}\|$ and $E := \max_{\boldsymbol{x}} \epsilon(\boldsymbol{x})$.

By Condition (E1)

$$\begin{aligned}
&\frac{\epsilon(\boldsymbol{x}_{t+1})^2}{2} - \frac{\epsilon(\boldsymbol{x}_t)^2}{2} \\
&\leq \epsilon(\boldsymbol{x}_t)\big(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\big)^\top \nabla\epsilon(\boldsymbol{x}_t) + \frac{K}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 \\
&= \alpha_t\epsilon(\boldsymbol{x}_t)(\boldsymbol{v}_t - \boldsymbol{x}_t)^\top \boldsymbol{c}_t + \alpha_t\epsilon(\boldsymbol{x}_t)(\boldsymbol{v}_t - \boldsymbol{x}_t)^\top\left[\nabla\epsilon(\boldsymbol{x}_t) - \boldsymbol{c}_t\right] + \frac{K}{2}\alpha_t^2\|\boldsymbol{v}_t - \boldsymbol{x}_t\|^2 \\
&\leq \alpha_t\epsilon(\boldsymbol{x}_t)(\boldsymbol{x}^\star - \boldsymbol{x}_t)^\top \boldsymbol{c}_t + \alpha_t\epsilon(\boldsymbol{x}_t)\|\boldsymbol{v}_t - \boldsymbol{x}_t\|\|\nabla\epsilon(\boldsymbol{x}_t) - \boldsymbol{c}_t\| + \frac{K}{2}\alpha_t^2\|\boldsymbol{v}_t - \boldsymbol{x}_t\|^2 \\
&\leq \alpha_t\epsilon(\boldsymbol{x}_t)(\boldsymbol{x}^\star - \boldsymbol{x}_t)^\top \boldsymbol{c}_t + \alpha_t ED\|\nabla\epsilon(\boldsymbol{x}_t) - \boldsymbol{c}_t\| + \alpha_t^2\frac{KD^2}{2}
\end{aligned} \tag{EC.28}$$

Notice that, since $m_t \geq (2\sigma E/KD\alpha_t)^2$,

$$\mathbb{E}\big[\|\nabla l(\boldsymbol{x}_t) - \boldsymbol{c}_t\|\,|\mathcal{F}_t\big] \leq \sqrt{\mathbb{E}\left[\|\nabla l(\boldsymbol{x}_t) - \boldsymbol{c}_t\|^2\,|\mathcal{F}_t\right]} \leq \frac{\sigma}{\sqrt{m_t}} \leq \alpha_t\frac{KD}{2E}\,. \tag{EC.29}$$

Taking expectations in (EC.28) gives

$$\begin{aligned}
&\mathbb{E}\left[\frac{\epsilon(\boldsymbol{x}_{t+1})^2}{2}\right] - \mathbb{E}\left[\frac{\epsilon(\boldsymbol{x}_t)^2}{2}\right] \\
&\leq \alpha_t\mathbb{E}\left[\epsilon(\boldsymbol{x}_t)\boldsymbol{c}_t^\top(\boldsymbol{x}^\star - \boldsymbol{x}_t)\right] + \alpha_t ED\mathbb{E}\left[\|\nabla\epsilon(\boldsymbol{x}_t) - \boldsymbol{c}_t\|\right] + \alpha_t^2\frac{KD^2}{2} \\
&\leq \alpha_t\mathbb{E}\left[\epsilon(\boldsymbol{x}_t)\nabla\epsilon(\boldsymbol{x}_t)^\top(\boldsymbol{x}^\star - \boldsymbol{x}_t)\right] + \alpha_t ED\mathbb{E}\left[\|\nabla\epsilon(\boldsymbol{x}_t) - \boldsymbol{c}_t\|\right] + \alpha_t^2\frac{KD^2}{2} \\
&\leq \alpha_t\mathbb{E}\left[\left(\frac{\nabla\epsilon(\boldsymbol{x}_t)^2}{2}\right)^\top(\boldsymbol{x}^\star - \boldsymbol{x}_t)\right] + \alpha_t^2 KD^2 \\
&\leq -\alpha_t\frac{\epsilon(\boldsymbol{x}_t)^2}{2} + \alpha_t^2 KD^2
\end{aligned}$$

In the third equality above, we apply (EC.29). In the final equality, we apply the convexity of $\epsilon(\boldsymbol{x})^2$. Thus, we see that

$$\mathbb{E}\left[\frac{\epsilon(\boldsymbol{x}_{t+1})^2}{2}\right] \leq (1-\alpha_t)\mathbb{E}\left[\frac{\epsilon(\boldsymbol{x}_t)^2}{2}\right] + \alpha_t^2 KD^2\,.$$

Consequently, by Lemma EC.5 (and Lemma EC.4) given below

$$\limsup_{t\to\infty}\frac{1}{\alpha_t}\mathbb{E}[\epsilon(\boldsymbol{x}_{t+1})^2/2] < \infty$$

Thus

$$\limsup_{t\to\infty}\frac{\mathbb{E}[\epsilon(\boldsymbol{x}_{t+1})]}{\sqrt{\alpha_t}} \leq \left(\frac{\mathbb{E}[\epsilon(\boldsymbol{x}_{t+1})^2]}{\alpha_t}\right)^{1/2} < \infty\,.$$

□

We require the following techincal lemma which we then extend.

LEMMA EC.4. *If ξ_n is a positive sequence such that*

$$\xi_{n+1} \le \xi_n\Big(1 - A\alpha_n\Big) + \alpha_n B$$

and

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \qquad \limsup_{n\to\infty} \alpha_n \le 0$$

then

$$\limsup_{n\to\infty} \xi_n \le \frac{B}{A}\,.$$

Proof. Rearranging gives

$$\xi_{n+1} - \xi_n \le -\alpha_n(A\xi_n - B).$$

If $\xi_n > B/A + \epsilon$ for some $\epsilon > 0$ then

$$\xi_{n+1} - \xi_n \le -\alpha_n(A\xi_n - B) \le -\alpha_n(A[B/A + \epsilon] - B) = -\alpha_n A\epsilon\,.$$

So ξ_n is decreasing when $\xi_n > B/A + \epsilon$ holds and, since $\sum_n \alpha_n = \infty$, there exists N s.t. $\xi_N \le B/A + \epsilon$. Let N_0 be the first value of N where $\xi_N \le B/A + \epsilon$ occurs.

Notice, ξ_n can only increase when $\xi_n \le B/A + \epsilon$, and since ξ_n is a positive then

$$\xi_{n+1} \le \xi_n + \alpha_n B\,.$$

Thus, we see that

$$\xi_n \le \frac{B}{A} + \epsilon + \alpha_n A\epsilon, \qquad \forall n \ge N_0\,.$$

Therefore

$$\limsup_{n\to\infty} \xi_n \le \frac{B}{A} + \epsilon + \limsup_{n\to\infty} \alpha_n B \le \frac{B}{A} + \epsilon\,.$$

Since ϵ is arbitrary the results holds. □

The following is an extension of the above lemma. Note for $\beta_n = \alpha_n = a/(u+t)^\gamma$ below and , for $\gamma < 1$, we can take $C = 0$ below. (We can consider the case $\gamma = 1$, but we require to take a sufficiently small.)

LEMMA EC.5. *If ξ_n is a positive sequence such that*

$$\xi_{n+1} \le \xi_n\Big(1 - A\alpha_n\Big) + \alpha_n\beta_n B$$

and

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \qquad \lim_{n\to\infty} \alpha_n = 0, \quad \frac{\beta_n}{\beta_{n+1}} \le (1 + C\alpha_n)$$

with $A > C$ then

$$\limsup_{n\to\infty} \frac{\xi_n}{\beta_n} \le \frac{A - C}{B}.$$

Proof. Since $\lim_{n\to\infty}\alpha_n = 0$, take N such that $\alpha_n < \delta$ for all $n \geq N$.

