Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

A SMOOTHED AUGMENTED LAGRANGIAN FRAMEWORK FOR CONVEX OPTIMIZATION WITH NONSMOOTH STOCHASTIC CONSTRAINTS

Peixuan Zhang[1] and Uday V. Shanbhag[2]

[1]Dept. of Industrial and Manufacturing Eng., Pennsylvania State University, State College, PA, USA
[2]Dept. of Industrial and Operations Eng., University of Michigan, Ann Arbor, MI, USA

ABSTRACT

We consider a convex stochastic optimization problem where both the objective and constraints are convex but possibly complicated by uncertainty and nonsmoothness. We present a smoothed sampling-enabled augmented Lagrangian (AL) framework that relies on inexact solutions to the AL subproblem, obtainable via a stochastic approximation framework. Under a constant penalty parameter, the dual suboptimality is shown to diminish at a sublinear rate while primal infeasibility and suboptimality both diminish at a slower sublinear rate.

1 INTRODUCTION

Consider the convex optimization problem with possibly nonsmooth expectation-valued constraints.

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x},\boldsymbol{\xi})] \,|\, g_i(\mathbf{x}) \triangleq \mathbb{E}\left[\tilde{g}_i(\mathbf{x},\boldsymbol{\xi})\right] \leq 0, i = 1,\cdots,m, \right\} \tag{NSCopt}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed and convex, $\boldsymbol{\xi} : \Omega \to \mathbb{R}^d$ is a d-dimensional random variable, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the probability space, $\Xi \triangleq \{\, \xi(\omega) \mid \omega \in \Omega \,\}$, and for any $\xi \in \Xi$, $\tilde{f}(\bullet, \xi)$ and $\tilde{g}_i(\bullet, \xi)$ are real-valued possibly nonsmooth (but smoothable (see Def. 1)) convex functions on $\mathcal{X}$ for $i = 1, \cdots, m$. A host of applications in engineering, economics and statistics can be formulated as (NSCopt); these include optimization problems with risk constraints as well as a range of problems in statistical learning including Lasso regression and Neyman-Pearson classification. In general, projection-based approaches cannot contend with such avenues; inspired by the success of augmented Lagrangian (AL) techniques for constrained optimization problems, we consider the development of a sampling-enabled AL framework.

1.1 Prior Research

Before proceeding, we briefly some relevant review prior research. (a). **Augmented Lagrangian Methods** The AL Method originates from (Hestenes 1969) and (Powell 1969) in convex settings while subsequent work by Rockafellar in the papers (Rockafellar 1973; Rockafellar 1976) provides a comprehensive theoretical underpinning coupled with rate guarantees. In fact, AL schemes represented a basis for `minos`, a nonlinear programming solver developed by Murtaugh and Saunders (Murtagh and Saunders 1978). Over the last fifteen years, there has been a pronounced effort in developing inexact AL schemes with complexity guarantees for addressing deterministic convex optimization problems with possibly composite objectives and either conic or more general constraints (Aybat and Iyengar 2013; Lan and Monteiro 2013; Necoara et al. 2019; Xu 2021). When f and g are expectation-valued, there has been far less research. The only two available schemes are provided in (Zhang et al. 2023; Zhang et al. 2022) and both are equipped with a rate of $\mathcal{O}(\frac{1}{\sqrt{K}})$, but the first algorithm necessitates solving the AL problem exactly in finite time. Given that the latter is a compositional expectation-valued problem in that subproblem objective is characterized by the squared norm of the constraint expectation, this is generally not possible. *Instead, in this paper, we develop an inexact framework that requires approximate solutions of a smoothed AL problem, where the* *smoothing allows for employing first-order techniques for computing such an approximate solution in finite time. Furthermore, our new algorithm attains sublinear convergence for constant penalty parameters and linear convergence for geometrically increasing penalty parameters.* (b) **Alternate schemes.** An alternate cooperative stochastic approximation (CSA) framework for addressing convex optimization problems with a single expectation constraint is adopted in (Lan and Zhou 2020), while in (Yan and Xu 2022), the authors investigate an adaptive primal-dual stochastic gradient method (APriD). Both CSA and APriD obtain a rate of $\mathcal{O}(1/\sqrt{K})$ in terms of expected sub-optimality. (c). **Smoothing techniques.** In (Jalilzadeh et al. 2022), the authors develop a smoothed, accelerated, and variance-reduced scheme that achieves the optimal rate of $\mathcal{O}(1/k)$ and an optimal oracle complexity of $\mathcal{O}(1/\epsilon^2)$ for nonsmooth stochastic convex problems with projection-friendly convex constraint sets. Related research on smoothing approaches may be found in the same reference.

Gap: No efficient AL schemes for convex programs with nonsmooth expectation-valued constraints.

1.2 Contributions and Organization

Motivated by the aforementioned gap, we present a smoothed variance-reduced AL method (**VR-AL**) that allows for expectation-valued objectives and constraints with nonsmooth (but smoothable) integrands. In contrast with traditional AL schemes, the proposed scheme employs both penalty and smoothing parameter sequences to develop a smooth AL subproblem and allows for inexact resolution of the stochastic AL subproblem. In Section 2, we present the framework while in Section 3, under a constant penalty parameter, we show that the dual suboptimality, primal suboptimality, and primal infeasibility diminish at the rate of $\mathcal{O}(1/K)$, $\mathcal{O}(1/\sqrt{K})$, and $\mathcal{O}(1/\sqrt{K})$, respectively. Analogous geometric rates are provided when the penalty parameter sequence grows at a geometric rate. The paper concludes in Section 4 where we comment on flexibility of the framework in accommodating convex constraints and weakly convex objectives.

2 A SMOOTHED STOCHASTIC AUGMENTED LAGRANGIAN FRAMEWORK

In this section, we provide some preliminaries required for generalizing the augmented Lagrangian framework to regimes with nonsmooth expectation-valued constraints. We begin by providing some background. Corresponding to problem (NSCopt), we may define $\mathcal{L}_0(\mathbf{x},\lambda)$ as $\mathcal{L}_0(\mathbf{x},\lambda) \triangleq f(\mathbf{x}) + \lambda^\top g(\mathbf{x})$ where $g(\mathbf{x}) \triangleq \mathbb{E}\left[\,\tilde{g}(\mathbf{x},\boldsymbol{\xi})\,\right]$ and $\lambda \geq 0$. This allows for denoting the set of minimizers of $\mathcal{L}_0(\mathbf{x},\lambda)$ by $\mathcal{X}^*(\lambda)$, the dual function by $\mathcal{D}_0(\lambda)$, and the dual solution set by Λ^*, each of which is defined next.

$$\mathcal{X}^*(\lambda) \triangleq \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_0(\mathbf{x},\lambda), \mathcal{D}_0(\lambda) \triangleq \inf_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_0(\mathbf{x},\lambda), \text{ and } \Lambda^* \triangleq \arg\max_{\lambda\geq 0} \mathcal{D}_0(\lambda).$$

By adding a slack variable $\mathbf{v}\in\mathbb{R}^m$, we may recast (NSCopt) as $\left\{\min_{\mathbf{x}\in\mathcal{X},\mathbf{v}\geq 0}\left\{\, f(\mathbf{x}) \;\middle|\; g(\mathbf{x})+\mathbf{v}=0\,\right\}\right\}$, where $\lambda\in\mathbb{R}^m$ denotes the Lagrange multiplier associated with the constraint $g(\mathbf{x})+\mathbf{v}=0$. Then the augmented Lagrangian (AL) function, denoted by $\mathcal{L}_\rho$, is defined as

$$\mathcal{L}_\rho(\mathbf{x},\lambda) \triangleq \min_{\mathbf{v}\geq 0}\left\{ f(\mathbf{x}) + \lambda^\top(g(\mathbf{x})+\mathbf{v}) + \tfrac{\rho}{2}\|g(\mathbf{x})+\mathbf{v}\|^2 \right\}. \tag{1}$$

Similarly the dual problem corresponding to minimizing $\mathcal{L}_\rho(\bullet,\lambda)$ (the augmented dual problem) is defined as $\mathcal{D}_\rho(\lambda) \triangleq \inf_{\mathbf{x}\in\mathcal{X}}\mathcal{L}_\rho(\mathbf{x},\lambda)$. The next result derives the $\nabla_\lambda\mathcal{L}_\rho(\lambda)$ and $\nabla_\lambda\mathcal{D}_\rho(\lambda)$, where Π_+ and d_- denote the Euclidean projection onto $\mathbb{R}^n_+$ and the distance to the $\mathbb{R}^n_-$, respectively.

Lemma 1 ((Zhang et al. 2024) Lemma 1,2) Consider $\mathcal{L}_\rho$ defined in (1) for $\rho > 0$, $\mathbf{x}\in\mathcal{X}$ and $\lambda \geq 0$. Then the following hold.

(i) $\mathcal{L}_\rho$ and $\nabla_\lambda \mathcal{L}_\rho$ can be expressed as follows.

$$\mathcal{L}_\rho(\mathbf{x},\lambda) = \left(f(\mathbf{x}) + \tfrac{\rho}{2}\left(d_-\left(\tfrac{\lambda}{\rho} + g(\mathbf{x})\right)\right)^2 - \tfrac{1}{2\rho}\|\lambda\|^2 \right) \text{ and } \nabla_\lambda \mathcal{L}_\rho(\mathbf{x},\lambda) = \left(-\tfrac{\lambda}{\rho} + \Pi_+\left(\tfrac{\lambda}{\rho} + g(\mathbf{x})\right)\right).$$

(ii) $\mathcal{D}_\rho$ is a C^1, concave, and the Moreau envelope of $\mathcal{D}_0$, defined as $\mathcal{D}_\rho(\lambda) \triangleq \max\limits_{u\in\mathbb{R}^m} \left[\mathcal{D}_0(u) - \frac{1}{2\rho}\|u-\lambda\|^2\right]$. Furthermore, $\nabla_\lambda \mathcal{D}_\rho(\lambda) = \frac{1}{\rho}\left(q_\rho(\lambda) - \lambda\right),$ where $q_\rho(\lambda) \triangleq \arg\max\limits_u \left[\mathcal{D}_0(u) - \frac{1}{2\rho}\|u-\lambda\|^2\right]$.

To address nonsmoothness. we consider a smoothing f_η and g_η corresponding to f and g, defined next.

Definition 1 (Beck and Teboulle 2012) *Consider a closed, convex, proper function $h : \mathbb{R}^n \to \mathbb{R}$. A convex function is said to be (α,β)-smoothable if for any $\eta > 0$, there exists a convex C^1 function h_η such that*

$$\begin{aligned} \|\nabla_{\mathbf{z}} h_\eta(\mathbf{z}_1) - \nabla_{\mathbf{z}} h_\eta(\mathbf{z}_2)\| &\le \tfrac{\alpha}{\eta}\|\mathbf{z}_1 - \mathbf{z}_2\|, \quad \forall \mathbf{z}_1,\mathbf{z}_2 \in \mathbb{R}^n \\ h_\eta(\mathbf{z}) \le h(\mathbf{z}) &\le h_\eta(\mathbf{z}) + \eta\beta, \qquad \forall \mathbf{z} \in \mathbb{R}^n. \end{aligned}$$

Remark When $h(\mathbf{x}) \triangleq \mathbb{E}[\tilde{h}(\mathbf{x},\boldsymbol{\xi})]$, then smoothed function h_η is defined as $h_\eta(\mathbf{x}) = \mathbb{E}[\tilde{h}_\eta(\mathbf{x},\boldsymbol{\xi})]$, where $\tilde{h}_\eta(\bullet,\xi)$ denotes the smoothing of $h(\bullet,\xi)$.