Now defining $\xi'_n = \xi_n/\beta_n$ for $n \geq N$ gives

$$\begin{aligned}
\xi'_{n+1} = \frac{\xi_{n+1}}{\beta_{n+1}} &\leq \frac{\beta_n}{\beta_{n+1}}\Big(1 - A\alpha_n\Big)\xi'_n + \alpha_n \frac{\beta_n}{\beta_{n+1}} B \\
&\leq (1 + C\alpha_n)\,(1 - A\alpha_n)\,\xi'_n + \alpha_n\,(1 + C\alpha_n)\,B \\
&\leq (1 - (A - C + \delta)\alpha_n)\,\xi'_n + \alpha_n(1 + C\delta)B \\
&= (1 - A'\alpha_n)\xi'_n + \alpha_n B'
\end{aligned}$$

where we define $A' = A - C + \delta$ and $B' = (1 + C\delta)B$. Applying Lemma EC.4 gives

$$\limsup_{n\to\infty} \xi'_n \leq \frac{A'}{B'},$$

which recalling the definitions of ξ'_n, A',B' and recalling that δ is arbitrary gives the result. □

EC.2.6. Appendix to Section 3.6 : Linear Convergence Proofs

As discussed, our proof follows the main argument of Theorem 3.2 of Davis et al. (2019). We divide the procedure into S stages. We consider PSGD with constant step size within each stage, as defined in (12). The task of each stage is to half the error with the optimum. We apply our bound Lemma EC.6, which is a stronger concentration bound than Theorem 4.1, used in Davis et al. (2019). This leads to some improvements in the bounds found there.

EC.2.6.1. Exponential Concentration for constant step-size and unbounded state-space. Below, we state an exponential concentration bound for constant step sizes. We do not require the function $f(x)$ or the set $\mathcal{X}$ to be bounded (or constrained) for this result to hold.

Lemma EC.6. *For constant step sizes α*

$$\mathbb{P}(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star) \geq z | \mathcal{F}_0) \leq e^{-\frac{Q}{\alpha}z}\left\{ e^{\frac{Q}{\alpha}(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star))} e^{-t\frac{Q}{\alpha}\kappa/2} + D\frac{e^{Q\kappa/2}}{1 - e^{-Q\kappa/2}} e^{QB} \right\}.$$

Proof. There are no boundedness assumptions placed in Lemma 4. We restate the conclusion of that result here:

$$\mathbb{E}[e^{\eta L_{t+1}} | \mathcal{F}_{T_0}] \leq \mathbb{E}[e^{\eta L_{T_1}} | \mathcal{F}_{T_0}] \prod_{k=T_1}^{t} \rho_t + D \sum_{\tau=T_1+1}^{t+1} \prod_{k=\tau}^{t} \rho_k, \tag{EC.30}$$

for $t \geq T_1 \geq T_0$. If we consider the above terms for constant step sizes $\alpha = \alpha_t$ then

$$\begin{aligned}
L_{t+1} &= f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^\star) - \alpha B \\
\rho_t = \rho &:= e^{-\alpha\eta\kappa + \alpha^2\eta^2 E} \leq e^{-\alpha\eta\frac{\kappa}{2}} \qquad \text{for} \qquad \alpha\eta \leq Q \\
T_0 &= \min\{t : \frac{\alpha_t - \alpha_{t+1}}{\alpha_t} \leq \frac{\kappa}{2B}\} = 0 \\
T_1 &= 0,
\end{aligned}$$

also

$$\prod_{k=0}^{t}\rho_k=\rho^{t+1}\le 1 \quad \text{and} \quad \sum_{l=1}^{t+1}\prod_{k=l}^{t}\rho_k=\sum_{l=1}^{t+1}\rho^{t-l}\le\sum_{l=-1}^{\infty}\rho^l=\frac{\rho^{-1}}{1-\rho}\,.$$

with these terms the above expression (EC.30) gives the requied bound

$$\mathbb{E}[e^{\eta(f(\boldsymbol{x}_{t+1})-f(\boldsymbol{x}^\star))}|\mathcal{F}_0]\le e^{\eta(f(\boldsymbol{x}_0)-f(\boldsymbol{x}^\star))}\rho^{t+1}+D\frac{\rho^{-1}}{1-\rho}e^{\alpha\eta B}\,.$$

Applying Markov's inequality gives

$$\mathbb{P}\left(f(\boldsymbol{x}_{t+1})-f(\boldsymbol{x}^\star)\ge z\right)\le e^{-\eta z}\left\{e^{\eta(f(\boldsymbol{x}_0)-f(\boldsymbol{x}^\star))}\rho^{t+1}+D\frac{\rho^{-1}}{1-\rho}e^{\alpha\eta B}\right\}.$$

Taking $\eta=Q/\alpha$ gives the required bound. □

EC.2.6.2. Linear Convergence under Exponential Concentration We note that while we generally assume that the set $\mathcal{X}$ is bounded, the above lemma and the linear convergence results of this section apply to unbounded constraint sets. We now prove Theorem 5.

THEOREM 5. *We assume that $\mathcal{X}$ is a convex set that may be unbounded. Assume Conditions* (C1) *and* (C2) *hold for a stochastic approximation procedure with rates given in* (12)*:*
a) If, for $\hat{\epsilon}>0$ and $\hat{\delta}\in(0,1)$, we set

$$S=\log\left(\frac{F}{\hat{\epsilon}}\right),\quad \hat{\alpha}_s=\frac{2^{-s}F\kappa}{E\log\left(\frac{RS}{\hat{\delta}}\right)},\quad \textit{and}\quad t_s=\left\lceil\frac{2}{\kappa^2}\log\left(\frac{RS}{\hat{\delta}}\right)\right\rceil$$

then with probability greater than $1-\hat{\delta}$ it holds that $\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\hat{\boldsymbol{x}}_S-\boldsymbol{x}\|\le\hat{\epsilon}$. Moreover, the number of iterations (12) *required to achieve this bound is*

$$\left\lceil\log_2\left(\frac{F}{\hat{\epsilon}}\right)\right\rceil\left\lceil\frac{2}{\kappa^2}\log\left(\frac{R}{\hat{\delta}}\right)+\log\left(\left\lceil\log_2\left(\frac{F}{\hat{\epsilon}}\right)\right\rceil\right)\right\rceil.$$

(Above $F=\min_{x^\star\in\mathcal{X}^\star}\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|$ and R and E are time-independent constants that depend on the constants given in Conditions (C1) *and* (C2)*.)*

b) For $\hat{\alpha}_s=\frac{a}{2^s\log(s+1)}$ and $t_s=\log^2(s+1)$, there exists positive constants A and M such that $\forall\hat{\delta}\in(0,1)$ if $a\ge A/\hat{\delta}$ then

$$\mathbb{P}\left(\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\hat{\boldsymbol{x}}_s-\boldsymbol{x}\|\le 2^{-s}M,\quad\forall s\in\mathbb{N}\right)\ge 1-\hat{\delta}\,.$$

Proof of Theorem 5. First, we recall some notation: $f(\hat{\boldsymbol{x}}_s):=\min_{\boldsymbol{x}\in\mathcal{X}^\star}\|\hat{\boldsymbol{x}}_s-\boldsymbol{x}\|$ and $F=f(\boldsymbol{x}_0)-\min_{x\in\mathcal{X}}f(x)$. The constants D and E are the moment generating function constants as defined in Lemma 3 and Lemma 4, respectively.