We now present our main assumption.

Assumption 2 *(a) The function $\tilde{f}(\bullet,\xi)$ is an (α,β)-smoothable real-valued function for any ξ. (b) For $i = 1,\ldots,m$, the constraint function $\tilde{g}_i(\bullet,\xi)$ is an (α,β)-smoothable real-valued function. (c) There exists a point $(\mathbf{x}^*,\lambda^*)$ satisfying the KKT conditions. (d) The set $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex and compact set. (e) There exists a vector $\bar{\mathbf{x}} \in \mathcal{X}$ such that $g_i(\bar{\mathbf{x}}) < 0$ for $i = 1,\ldots,m$.*

The above assumption ensures that f and g_i are (α,β) smoothable for $i = 1,\cdots,m$. We now consider the smoothed counterpart of (NSCopt), given by (NSCopt$_\eta$) and defined as

$$\min_{\mathbf{x}\in\mathcal{X},\mathbf{v}\ge 0} \left\{ f_\eta(\mathbf{x}) \mid g_\eta(\mathbf{x}) + \mathbf{v} = 0 \right\}. \tag{NSCopt$_\eta$}$$

The resulting smoothed Lagrangian function $\mathcal{L}_{\eta,0}$ and the dual function $\mathcal{D}_{\eta,0}(\lambda)$ are defined as follows.

$$\mathcal{L}_{\eta,0}(\mathbf{x},\lambda) \triangleq \begin{cases} f_\eta(\mathbf{x}) + \lambda^\top g_\eta(\mathbf{x}) & \lambda \ge 0 \\ -\infty, & \text{otherwise} \end{cases} \quad \text{and} \quad \mathcal{D}_{\eta,0}(\lambda) \triangleq \inf_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\eta,0}(\mathbf{x},\lambda).$$

where $g_\eta(\mathbf{x}) \triangleq \mathbb{E}[\tilde{g}_\eta(\mathbf{x},\boldsymbol{\xi})]$. Then the smoothed augmented Lagrangian function $\mathcal{L}_{\eta,\rho}$ is defined as

$$\mathcal{L}_{\eta,\rho}(\mathbf{x},\lambda) \triangleq \min_{\mathbf{v}\ge 0} \left\{ f_\eta(\mathbf{x}) + \lambda^\top (g_\eta(\mathbf{x}) + \mathbf{v}) + \tfrac{\rho}{2}\|g_\eta(\mathbf{x}) + \mathbf{v}\|^2 \right\}.$$

We may now define $\mathcal{D}_{\eta,\rho}$ and $q_{\eta,\rho}$ as follows where $q_{\eta,\rho}(\lambda) \triangleq \mathrm{argmax}_u \left[\mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho}\|u-\lambda\|^2\right]$.

$$\mathcal{D}_{\eta,\rho}(\lambda) = \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_{\eta,0}(u) - \tfrac{1}{2\rho}\|u-\lambda\|^2\right] \text{ and } \nabla_\lambda \mathcal{D}_{\eta,\rho}(\lambda) = \tfrac{1}{\rho}\left(q_{\eta,\rho}(\lambda) - \lambda\right).$$

Next, we presented some important results about the smoothed Lagrangian, dual, and augmented dual to their nonsmooth counterparts from (Zhang et al. 2024, Lemma 3-4, Proposition 21).

Lemma 3 *Suppose Assumption 2 holds. For any $\lambda \in \mathbb{R}^m_+$ and $\mathbf{x} \in \mathcal{X}$, any $\eta,\rho > 0$.*
(i) $\mathcal{L}_{\eta,0}(\mathbf{x},\lambda) \le \mathcal{L}_0(\mathbf{x},\lambda) \le \mathcal{L}_{\eta,0}(\mathbf{x},\lambda) + \eta\left(\|\lambda\| m + 1\right)\beta$;
(ii) $|\mathcal{D}_{\eta,0}(\lambda) - \mathcal{D}_0(\lambda)| \le \eta\left(\|\lambda\| m + 1\right)\beta$;
(iii) $|\mathcal{D}_{\eta,\rho}(\lambda) - \mathcal{D}_\rho(\lambda)| \le \eta\left(\|\lambda\| m + 1\right)\beta$.