We define the event $\mathcal{E}_s:=\{f(\hat{\boldsymbol{x}}_s)\le 2^{-s}F\}$. So $\mathbb{P}(\mathcal{E}_0)=1$. We inductively analyze $\mathbb{P}(\mathcal{E}_s)$. Notice

$$\mathbb{P}(\mathcal{E}_s)\ge\mathbb{P}(\mathcal{E}_s|\mathcal{E}_{s-1})\mathbb{P}(\mathcal{E}_{s-1})=\left(1-\mathbb{P}(\mathcal{E}_s^c|\mathcal{E}_{s-1})\right)\mathbb{P}\left(\mathcal{E}_{s-1}\right)\ge\mathbb{P}(\mathcal{E}_{s-1})-\mathbb{P}\left(\mathcal{E}_s^c|\mathcal{E}_{s-1}\right). \tag{EC.31}$$

By Lemma EC.6, for $\hat{x}_{s-1}$ such that $f(\hat{\boldsymbol{x}}_{s-1}) \le 2^{-s+1}F$ we have

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_s^c|\hat{\boldsymbol{x}}_{s-1}\right) = \mathbb{P}\left(f(\hat{\boldsymbol{x}}_s) \ge 2^{-s}F|\hat{\boldsymbol{x}}_{s-1}\right) &\le e^{-\frac{Q}{\hat{\alpha}_s}z}\left[e^{\frac{Q}{\hat{\alpha}_s}(f(\hat{\boldsymbol{x}}_{s-1})-f(\boldsymbol{x}^\star))}e^{-t\frac{Q}{\hat{\alpha}_s}\kappa/2} + D\frac{e^{Q\kappa/2}}{1-e^{-Q\kappa/2}}e^{QB}\right] \\ &\le e^{-\frac{Q}{\hat{\alpha}_s}z}\left[\exp\left\{\frac{Q}{\hat{\alpha}_s}2^{-s+1}F - t\frac{Q}{\hat{\alpha}_s}\kappa/2\right\} + D\frac{e^{Q\kappa/2}}{1-e^{-Q\kappa/2}}e^{QB}\right].\end{aligned}$$

(Here we apply Lemma EC.6 for times $t = T_{s-1}, ..., T_s - 1$ with expectation $\mathbb{E}[\cdot]$ given by $\mathbb{E}[\cdot|\hat{x}_{s-1}]$.)

Notice that the term in curly brackets above is negative iff $t_s \ge 2^{-s+1}F/\kappa\hat{\alpha}_s$. If this holds then

$$\mathbb{P}\left(\mathcal{E}_s^c|\mathcal{E}_{s-1}\right) \le Re^{-2^{-s}F\kappa/2E\hat{\alpha}} \qquad \text{where} \quad R := 1 + D\frac{e^{Q\kappa/2}}{1-e^{-Q\kappa/2}}e^{QB}.$$

Applying this to (EC.31), $\mathbb{P}(\mathcal{E}_s) \ge \mathbb{P}(\mathcal{E}_{s-1}) - \mathbb{P}(\mathcal{E}_s^c|\mathcal{E}_{s-1}) \ge \mathbb{P}(\mathcal{E}_{s-1}) - Re^{-2^{-s}F\kappa/2E\hat{\alpha}_s}$. So we have

$$\mathbb{P}(\mathcal{E}_S) \ge 1 - \sum_{s=1}^{S} Re^{-2^{-s}F\kappa/2E\hat{\alpha}_s}. \tag{EC.32}$$

The total number of computations/samples required is $\sum_{s=1}^{S} t_s \ge \sum_{s=1}^{S} 2^{-s}F/\kappa\hat{\alpha}_s$.

We now prove part a). Given the bounds above, we can optimize the number of samples to achieve a probability $1-\hat{\delta}$. That is we solve

$$\text{minimize} \quad \sum_{s=1}^{S}\frac{2^{-s+1}F}{\kappa\hat{\alpha}_s} \quad \text{such that} \quad \sum_{s=1}^{S} Re^{-2^{-s+1}F\kappa/2E\hat{\alpha}_s} \le \hat{\delta} \quad \text{over} \quad \hat{\alpha}_s > 0.$$

A short calculation shows that this is minimized by $\hat{\alpha}_s = 2^{-s}F\kappa/E\log(RS/\hat{\delta})$ and thus since $t_s \ge 2^{-s+1}F/\kappa\hat{\alpha}_s$ we define $t_s = \left\lceil \frac{2}{\kappa^2}\log\left(\frac{RS}{\hat{\delta}}\right)\right\rceil$ and the number of samples required here is $S \times t_s$ which equals $S\left\lceil \frac{2}{\kappa^2}\log\left(\frac{RS}{\hat{\delta}}\right)\right\rceil$. Since for an $\hat{\epsilon}$ approximation, we require S to be such that $\hat{\epsilon} \ge 2^{-S}F$, we take $S = \lceil \log_2(F/\hat{\epsilon})\rceil$. Thus we see that an $\hat{\epsilon}$ approximation can be achieved with a probability greater than $1-\hat{\delta}$ in a number of samples given by

$$\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right)\right\rceil\left\lceil \frac{2}{\kappa^2}\log\left(\frac{R}{\hat{\delta}}\right) + \log\left(\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right)\right\rceil\right)\right\rceil.$$

This gives the part a) of Theorem 5.

Notice, we can make the sum (EC.32) finite for $S = \infty$. Specifically if we take $\hat{\alpha}_s = a/2^s\log(s+1)$ and $t_s = (\log(s+1))^2$ then the Condition (C1) holds $\forall s \ge s_0$ for $s_0 = \lceil e^{2F/\kappa}\rceil + 1$ and thus

$$\sum_{s=s_0}^{\infty} Re^{-2^{-s+1}F\kappa/2E\hat{\alpha}_s} = \sum_{s=s_0}^{\infty} R\frac{1}{(s+1)^{aF\kappa/2E}} \le R\int_{s_0}^{\infty}\frac{1}{s^{aF\kappa/2E}}ds \le \frac{2RE}{aF\kappa}\cdot\frac{1}{s_0^{aF\kappa/2E}} \le \frac{2RE}{aF\kappa}.$$

The above sum is less than $\hat{\delta}$ for $a \ge RE/2\hat{\delta}F\kappa$. Letting $A = 2RE/F\kappa$ and $M = 2^{s_0}F$, we see that for $a \ge A/\hat{\kappa}$ gives

$$\mathbb{P}\left(\exists s \in \mathbb{N} \text{ s.t. } \min_{x\in\mathcal{X}^\star}\|x_s - x\| \ge 2^{-s}F\right) \le \sum_{s=1}^{\infty}\mathbb{P}\left(\mathcal{E}_s^c \cup \left(\cap_{s'\le s}\mathcal{E}_{s'}\right)\right) \le \sum_{s=1}^{\infty}\mathbb{P}\left(\mathcal{E}_s^c|\mathcal{E}_{s-1}\right) \le \sum_{s=1}^{\infty} Re^{-\frac{2^s F\kappa}{2E\hat{\alpha}}} \le \hat{\delta}.$$

Thus for learning rates $\hat{\alpha}_s = a/2^s\log(s+1)$ with $a \ge M/\hat{\delta}$ if it holds that $\mathbb{P}(\forall s, \min_{x\in\mathcal{X}^\star}\|x_s - x\| \le 2^{-s}M) \ge 1-\hat{\delta}$. This gives the 2nd part of Theorem 5. □