(iv) (Slater condition) For any $\eta > 0$, there exists $\bar{\mathbf{x}} \in \mathcal{X}$ such that $g_\eta(\bar{\mathbf{x}}) < 0$;
(v) The set of optimal multipliers Λ^* for (NSCopt) is bounded as per

$$\Lambda^* \subseteq \left\{ \lambda \geq 0 \,\Big|\, \sum_{i=1}^m \lambda_i \leq b_\lambda \right\} \textit{ where } b_\lambda \geq \tfrac{f(\bar{\mathbf{x}}) - \mathcal{D}_0^*}{\min_j \{-g_j(\bar{\mathbf{x}})\}};$$

(vi) For any $\eta > 0$, the set of optimal multipliers Λ_η^* for (NSCopt$_\eta$) is bounded as per

$$\Lambda_\eta^* \subseteq B_{\lambda,\eta} = \left\{ \lambda \geq 0 \,\Big|\, \sum_{i=1}^m \lambda_i \leq b_{\lambda,\eta} \right\} \textit{ where } b_{\lambda,\eta} \geq \tfrac{f(\bar{\mathbf{x}}) - \mathcal{D}_0^* + \eta(\beta + \tilde{C}^*)}{\min_j \{-g_j(\bar{\mathbf{x}})\}};$$

(vii) $\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\| \leq \sqrt{4\rho\eta\left(\|\lambda\|m + C_m\right)\beta}$, where C_m is a constant;
(viii) $\|\nabla_\lambda \mathcal{D}_{\eta,\rho}(\lambda) - \nabla_\lambda \mathcal{D}_\rho(\lambda)\| = \frac{1}{\rho}\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\| \leq \sqrt{\frac{4\eta(\|\lambda\|m + C_m)\beta}{\rho}}$.

We now present our scheme which enjoys two key distinctions with deterministic variants.
(i) The Lagrangian subproblem is a compositional stochastic optimization problem and an inexact (random) solution is available in finite time via stochastic approximation schemes by using N_k evaluations of a suitable first-order oracle defined in the next subsection. The inexactness is captured by an error sequence $\{\epsilon_k\}$ which is driven to zero at a suitable rate, a consequence of raising N_k to infinity.
(ii) The traditional update of the Lagrange multiplier requires an exact evaluation of $g(\mathbf{x})$. Since g is an expectation-valued function, an exact update is unavailable in finite time. Instead, we employ a sampled update reliant on M_k evaluations of a zeroth-order oracle, leading to an error w_k. By driving M_k to infinity at a suitable rate, the bias captured by w_k is driven to zero.

Our framework requires access to a stochastic first-order oracle $\mathcal{SO}^{\text{fir}}$ associated with $\nabla_{\mathbf{x}} \tilde{f}_\eta(\mathbf{x}, \xi)$ and $\nabla_{\mathbf{x}} \tilde{g}_{i,\eta}(\mathbf{x}, \xi)$ for $i = 1, \cdots, m$ and a stochastic zeroth-order oracle $\mathcal{SO}^{\text{zer}}$ for $g_{i,\eta}(\mathbf{x}, \xi)$ for $i = 1, \cdots, m$. To clarify the nature of the stochastic zeroth and first-order oracles, we present three nonsmooth random functions $\tilde{f}(\bullet, \xi)$ in Table 1 for which smoothings are defined and analyzed.

Table 1: Bounding the second moments for certain smoothings.

$\tilde{f}(x,\xi)$	$\tilde{f}_\eta(x,\xi)$	$\nabla \tilde{f}_\eta(x,\xi)$	$\mathbb{E}[\|\nabla_x \tilde{f}_\eta(x,\xi) - \nabla_x f_\eta(x)\|^2]$
$\tilde{f}_1(x,\xi) = \lambda(\xi)\|x\|_1$	$\sum_{i=1}^n h_\eta(x_i,\xi)$, where $h_\eta(x_i,\xi) = \left\{ \begin{array}{ll} \lambda^2(\xi)\frac{x_i^2}{2\eta}, & \lambda(\xi)\|x_i\| < \eta \\ \lambda(\xi)\|x_i\| - \eta/2, & o.w. \end{array} \right\}$	$[\nabla_{x_i} h_\eta(x_i,\xi)]_{i=1}^n$, where $\nabla_{x_i} h_\eta(x_i,\xi) = \left\{ \begin{array}{ll} \lambda^2(\xi)\frac{x_i}{\eta}, & \lambda(\xi)\|x_i\| < \eta \\ \lambda(\xi)x_i/\|x_i\|, & o.w. \end{array} \right\}$	$4n\mathbb{E}[\lambda^2(\xi)]$
$\tilde{f}_2(x,\xi) = \lambda(\xi)\|x\|_2$	$\sqrt{\lambda^2(\xi)\|x\|^2 + \eta^2} - \eta$	$\frac{\lambda^2(\xi)x}{\sqrt{\lambda^2(\xi)\|x\|^2+\eta^2}}$	$4\mathbb{E}[\lambda^2(\xi)]$
$\tilde{f}_3(x,\xi) = \max\limits_{1\leq i\leq n}\{h_i(x,\xi)\}$ where $h_i(x,\xi) = v_i + s_i c(\xi)^T x$	$\eta \log\left(\sum_{i=1}^n \exp(h_i(x,\xi)/\eta)\right)$	$\frac{\sum_{i=1}^n \nabla_x h_i(x,\xi)\exp(h_i(x,\xi)/\eta)}{\sum_{i=1}^n \exp(h_i(x,\xi)/\eta)}$	$4\mathbb{E}\left[\left(\max\limits_{1\leq i\leq n}\|s_i c(\xi)\|\right)^2\right]$,

Definition 2 (First and zeroth-order oracles) Given an $\mathbf{x} \in \mathcal{X}$ and $\eta > 0$, $\mathcal{SO}^{\text{zer}}$ returns a random vector $\tilde{g}_\eta(\mathbf{x}, \xi)$; (ii) Given $\mathbf{x} \in \mathcal{X}$, the $\mathcal{SO}^{\text{fir}}$ returns $\nabla_{\mathbf{x}} \tilde{f}_\eta(\mathbf{x}, \xi)$ and $\nabla_{\mathbf{x}} \tilde{g}_{i,\eta}(\mathbf{x}, \xi), i = 1, \ldots, m$.