EC.2.6.3. Application to Specific Stochastic Approximation Algorithms. The following is the equivalent linear convergence result for Kiefer-Wolfowotiz

COROLLARY EC.1 **(Linear Convergence in Projected Stochastic Gradient Descent)**. *We assume that $\mathcal{X}$ is a convex set that may be unbounded. Assume Conditions* (D1) *and* (D2) *hold for PSGD with rates given in* (12)*. If, for $\hat{\epsilon} > 0$ and $\hat{\delta} \in (0,1)$, we set*

$$S = \log\left(\frac{F}{\hat{\epsilon}}\right), \quad \hat{\alpha}_s = \frac{2^{-s} F \kappa}{E \log\left(\frac{GS}{\hat{\delta}}\right)}, \quad \textit{and} \quad t_s = \left\lceil \frac{2}{\kappa^2} \log\left(\frac{GS}{\hat{\delta}}\right) \right\rceil$$

then with probability greater than $1-\hat{\delta}$ it holds that $\min_{\boldsymbol{x}\in\mathcal{X}^\star} \|\hat{\boldsymbol{x}}_S - \boldsymbol{x}\| \le \hat{\epsilon}$. Moreover, the number of iterations (12) *required to achieve this bound is*

$$\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right) \right\rceil \left\lceil \frac{2}{\kappa^2} \log\left(\frac{G}{\hat{\delta}}\right) + \log\left(\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right)\right\rceil\right) \right\rceil .$$

COROLLARY EC.2 **(Linear Convergence of Kiefer-Wolfowitz)**. *Assume Conditions* (D1), (D2), (D3) *hold. For the KW algorithm,* (9)*, with step-sizes given in* (12)*: If, for $\hat{\epsilon} > 0$ and $\hat{\delta} \in (0,1)$, we set*

$$S = \log\left(\frac{F}{\hat{\epsilon}}\right), \quad \hat{\alpha}_s = \frac{2^{-s} F \kappa}{E \log\left(\frac{GS}{\hat{\delta}}\right)}, \quad \nu_s = \sqrt{\frac{\kappa}{3c}} \quad \textit{and} \quad t_s = \left\lceil \frac{2}{\kappa^2} \log\left(\frac{GS}{\hat{\delta}}\right) \right\rceil$$

then with probability greater than $1-\hat{\delta}$ it holds that $\min_{\boldsymbol{x}\in\mathcal{X}^\star} \|\hat{\boldsymbol{x}}_S - \boldsymbol{x}\| \le \hat{\epsilon}$. Moreover, the number of iterations (12) *required to achieve this bound is*

$$\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right) \right\rceil \left\lceil \frac{2}{\kappa^2} \log\left(\frac{G}{\hat{\delta}}\right) + \log\left(\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right)\right\rceil\right) \right\rceil .$$

The proof of Corollary EC.2 is identical to the proof of Theorem 5.

COROLLARY EC.3 **(Linear Convergence of Stochastic Frank-Wolfe)**. *We assume that $\mathcal{X}$ is a convex set that may be unbounded. Also, Assume Condition D1, D2, E1 and E2 hold. For the SFW algorithm with step-sizes given by* (12)*. If we set*

$$S = \log\left(\frac{F}{\hat{\epsilon}}\right), \quad \hat{\alpha}_s = \frac{2^{-s} F \kappa}{E \log\left(\frac{GS}{\hat{\delta}}\right)}, \quad m_s := \left\lceil \left(\frac{3\sigma}{\kappa\alpha}\right)^2 \right\rceil, \quad \textit{and} \quad t_s = \left\lceil \frac{2}{\kappa^2} \log\left(\frac{GS}{\hat{\delta}}\right) \right\rceil$$

then with probability greater than $1-\hat{\delta}$ it holds that $\min_{\boldsymbol{x}\in\mathcal{X}^\star} \|\hat{\boldsymbol{x}}_S - \boldsymbol{x}\| \le \hat{\epsilon}$. Moreover, the number of iterations (12) *required to achieve this bound is*

$$\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right) \right\rceil \left\lceil \frac{2}{\kappa^2} \log\left(\frac{G}{\hat{\delta}}\right) + \log\left(\left\lceil \log_2\left(\frac{F}{\hat{\epsilon}}\right)\right\rceil\right) \right\rceil .$$

EC.3. Appendix to Theoretical Results

EC.3.1. List of Notations for Theorem 1

There are several time-independent constants (usually denoted with a capital letter) in Theorem 1. We list these here.

B	Bound where the drift condition holds. See (C1).
$C = \dfrac{(1+H)}{2QG^n} + B$	Constant in Proposition 1
$D = \mathbb{E}[e^{\lambda Z}] = \mathbb{E}[e^{\lambda[Y+\kappa/2]}]$	Moment Generating Function, see (18).
$E = \mathbb{E}\left[\dfrac{e^{\lambda Z} - 1 - \lambda Z}{\lambda^2}\right]$	See Lemma 4
$F = \max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) - \min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$	
$G = \dfrac{1}{4^\gamma}, \quad \text{for } \alpha_t = \dfrac{a}{(u+t)^\gamma}$	See Lemma 5 and Proof in Section 4.1.2
$H = De^{\frac{\kappa QG^n}{2}}/(1 - e^{-\frac{\kappa QG^n}{2}})$	Constant in Proposition 1 Defined after (26).
$I = (1+H)e^{QG/(F/\alpha_{T_2} - B)}$	Constant in Theorem 1
$J = QG^n$	Exponent in Theorem 1
$K = \dfrac{I}{J}$	Constant in Theorem 1
$n = \begin{cases} 1 & \text{for } \gamma < 1 \\ \lceil \frac{\alpha_0 B + F}{a \log 2} \rceil & \text{for } \gamma = 1 \end{cases}$	
$Q = \lambda \wedge (\kappa/2E)$	
$T_0 = \min\left\{t \geq 0 : \dfrac{(\alpha_s - \alpha_{s+1})}{\alpha_s} < \kappa/2B, \forall s \geq t\right\}$	See Lemma 3 and its proof.
$T_1 = \begin{cases} \frac{\alpha_0 B + F}{\frac{a}{1-\gamma}\left[1 - \frac{1}{2^{1-\gamma}}\right]} & \text{for } \gamma < 1 \\ u2^n & \text{for } \gamma = 1 \end{cases}$	
$T_2 = T_0 \vee T_1$	
Y	Sub-exponential Random Variable defined in C2
$Z = Y + \dfrac{\kappa}{2}$	Random Variable defined in Lemma 3

EC.3.2. Technical Lemmas for the Proof of Proposition 1

LEMMA 3. *Given Conditions* (C1) *and* (C2) *hold, there exists a deterministic constant* T_0 *such that the sequence of random variables* $(L_t : t \geq T_0)$ *satisfies*

$$\mathbb{E}\big[L_{t+1} - L_t \big| \mathcal{F}_t\big]\mathbb{I}[L_t \geq 0] < -\alpha_t \kappa, \tag{17}$$

and

$$[|L_{t+1} - L_t||\mathcal{F}_t] \leq \alpha_t Z \quad \textit{where} \quad D := \mathbb{E}[e^{\lambda Z}] < \infty\,. \tag{18}$$

Proof of Lemma 3. Applying the definition of L_t and the drift Condition (C1) gives

$$\begin{aligned}\mathbb{E}[L_{t+1} - L_t|\mathcal{F}_t] &= \mathbb{E}[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)|\mathcal{F}_t] + (\alpha_t - \alpha_{t+1})B \\ &\leq -2\alpha_t\kappa + (\alpha_t - \alpha_{t+1})B \\ &\leq -2\alpha_t\kappa\left[1 - (\alpha_t - \alpha_{t+1})B/\alpha_t\kappa\right]\end{aligned}$$

Since $(\alpha_t - \alpha_{t+1})/\alpha_t \to 0$, there exists a constant T_0 such that $(\alpha_t - \alpha_{t+1})/\alpha_t < \kappa/2B$ for all $t \geq T_0$. Specifically we can take $T_0 = \min\{t \geq 0 : (\alpha_s - \alpha_{s+1})/\alpha_s < \kappa/2B, \forall s \geq t\}$. This gives the first drift condition (17).