We now formally state the variance-reduced augmented Lagrangian scheme where $\mathcal{F}_k$ is defined as

$$\mathcal{F}_0 \triangleq \{\mathbf{x}_0\}, \mathcal{F}_k \triangleq \mathcal{F}_{k-1} \cup \{\tilde{g}_{\eta_k}(\mathbf{x}_k, \xi_j)\}_{j=1}^{M_k} \cup \left\{ \nabla_x \tilde{f}_{\eta_k}(\mathbf{x}, \xi_j) \cup \{\nabla_x \tilde{g}_{i,\eta_k}(\mathbf{x}, \xi_j)\}_{i=1}^m \right\}_{j=1}^{N_k}, \quad k \geq 1.$$

Variance-reduced augmented Lagrangian scheme (VR-AL). Given $\mathbf{x}_0, \lambda_0$ and sequences $\{\rho_k, \epsilon_k, \eta_k, N_k, M_k\}$. For $k = 1, \cdots, K$,

[1] $\mathbf{x}_{k+1}$ satisfies $\mathbb{E}[\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) \mid \mathcal{F}_k] \leq \epsilon_k \eta_k^b \quad a.s.$ with N_k evals of $\mathcal{SO}^{\text{fir}}$;

[2] $\lambda_{k+1} = \lambda_k + \rho_k \left(\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) + w_k\right),$ where w_k requires M_k evals of $\mathcal{SO}^{\text{zer}}$.

In (**VR-AL**), step [1] requires inexactly solving an stochastic optimization problem such that the sub-optimality is within $\epsilon_k \eta_k^b$ by leveraging N_k samples of the first-order oracle $\mathcal{SO}^{\rm fir}$ (equivalently requiring taking N_k steps of a stochastic gradient scheme). The "a.s." requirement is introduced since the suboptimality metric is a conditional expectation. Step [2] captures a Lagrange multiplier update which employs a mini-batch of M_k samples of the ZO oracle.

Lemma 4 Consider the sequence generated by Algo. **VR-AL**. Then for any $k > 0$, step [2] is equivalent to

$$\lambda_{k+1} = \Pi_+ \left(\lambda_k + \rho_k \bar{g}_{\eta_k, M_k}(\mathbf{x}_k) \right).$$

Proof. We may observe that

$$\begin{aligned} \lambda_{k+1} &= \lambda_k + \rho_k \left(\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) + w_k) \right) = \lambda_k + \rho_k \left(-\tfrac{\lambda_k}{\rho_k} + \Pi_+ \left(\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1}) \right) + w_k) \right) \\ &= \rho_k \Pi_+ \left(\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1}) \right) + \rho_k w_k = \Pi_+ \left(\lambda_k + \rho_k g_{\eta_k}(\mathbf{x}_k) \right) + \rho_k w_k. \end{aligned} \tag{2}$$

Suppose w_k is defined as

$$w_k \triangleq \left(\Pi_+ \left(\tfrac{\lambda_k}{\rho_k} + [\bar{g}_{\eta_k,M_k}(\mathbf{x}_{k+1})] \right) - \Pi_+ \left(\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1}) \right) \right). \tag{3}$$

Then by substituting (3) in (2), we obtain that $\lambda_{k+1} = \Pi_+ \left(\lambda_k + \rho_k \bar{g}_{\eta_k,M_k}(\mathbf{x}_k) \right).$ □

We now impose moment assumptions on the ZO oracle and a requirement on the parameter sequences.

Assumption 5 [ZO oracle, Parameter sequences] (i) For any $j \in [m]$ and $\mathbf{x} \in \mathcal{X}$, $\mathbb{E}[\|\tilde{g}_{\eta,j}(\mathbf{x},\boldsymbol{\xi}) - g_{\eta,j}(\mathbf{x})\|^2] \leq \nu_j^2$; (ii) Suppose $\{\rho_k, \epsilon_k, \eta_k, M_k, N_k\}$ be such that $\left\{ \sqrt{2\rho_k \epsilon_k \eta_k^b} + \frac{\nu_g \rho_k}{\sqrt{M_k}} + 2\sqrt{\eta_k \rho_k} \right\}$ be summable.

While we do not impose an assumption on the FO oracles, to ensure that step [1] in (**VR-AL**) is well-defined, we implicitly need suitably conditional unbiasedness requirement and a bound on the conditional second moments. This will be clarified when the complexity analysis is carried out in future work. We may now derive the following bound.

Lemma 6 Suppose Assumption 5 holds. For any k, $\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \leq \nu_G^2 / M_k$ with a constant ν_G.

Proof. By empploying the nonexpansivity of the Euclidean projection,

$$\begin{aligned} \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] &\leq \mathbb{E}\left[\left\| \Pi_+ \left(\tfrac{\lambda_k}{\rho_k} + [\bar{g}_{\eta_k,M_k}(\mathbf{x}_{k+1})] \right) - \Pi_+ \left((\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1}) \right) \right\|^2 \mid \mathcal{F}_k \right] \\ &\leq \mathbb{E}\left[\|\bar{g}_{\eta_k,M_k}(\mathbf{x}_{k+1}) - g_{\eta_k}(\mathbf{x}_{k+1})\|^2 \mid \mathcal{F}_k \right] \leq \tfrac{\nu_G^2}{M_k}. \end{aligned}$$

□

Unless mentioned otherwise, Assumptions 2 and 5 hold throughout this paper.

3 RATE ANALYSIS OF STOCHASTIC AUGMENTED LAGRANGIAN SCHEME

3.1 Preliminary Results

We begin by deriving the following bound, an extension of the result proved in (Rockafellar 1973).

Lemma 7 Consider the sequence $\{(\mathbf{x}_k, \lambda_k)\}$ generated by (**VR-AL**). Suppose $\mathbf{x}_{k+1}$ satisfies $\mathbb{E}[\mathcal{L}_{\rho_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\rho_k}(\lambda_k) \mid \mathcal{F}_k] \leq \epsilon_k \eta_k^b$ with $b \geq 0$. Then the following holds a.s. for any $k \geq 0$.

$$\mathbb{E}\left[\|\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(x_{k+1}, \lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k)\|^2 \mid \mathcal{F}_k \right] \leq \tfrac{2\epsilon_k \eta_k^b}{\rho_k}.$$

Proof. We employ the arguments in (Rockafellar 1976; Zhang et al. 2024) that

$$\mathcal{D}_{\eta_k,\rho_k}(\lambda) + \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda)^\top (u-\lambda) \geq \mathcal{D}_{\eta_k,\rho_k}(u) \geq \mathcal{D}_{\eta_k,\rho_k}(\lambda) + (u-\lambda)^\top \nabla \mathcal{D}_{\eta_k,\rho_k}(\lambda) - \tfrac{1}{2\rho_k}\|u-\lambda\|^2.$$

Next, by the concavity of $\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x},\bullet)$, we have that

$$\begin{aligned} \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) + (w-\lambda_k)^\top \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) &\geq \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},w) \geq \mathcal{D}_{\eta_k,\rho_k}(w) \\ &\geq \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) + (w-\lambda_k)^\top \nabla \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) - \tfrac{1}{2\rho_k}\|w-\lambda_k\|^2 \\ \implies \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) - \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) &\geq (w-\lambda_k)^\top (\nabla \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)) - \tfrac{1}{2\rho_k}\|w-\lambda_k\|^2 \\ &= u^\top (\nabla \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)) - \tfrac{1}{2\rho_k}\|u\|^2. \end{aligned}$$

But since this inequality holds for all u, we have that

$$\begin{aligned} \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) - \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) &\geq \sup_{u\in\mathbb{R}^m} \left\{ u^\top (\nabla \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)) - \tfrac{1}{2\rho_k}\|u\|^2 \right\} \\ &= \tfrac{\rho_k}{2}\|\nabla \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)\|^2. \end{aligned}$$

Recall that $\mathbf{x}_{k+1}$ satisfies $\mathbb{E}\left[\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) - \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) \mid \mathcal{F}_k\right] \leq \epsilon_k \eta_k^b$ in an a.s. sense, implying that $\mathbb{E}\left[\|\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k)\|^2 \mid \mathcal{F}_k\right] \leq \frac{2\epsilon_k\eta_k^b}{\rho_k}$ holds in an a.s. sense. □

We now derive a bound on the multiplier sequence $\{\lambda_k\}$ in an expected-value sense.

Lemma 8 (Bound on λ_k) Suppose $\{(\mathbf{x}_k,\lambda_k)\}$ is generated by (**VR-AL**). Then the following hold. (a) $\{\lambda_k\}$ is a convergent sequence in an a.s. sense; (b) For any K, we have that

$$\mathbb{E}\left[\|\lambda_K - \lambda^*\|\right] \leq \sum_{k=0}^{K-1} \left(\sqrt{2\rho_k\epsilon_k\eta_k^b} + \frac{\nu_g\rho_k}{\sqrt{M_k}} + 2\sqrt{\eta_k\rho_k(\|\lambda^*\|m+1)B} \right) + \|\lambda_0 - \lambda^*\| \leq B_\lambda.$$

Proof. We begin by deriving a bound on $\|\lambda_{k+1}-\lambda^*\|$ as follows

$$\begin{aligned} \|\lambda_{k+1}-\lambda^*\| &\leq \|\lambda_{k+1} - q_{\eta_k,\rho_k}(\lambda_k)\| + \|q_{\eta_k,\rho_k}(\lambda_k) - q_{\eta_k,\rho_k}(\lambda^*)\| + \|q_{\eta_k,\rho_k}(\lambda^*) - q_{\rho_k}(\lambda^*)\| + \|q_{\rho_k}(\lambda^*) - \lambda^*\| \\ &\leq \|\lambda_{k+1} - q_{\eta_k,\rho_k}(\lambda_k)\| + \|\lambda_k - \lambda^*\| + \|q_{\eta_k,\rho_k}(\lambda^*) - q_{\rho_k}(\lambda^*)\| + 0, \end{aligned}$$

where the first inequality is a result of the triangle inequality while the second inequality is a result of the non-expansivity of $q_{\eta,\rho}(\bullet)$ and by noting that $\lambda^* = q_\rho(\lambda^*)$. We now derive a bound on $\|\lambda_{k+1} - q_{\eta_k,\rho_k}(\lambda_k)\|$.

$$\begin{aligned} \|\lambda_{k+1} - q_{\eta_k,\rho_k}(\lambda_k)\| &= \|\lambda_k + \rho_k\left(\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) + w_k)\right) - q_{\eta_k,\rho_k}(\lambda_k)\| \\ &= \|\lambda_k + \rho_k\left(\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) + w_k)\right) - \rho_k \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) - \lambda_k\| \\ &\leq \rho_k\|\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k)\| + \rho_k\|w_k\|. \end{aligned} \tag{4}$$