For the second condition, for $t \geq T_0$ with T_0 as just defined:

$$\begin{aligned}\left[|L_{t+1} - L_t|\middle|\mathcal{F}_t\right] &\leq \left[|f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)|\middle|\mathcal{F}_t\right] + |\alpha_{t+1} - \alpha_t|B \\ &\leq \alpha_t Y + \alpha_t\frac{(\alpha_t - \alpha_{t+1})}{\alpha_t}B \\ &\leq \alpha_t(Y + \kappa/2)\,.\end{aligned}$$

Taking $Z = Y + \kappa/2$, it is clear that condition (18) holds for Z as an immediate consequence of the boundedness condition on Y in (C2). □

Lemma 5. *If α_t, $t \in \mathbb{Z}_+$, is a decreasing positive sequence, then*

$$\min_{s=\hat{t},\ldots,t}\left\{\frac{\sum_{k=s}^t \alpha_k}{\sum_{k=s}^t \alpha_k^2}\right\} = \frac{\sum_{k=\hat{t}}^t \alpha_k}{\sum_{k=\hat{t}}^t \alpha_k^2}\,. \tag{21}$$

Moreover, if α_t, $t \in \mathbb{Z}_+$ satisfies the learning rate condition (13) *then*

$$\frac{1}{\alpha_{\lfloor t/2^n \rfloor}} \geq \frac{G^n}{\alpha_t} \qquad \textit{and} \qquad \min_{s=\lfloor t/2^n \rfloor,\ldots,t}\left\{\frac{\sum_{k=s}^t \alpha_k}{\sum_{k=s}^t \alpha_k^2}\right\} \geq \frac{G^n}{\alpha_t} \tag{22}$$

for some constant $G \in (0,1]$ and for $n \in \mathbb{N}$ such that $t/2^n > 1$.

Proof of Lemma 5. It is straight-forward to show that for $a, a', A, A' > 0$

$$\frac{a}{a'} \leq \frac{A}{A'} \quad \text{if and only if} \quad \frac{a+A}{a'+A'} \leq \frac{A}{A'}\,. \tag{EC.33}$$

[Note that both expressions above are equivalent to $AA' + aA' \leq AA' + a'A$.]

Take positive numbers a_s, a'_s, $s = 1, ..., t$. If

$$\frac{a_s}{a'_s} \leq \frac{a_k}{a'_k}$$

for $k = s+1, ..., t$, then

$$\sum_{k=s+1}^t a'_k a_s \leq \sum_{k=s+1}^t a_k a'_s. \tag{EC.34}$$

Thus,

$$\frac{a_s}{a'_s} \leq \frac{\sum_{k=s+1}^{t} a_k}{\sum_{k=s+1}^{t} a'_k}.$$

Thus applying (EC.33) with $A = \sum_{k=s+1}^{t} a_k$ and $A' = \sum_{k=s+1}^{t} a'_k$ gives

$$\frac{\sum_{k=s}^{t} a_k}{\sum_{k=s}^{t} a'_k} \leq \frac{\sum_{k=s+1}^{t} a_k}{\sum_{k=s+1}^{t} a'_k}. \tag{EC.35}$$

Finally, taking $a_k = \alpha_k$ and $a'_k = \alpha_k^2$, we see that (EC.34) holds since α_t is decreasing. Thus, from (EC.35), we see that the result (21) holds.

If the condition (13) holds then $\liminf_{t\to\infty} \alpha_{2t}/\alpha_t > 0$ implies

$$\frac{\alpha_t}{\alpha_{\lfloor t/2 \rfloor}} > \sqrt{G} \tag{EC.36}$$

for some $1 \geq G > 0$. Thus

$$\frac{\alpha_t}{\alpha_{\lfloor t/2^n \rfloor}} = \frac{\alpha_t}{\alpha_{\lfloor t/2 \rfloor}} \times ... \times \frac{\alpha_{\lfloor t/2^{n-1} \rfloor}}{\alpha_{\lfloor t/2^n \rfloor}} \geq G^{\frac{n}{2}} \geq G^n. \tag{EC.37}$$

Since the sequence is decreasing and (EC.36) holds, we have that

$$\frac{\sum_{k=\lfloor t/2^n \rfloor}^{t} \alpha_k}{\sum_{k=\lfloor t/2^n \rfloor}^{t} \alpha_k^2} \geq \frac{(t - \lfloor t/2^n \rfloor)\alpha_t}{(t - \lfloor t/2^n \rfloor)\alpha_{\lfloor t/2^n \rfloor}^2} = \frac{\alpha_t^2}{\alpha_{\lfloor t/2^n \rfloor}^2} \frac{1}{\alpha_t} = \frac{\alpha_t^2}{\alpha_{\lfloor t/2 \rfloor}^2} \times ... \times \frac{\alpha_{\lfloor t/2^{n-1} \rfloor}^2}{\alpha_{\lfloor t/2^n \rfloor}^2} \frac{1}{\alpha_t} \geq \frac{G^n}{\alpha_t}.$$

Applying this to (21) with $s = \lfloor t/2^n \rfloor$ gives

$$\min_{s=\lfloor t/2^n \rfloor, \ldots, t} \left\{ \frac{\sum_{k=s}^{t} \alpha_k}{\sum_{k=s}^{t} \alpha_k^2} \right\} \geq \frac{G^n}{\alpha_t}.$$

Thus the above along with (EC.37) proves that (22) holds as required. □

LEMMA EC.7. *For $\alpha_t = a/(u+t)^\gamma$ with $0 \leq \gamma < 1$ Taking*

$$n = \begin{cases} 1, & \text{for } \gamma < 1, \\ 1 + \left\lceil \frac{\alpha_0 B + F}{a \log 2} \right\rceil, & \text{for } \gamma = 1, \end{cases} \qquad \text{and} \qquad T_1 = \begin{cases} u + \frac{2^{1+\gamma}}{a u^{-\gamma}} [\alpha_0 B + F], & \text{for } \gamma < 1, \\ u 2^n, & \text{for } \gamma = 1, \end{cases}$$

it holds that

$$\sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \geq \alpha_0 B + F, \qquad \forall t \geq T_1. \tag{EC.38}$$

Proof. We consider the case of $\gamma < 1$ separately from the case where $\gamma = 1$.