Recall by the conditional variant of Jensen's inequality, we have that

$$\begin{aligned} \left(\mathbb{E}\left[\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\tilde{\lambda}) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(x_{k+1},\tilde{\lambda}))\| \mid \mathcal{F}_k\right]\right)^2 &\leq \mathbb{E}\left[\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\tilde{\lambda}) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(x_{k+1},\tilde{\lambda}))\|^2 \mid \mathcal{F}_k\right] \\ &\leq \tfrac{2\epsilon_k\eta_k^b}{\rho_k} \\ \implies \mathbb{E}\left[\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\tilde{\lambda}) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(x_{k+1},\tilde{\lambda}))\| \mid \mathcal{F}_k\right] &\leq \sqrt{\tfrac{2\epsilon_k\eta_k^b}{\rho_k}}. \end{aligned}$$

Taking expectations conditioned on $\mathcal{F}_k$ on both sides of (4),

$$\begin{aligned}\mathbb{E}\left[\|\lambda_{k+1} - q_{\eta_k,\rho_k}(\lambda_k)\| \mid \mathcal{F}_k\right] &\leq \rho_k \mathbb{E}\left[\|\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k)\| \mid \mathcal{F}_k\right] + \rho_k \mathbb{E}\left[\|w_k\| \mid \mathcal{F}_k\right] \\ &\leq \sqrt{2\rho_k \epsilon_k \eta_k^b} + \tfrac{\nu_G \rho_k}{\sqrt{M_k}}.\end{aligned}$$

From Lemma 3, $\|q_{\eta_k,\rho_k}(\lambda^*) - q_{\rho_k}(\lambda^*)\| \leq 2\sqrt{\rho_k \eta_k (\|\lambda^*\|m + C_m)B}$. Consequently, we obtain that

$$\mathbb{E}\left[\|\lambda_{k+1} - \lambda^*\| \mid \mathcal{F}_k\right] \leq \sqrt{2\rho_k \epsilon_k \eta_k^b} + \tfrac{\nu_G \rho_k}{\sqrt{M_k}} + 2\sqrt{\rho_k \eta_k (\|\lambda^*\|m + C_m)B} + \left[\|\lambda_k - \lambda^*\|\right].$$

It follows from the Robbins-Siegmund Lemma that if $\sqrt{2\rho_k \epsilon_k \eta_k^b} + \frac{\nu_G \rho_k}{\sqrt{M_k}} + 2\sqrt{\rho_k \eta_k (\|\lambda^*\|m + C_m)B}$ is summable, then $\{\|\lambda_k - \lambda^*\|\}$ converges almost surely to a nonnegative random variable. It follows that $\{\lambda_k\}$ is convergent almost surely and is therefore bounded almost surely.
(b) Taking unconditional expectations and summing from $k = 0, \cdots, K-1$, we obtain that $\mathbb{E}\left[\|\lambda_K - \lambda^*\|\right] \leq \sum_{k=0}^{K-1} \left(\sqrt{2\rho_k \epsilon_k \eta_k^b} + \frac{\nu_G \rho_k}{\sqrt{M_k}} + 2\sqrt{\eta_k \rho_k (\|\lambda^*\|m + 1)B}\right) + \|\lambda_0 - \lambda^*\| \leq B_\lambda$. □

3.2 Rate Analysis Under Constant ρ_k

Proposition 9 (Dual sub-optimality) Consider the sequence $\{(\mathbf{x}_k, \lambda_k)\}$ generated by (**VR-AL**). Suppose $\rho_k = \rho$ for every $k \geq 0$. Then the following holds for $\bar{\lambda}_K \triangleq \frac{\sum_{i=0}^{K-1} \lambda_i}{K}$ and for any $K > 0$.

$$\mathbb{E}\left[f^* - \mathcal{D}_\rho(\bar{\lambda}_K)\right] \leq \tfrac{1}{K}\mathbb{E}[\|\lambda_0 - \lambda^*\|^2 + \tfrac{1}{K}\sum_{k=0}^{K-1}\left(\left(\tfrac{\nu_G}{\sqrt{M_k}} + \tfrac{\sqrt{2\epsilon_k \eta_k^b}}{\sqrt{\rho}} + \eta_k m \beta\right) B_\lambda + 2\eta_k (b_\lambda m + 1)\beta\right).$$

Proof. Recall that $\mathcal{D}_{\eta_k,\rho}$ is the Moreau envelope of $\mathcal{D}_{\eta_k,0}$. Consequently, $\nabla_\lambda \mathcal{D}_{\eta_k,\rho}$ is $\frac{1}{\rho}$-Lipschitz. We may then claim the following.

$$\begin{aligned}-\mathcal{D}_{\eta_k,\rho}(\lambda_{k+1}) &\leq -\mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{D}_{\eta,\rho}(\lambda_k)^\top (\lambda_{k+1} - \lambda_k) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\ &\leq -\mathcal{D}_{\eta_k,\rho}(\lambda^*) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k)^\top (\lambda_{k+1} - \lambda^*) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2,\end{aligned}$$

where $-\mathcal{D}_{\eta_k,\rho}(\lambda^*) \geq -\mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda)^\top (\lambda^* - \lambda_k)$. It follows that