First we take $\gamma < 1$ and $n = 1$. In the following expression, we take $t = xu$ with $x \geq 1$,

$$\sum_{s=\lfloor t/2 \rfloor}^{t} \alpha_s \geq \frac{t}{2} \alpha_t = \frac{a}{2} \frac{t}{(u+t)^\gamma} = \frac{a}{2} u^{1-\gamma} \frac{x}{(1+x)^\gamma} \geq \frac{a u^{1-\gamma}}{2^{1+\gamma}} x = a \frac{u^{-\gamma}}{2^{1+\gamma}} t \tag{EC.39}$$

Thus, $t = ux$ with $x \geq 1$ and right-hand side of (EC.39) is greater than $\alpha_0 B + F$ for

$$T_1 = \frac{2^{1+\gamma}}{a u^{-\gamma}} [\alpha_0 B + F] + u,$$

and for any t such that $t \geq T_1$. This completes the proof for $\gamma < 1$

Second, we take $\gamma = 1$. We assume that $t \geq T_1 := 2^n u$ and we will take $n = 1 + \left\lceil \frac{\alpha_0 B + F}{a \log 2} \right\rceil$.

$$\sum_{s=\lfloor t/2^n \rfloor}^{t} \alpha_s \geq \int_{t/2^n}^{t} \frac{a}{u+s} ds = a \log\left(\frac{u+t}{u+t2^{-n}}\right) = an \log 2 + a \log\left(\frac{u+t}{u2^n + tk}\right) \geq an \log 2 + a \log \frac{1}{2}.$$

The last inequality above holds since $t \geq T_1 := 2^n u$. Notice that

$$an \log 2 + a \log \frac{1}{2} = a(n-1) \log 2 \geq \alpha_0 B + F, \qquad \text{for } n = 1 + \left\lceil \frac{\alpha_0 B + F}{a \log 2} \right\rceil.$$

Thus the required bound (EC.38) holds for n and T_1 as specified for $\gamma = 1$. □

EC.4. Appendix to Applications and Numerical Examples

This section aims to provide a simple application of the main results of Theorem 1 and Theorem 2. Given the importance of Linear Programming (LP) and Markov Decision Processes (MDP) in operations research, we briefly explore these problem settings. However, we emphasize that linear objectives are a special case of the results proven in Theorem 1 and Theorem 2. The results are proved under conditions that apply to non-smooth, non-convex objectives and general convex constraints. We refer to Birge and Louveaux (2011) and Shapiro et al. (2021) as standard texts on stochastic linear programming. For the linear programming formulation of MDPs, we refer to Schweitzer and Seidmann (1985).

EC.4.1. Linear Programming

Here we consider a linear program in which the cost function that we wish to minimize must be sampled and where the optimization constraints are deterministic. We are interested in solving a linear program of the form

$$\text{minimize } \bar{\boldsymbol{c}}^\top \boldsymbol{x} \;\; \text{subject to } H\boldsymbol{x} \leq \boldsymbol{b} \text{ over } \boldsymbol{x} \in \mathbb{R}^d, \tag{EC.40}$$

where $\bar{\boldsymbol{c}} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$, $H \in \mathbb{R}^{p \times d}$ and $\boldsymbol{b} \in \mathbb{R}^p$. We assume $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d : H\boldsymbol{x} \leq \boldsymbol{b}\}$ is a bounded polytope.

We suppose that the constraint set $\mathcal{X}$ is deterministic and known, however, the cost vector $\bar{\boldsymbol{c}}$ is unknown but can be sampled.

Specifically, we let $\boldsymbol{c}_t$, $t \in \mathbb{Z}_+$, be an independent, mean $\bar{\boldsymbol{c}}$, sub-exponential random vectors in $\mathbb{R}^d$. That is

$$\mathbb{E}[\boldsymbol{c}_t | \mathcal{F}_t] = \bar{\boldsymbol{c}} \qquad \text{and} \qquad \sup_{t \in \mathbb{Z}_+} \mathbb{E}\left[e^{\lambda \|\boldsymbol{c}_t\|} \middle| \mathcal{F}_t\right] < \infty \tag{EC.41}$$

for some $\lambda > 0$. We then apply projected stochastic gradient descent (6). Notice that Condition (D1) is satisfied by (EC.41). Further Condition (D2) holds for any linear program. This is a consequence of the following technical lemma.

LEMMA EC.8. *If $\mathcal{X}$ is a bounded polytope and $\mathcal{X}^\star = argmin_{\boldsymbol{x}\in\mathcal{X}} \bar{\boldsymbol{c}}^\top \boldsymbol{x}$., then there exists a constant $K > 0$ such that*

$$\frac{\bar{\boldsymbol{c}}^\top(\boldsymbol{x}-\boldsymbol{x}^\star)}{||\bar{\boldsymbol{c}}||||\boldsymbol{x}-\boldsymbol{x}^\star||} \geq K,$$

for $\boldsymbol{x}^\star$ the projection of $\boldsymbol{x}$ onto $\mathcal{X}^\star$. Thus Condition (D2) *holds for PSGD applied to the LP* (EC.40).

The proof of Lemma EC.8 requires some careful bounding between optimal solutions and sub-optimal extreme points. The proof is given below. The result bounds the angle between optimal and sub-optimal points for a polytope.

Proof of Lemma EC.8. We assume without loss of generality that $\bar{\boldsymbol{c}}^\top \boldsymbol{x}^\star = 0$ and $||\bar{\boldsymbol{c}}|| = 1$. Let $\mathcal{E}$ be the extreme points of $\mathcal{X}$. Let $\mathcal{E}^\star$ be the extreme points in $\mathcal{X}^\star$. Then let $\mathcal{E}' := \mathcal{E} \setminus \mathcal{E}^\star$ and $\mathcal{X}'$ is the convex closure of $\mathcal{E}'$. Let $a := \min_{\boldsymbol{x}\in\mathcal{X}'} \bar{\boldsymbol{c}}^\top \boldsymbol{x}$ and $D := \max_{\boldsymbol{x}^\star\in\mathcal{X}^\star, \boldsymbol{x}'\in\mathcal{X}'} ||\boldsymbol{x}^\star - \boldsymbol{x}'||$. We will show we can take $K := a/D$.

For all $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}^\star$, $\boldsymbol{x}$ must be a convex combination of a point in $\mathcal{X}^\star$ and a point in $\mathcal{X}'$. Specifically,

$$\boldsymbol{x} = (1-p)\boldsymbol{x}_0 + p\boldsymbol{x}_1, \tag{EC.42}$$

for $\boldsymbol{x}_0 \in \mathcal{X}^\star$ and $\boldsymbol{x}_1 \in \mathcal{X}'$ and $p \in (0,1]$. Then, as required,

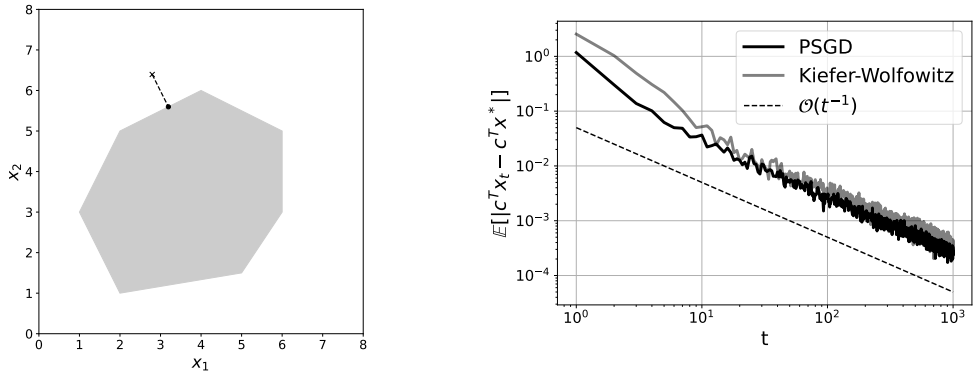
$$\frac{\bar{\boldsymbol{c}}^\top \boldsymbol{x}}{||\boldsymbol{x}-\boldsymbol{x}^\star||} \geq \frac{\bar{\boldsymbol{c}}^\top(\boldsymbol{x}-\boldsymbol{x}_0)}{||\boldsymbol{x}-\boldsymbol{x}_0||} = \frac{\bar{\boldsymbol{c}}^\top(\boldsymbol{x}_1-\boldsymbol{x}_0)}{||\boldsymbol{x}_1-\boldsymbol{x}_0||} \geq \frac{a}{D} = K > 0\,.$$

The first inequality above uses the fact that $\boldsymbol{x}^\star$ is closest to $\boldsymbol{x}$. The equality applies (EC.42). Then finally, we apply the definitions of a, D and K. □

Thus, we see that both Theorem 2 and Theorem 5 hold in the context of linear programming problems with an unknown objective function.

EC.4.1.1. Polytope Example We consider the problem with two variables with the constraints being the polytope in Figure EC.2(a). We assume that the cost vector $\bar{\boldsymbol{c}} = [4,6]^\top$ is unknown but can be sampled from a joint Gaussian distribution of independent random variables with mean vector $\bar{\boldsymbol{c}}$ and variance 1. This problem is analytically tractable. Given the costs, we can calculate the reference solution to be $x^\star = [2,1]^T$.

The convergence rate for the PSGD and Kiefer-Wolfowitz should be $O(1/t^\gamma)$ in expectation when the error is measured by the L^1-norm. Evidence for the convergence rate is shown in Figure EC.2(b). Increasing the batch size above 50 substantially reduces the noise of sampling costs, and the algorithm may perform better than $O(1/t)$. In this case, the algorithm converges, reaching the optimum solution after 7 iterations. This occurs because the chance of observing any sample perturbing the stochastic gradient descent algorithm away from the optimal point is a rare event. However, when there is a non-negligible probability of an iteration leaving the optimal point, then the $O(1/t)$ is found as anticipated.



(a) Polytope (b) Convergence for batch size $B = 5$

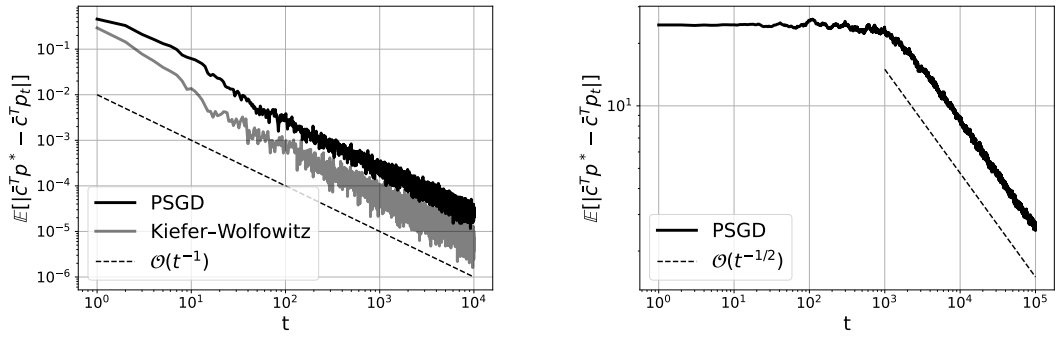
Figure EC.2 Polytope of the two variables linear programming problem and convergence of projected stochastic gradient descent on the two variables linear programming example. In Figure EC.2(a), the shaded area is the bounded polytope and the cross is one of the points of iterations, and the black point is the corresponding projection. In Figure EC.2(b), expectation is computed over 1000 realizations. The parameters of step size are chosen as $a = 1, u = 1$ and $\gamma = 1$ such that $\alpha_t = 1/(1+t)$. The costs $\boldsymbol{c}_t$ are computed with batch size $B = 5$. The parameter $v = 1$ is chosen for Kiefer-Wolfowitz. The fitted slope is -1.02 and -1.05 for PSGD and Kiefer-Wolfowitz.

EC.4.1.2. Probability Simplex This section considers a higher dimension for the optimization over the probability simplex as an example. There are simple, efficient algorithms for projection onto the probability simplex (Duchi et al. 2008). The problem that we solve is formulated as follows

$$\text{minimize} \quad p_1\bar{c}_1 + p_2\bar{c}_2 + ... + p_n\bar{c}_n = \bar{\boldsymbol{c}}^{\boldsymbol{T}}\boldsymbol{p} \quad \text{subject to} \quad \sum_{i=1}^{n} p_i = 1 \quad \text{over} \quad p_i \geq 0,\ \forall i = 1, ..., n,$$

where $\bar{c}_1 < \bar{c}_2 < ... < \bar{c}_n$ and $n = 50$. We label the polytope due to the constraint as $\mathcal{P}$ and suppose that the cost vector $\bar{\mathbf{c}}$ is unknown but can be sampled from a normal distribution with a certain mean vector and covariance matrix. In particular, for $t \in \mathbb{Z}_+$, we apply the stochastic gradient descent iteration: $\boldsymbol{p}_{t+1} = \Pi_{\mathcal{P}}(\boldsymbol{p}_t - \alpha_t \boldsymbol{c_t})$, where $\mathbf{c_t} \sim \mathcal{N}(\bar{\mathbf{c}}, \mathbf{1})$ and $\alpha_t = a/t$ with $a > 0$. According to the special settings above, the minimum of this problem is $\boldsymbol{p}^* = (1, 0, ..., 0)$. We expect that $\mathbb{E}\left[|\bar{\boldsymbol{c}}^{\boldsymbol{T}}\boldsymbol{p_t} - \bar{\boldsymbol{c}}^{\boldsymbol{T}}\boldsymbol{p}^*|\right] = O\left(1/t\right)$. Figure EC.3(a) confirms that the PSGD and Kiefer-Wolfowitz converge with an order of -1.

However, a itltohught falls somewhat outside the scope of this paper's results, it is also possible to consider the multi-armed bandit variation of this problem. Here, the natural generalization of the projected gradient descent algorithm applies importance sampling. Here we sample an index i_t according to the distribution $\boldsymbol{p}_t$ and apply the updated $p_{i,t+1} = p_{i,t} - \alpha_t \frac{c_{i,t}}{p_{i,t}} \mathbb{I}[i = i_t]$ for $i = 1, ..., n$. Simulations suggest a rate of convergence of the order of $O(1/\sqrt{t})$, see Figure EC.3(b).



(a) Probability simplex example (b) Multi-arm bandit problem

Figure EC.3 Convergence of projected stochastic gradient descent on the probability simplex example and multi-arm bandit problem. Expectations are computed over 100 realizations. The parameter $v = 1$ is chosen for Kiefer-Wolfowitz. The step size parameter is chosen as $a = 1$ such that $\alpha_t = 1/t$ for both. In Figure EC.3(a), the fitted slope is -1.01 and -1.01 for PSGD and Kiefer-Wolfowitz. In Figure EC.3(b), the fitted slope is -0.475.

EC.4.2. Markov Decision Processes

We now optimize a discounted Markov Decision Process (MDP) using the results from the last section. Here we use a linear programming approach to give the convergence of a simple policy gradient algorithm for an MDP in which the system dynamics are known but the costs are unknown.

An MDP can be formulated as a linear program, where the primal form of this linear program solves for the optimal value function, and the dual form finds the optimal occupancy measure. In this linear programming formulation, the dual problem takes the form:

$$\begin{aligned} \text{minimize} \quad & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \bar{c}(s,a)x(s,a) && \text{(Dual)} \\ \text{subject to} \quad & \sum_{a\in\mathcal{A}} x(s',a) = \xi(s') + \beta \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x(s,a)P(s'|s,a), \qquad \forall s' \in \mathcal{S} \\ \text{over} \quad & (x(s,a) : s\in\mathcal{S}, a\in\mathcal{A}) \in \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}. \end{aligned}$$

Here $(\xi(s) : s \in \mathcal{S})$ is a positive vector. We assume that the dynamics as given by $(P(s'|s,a) : a \in \mathcal{A},\ s, s' \in \mathcal{S})$ are known but costs $(\bar{c}(s,a) : a \in \mathcal{A}, s \in \mathcal{S})$ are unknown and must be sampled, then above we have a linear program with an unknown objective and known constraints. For this reason, we can apply the analysis developed in the last section.

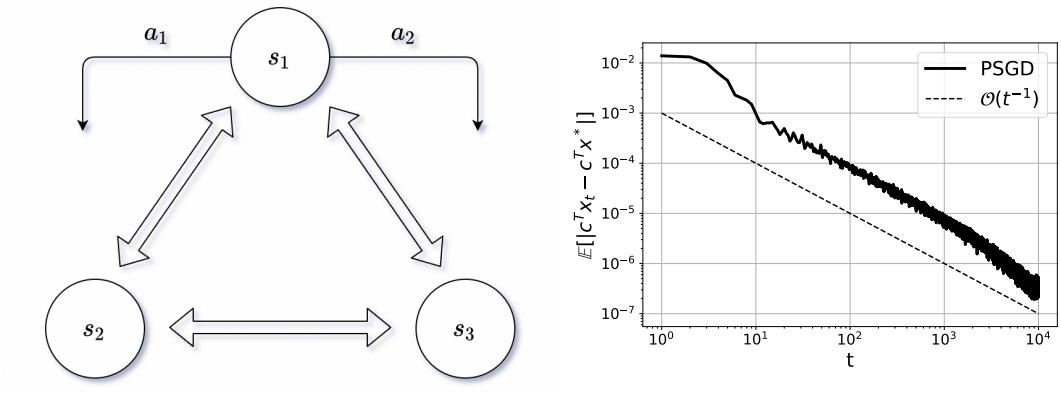
Here we assume that we can sample costs $\hat{\boldsymbol{c}} = (\hat{c}(s,a) : s \in \mathcal{S}, a \in \mathcal{A})$ where the states and actions are distributed according to some predetermined probability distribution $\boldsymbol{\pi} = (\pi(s,a) : a \in \mathcal{A}, s \in \mathcal{S})$. There are several ways of sampling the cost vector $\boldsymbol{c}_t$ for each t. The most straightforward one is as follows. For each t, the cost $\boldsymbol{c}_t = (c_t(s,a) : s \in \mathcal{S},\ a \in \mathcal{A})$ is sampled by first taking IID sample

(s_t, a_t), with distribution $\boldsymbol{\pi} = (\pi(s,a) : s \in \mathcal{S}, a \in \mathcal{A})$ where $\pi(s,a) > 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and then defining

$$c_t(s,a) = \frac{\hat{c}(s_t, a_t)}{\pi(s_t, a_t)} \mathbb{I}[(s_t, a_t) = (s,a)]. \tag{EC.43}$$

We allow for the possibility of averaging batches of costs of the form (EC.43). We then consider the projected gradient descent algorithm $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}\left(\boldsymbol{x}_t - \alpha_t \boldsymbol{c}_t\right)$. The projection above is onto the constraint set of the dual problem (Dual). Our above observation on Linear Programs holds here. Specifically, Theorem 2 and Theorem 5 hold for this PSGD algorithm.

EC.4.2.1. Three-state two-action Markov decision process We now consider the first reinforcement learning application of our results, a relatively simple MDP. We consider an MDP with three states $\mathcal{S} = \{s_1, s_2, s_3\}$. In each state, there are two actions $\mathcal{A} = \{a_1, a_2\}$, corresponding to move anticlockwise (a_1) and clockwise (a_2). Figure EC.4(a) shows the states and actions. When we choose to take an action, the probability of going to the desired state is 2/3; otherwise, one of the states uniformly at random. We assume that the costs $c(s,a)$ are independent normally distributed with $c(s_i, a_j) \sim N(i, 1)$, for $i = 1, 2, 3$. The states and actions are sampled according to the predetermined probability distribution $\boldsymbol{\pi} = (\pi(s,a) = 1/6 : s \in \mathcal{S}, a \in \mathcal{A})$. Figure EC.4(b) demonstrates the correct convergence rate as predicted.



(a) Three-state two-action MDP graph (b) Three-state two-action MDP example

Figure EC.4 The three-state two-action MDP graph and convergence of projected stochastic gradient descent on the three-state two-action MDP example. In Figure EC.4(b), the expectation is computed over 20 realizations. The costs $c_t(s,a)$ are computed with batch size $B = 200$. The parameters of step size are chosen as $a = 0.1, u = 1$ and $\gamma = 1$ such that $\alpha_t = 0.1/(1+t)$. The fitted slope is -1.29.

EC.4.2.2. Blackjack We now consider a larger tabular reinforcement learning problem for the game of Blackjack. Blackjack is a simple card game where a player is initially dealt two cards. The player is dealt cards sequentially before deciding to stop. The player must attempt to reach a total that is more than the dealer but not more than 21. The problem is described in more detail in Sutton and Barto (2018). The states of the problem depend on three factors which are: the player's current points (4–22); usable ace (with or without); dealer's showing card (1–10), which gives 290 states in total.

We label the states in sequence starting with s_1 being no usable ace, the player's current points 4 and dealer's showing card 1, and ending with s_{290} being usable ace, the player's current points 21 and dealer's showing card 10. The actions simply consist of hitting (a_1) and sticking (a_0). Denote the collection of states $\mathcal{S} = \{s_i : i = 1, ..., 290\}$ and the collection of actions $\mathcal{A} = \{a_i : i = 0, 1\}$.

We assume that the reward $\bar{\boldsymbol{r}} = (\bar{r}(s,a) : s \in \mathcal{S}, a \in \mathcal{A})$ can be sampled for each iteration of the projected stochastic gradient descent by carrying on the following procedure. We first simulate IID samples $(s_t^i, a_t^i),\ i = 1, ..., B$ from the distribution $\boldsymbol{\pi} = (\pi(s,a) = 1/580 : s \in \mathcal{A}, a \in \mathcal{A})$ and then define the cost similar as Equation (EC.43) with $\bar{c}(s_t^i, a_t^i) = -\bar{r}(s_t^i, a_t^i)$. In addition, according to the rules, it is reasonable to set the discount factor $\beta = 1$. Applying the PSGD with the learning rate of the form $\alpha_t = a/(b+t)^\gamma$, for $a, b > 0$ and $\gamma \in [0, 1]$, the projected stochastic gradient descent converges with a rate of $O(1/t^\gamma)$ in expectation. The rate $O(1/t)$ with $\gamma = 1$ is shown in Figure EC.5.

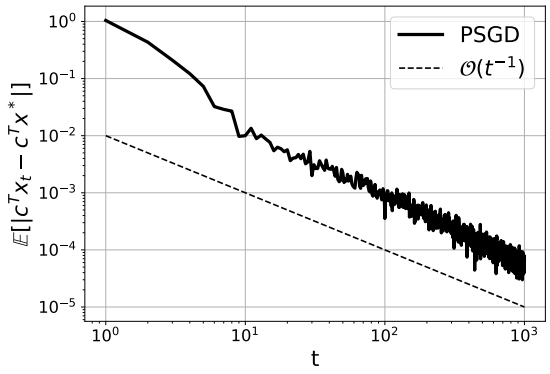


Figure EC.5 Convergence of projected stochastic gradient descent on the Blackjack example. The expectation is computed over 10 realizations. The costs $c_t(s,a)$ are computed with batch size $B = 200$. The parameters of step size are chosen as $a = 0.1, u = 1$ and $\gamma = 1$ such that $\alpha_t = 0.1/(1+t)$. The fitted slope is -1.23.