$$\begin{aligned}-\mathcal{D}_{\eta_k,\rho}(\lambda_{k+1}) \leq &- \mathcal{D}_{\eta_k,\rho}(\lambda^*) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)^\top (\lambda_{k+1} - \lambda^*) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\ &-(\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)^\top (\lambda_{k+1} - \lambda^*) \\ \leq &- \mathcal{D}_{\eta_k,\rho}(\lambda^*) - \tfrac{1}{\rho}(\lambda_{k+1} - \lambda_k)^\top (\lambda_{k+1} - \lambda^*) - w_k^\top (\lambda_{k+1} - \lambda^*) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\ &- (\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)^\top (\lambda_{k+1} - \lambda^*) \\ \leq &- \mathcal{D}_{\eta_k,\rho}(\lambda^*) - \tfrac{1}{\rho}(\lambda_{k+1} - \lambda_k)^\top (\lambda_{k+1} - \lambda^*) + \|w_k\|\|\lambda_{k+1} - \lambda^*\| + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\ &+ \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|\|\lambda_{k+1} - \lambda^*\| \\ = &- \mathcal{D}_{\eta_k,\rho}(\lambda^*) + \tfrac{1}{2\rho}(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2) + \|w_k\|\|\lambda_{k+1} - \lambda^*\| \\ &+ \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|\|\lambda_{k+1} - \lambda^*\| \\ \leq &- \mathcal{D}_p(\lambda^*) + \eta_k (\|\lambda^*\|m + 1)\beta + \tfrac{1}{2\rho}(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2) + \|w_k\|\|\lambda_{k+1} - \lambda^*\| \\ &+ \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|\|\lambda_{k+1} - \lambda^*\|.\end{aligned}$$

From Lemma 3, it follows that

$$
\begin{aligned}
&-\mathcal{D}_\rho(\lambda_{k+1}) \leq -\mathcal{D}_\rho(\lambda^*) + \eta_k(\|\lambda_{k+1}\|m+1)\beta + \eta_k(\|\lambda^*\|m+1)\beta + \tfrac{1}{2\rho}(\|\lambda_k-\lambda^*\|^2 - \|\lambda_{k+1}-\lambda^*\|^2) \\
&+ \|w_k\|\|\lambda_{k+1}-\lambda^*\| + \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\|\|\lambda_{k+1}-\lambda^*\| \\
&\leq -\mathcal{D}_\rho(\lambda^*) + \eta_k(\|\lambda_{k+1}-\lambda^*\|m)\beta + 2\eta_k(\|\lambda^*\|m+1)\beta + \tfrac{1}{2\rho}(\|\lambda_k-\lambda^*\|^2 - \|\lambda_{k+1}-\lambda^*\|^2) \\
&+ \|w_k\|\|\lambda_{k+1}-\lambda^*\| + \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\|\|\lambda_{k+1}-\lambda^*\| \\
&\leq -\mathcal{D}_\rho(\lambda^*) + 2\eta_k(\|\lambda^*\|m+1)\beta + \tfrac{1}{2\rho}(\|\lambda_k-\lambda^*\|^2 - \|\lambda_{k+1}-\lambda^*\|^2) \\
&+ \|w_k\|\|\lambda_{k+1}-\lambda^*\| + (\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\| + \eta_k m\beta)\,\|\lambda_{k+1}-\lambda^*\|.
\end{aligned}
$$

By summing from $k=0$ to $K-1$ and defining $f^* = f(\mathbf{x}^*)$, we obtain the following inequality.

$$
\begin{aligned}
\sum_{k=0}^{K-1} (-\mathcal{D}_\rho(\lambda_{k+1}) + f^*) &\leq \tfrac{1}{2\rho}(\|\lambda_0-\lambda^*\|^2 - \|\lambda_K-\lambda^*\|^2) + \sum_{k=0}^{K-1} 2\eta_k(b_\lambda m+1)\beta \\
&+ \sum_{k=0}^{K-1} (\|w_k\| + \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\| + \eta_k m\beta)\,\|\lambda_{k+1}-\lambda^*\|.
\end{aligned}
$$

Dividing by K, invoking the concavity of $\mathcal{D}_\rho$, and by taking expectations on both sides,

$$
\begin{aligned}
&\mathbb{E}\left[f^* - \mathcal{D}_\rho(\bar{\lambda}_K)\right] \leq \mathbb{E}\left[\tfrac{1}{K}\sum_{k=0}^{K-1}(f^* - \mathcal{D}_\rho(\lambda_{k+1}))\right] \leq \tfrac{1}{\rho K}\mathbb{E}[\|\lambda_0-\lambda^*\|^2 + \tfrac{1}{K}\sum_{k=0}^{K-1} 2\eta_k(b_\lambda m+1)\beta \\
&+ \tfrac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[(\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\| + \|w_k\| + \eta_k m\beta)\,\|\lambda_{k+1}-\lambda^*\|\right] \\
&\leq \tfrac{1}{K}\mathbb{E}[\|\lambda_0-\lambda^*\|^2 + \tfrac{1}{K}\sum_{k=0}^{K-1}\left(\left(\tfrac{\nu_G}{\sqrt{M_k}} + \tfrac{\sqrt{2\epsilon_k\eta_k^b}}{\sqrt{\rho}} + \eta_k m\beta\right)B_\lambda + 2\eta_k(b_\lambda m+1)\beta\right) \leq \tfrac{C_D}{K},
\end{aligned}
$$

where boundedness of λ_k follows from Lemma 8 and C_D is a constant. □

Next, we derive a rate statement on the infeasibility.

Proposition 10 (Rate on primal infeasibility) Consider the sequence $\{(\mathbf{x}_k,\lambda_k)\}$ generated by (**VR-AL**). Suppose $\rho_k = \rho$ for every $k \geq 0$. Then the following holds for $\bar{\mathbf{x}}_K = \frac{\sum_{i=0}^{K-1}\mathbf{x}_i}{K}$ and for any $K > 0$.

$$
\mathbb{E}\left[d_\_(g(\bar{\mathbf{x}}_K))\right] \leq \sqrt{\tfrac{m}{K}\sum_{i=0}^{K-1}\left(\tfrac{6\epsilon_i\eta_i^b}{\rho} + 3m^2\eta_i^2\beta^2\right) + \tfrac{6mC_D}{\rho K} + \tfrac{C_B}{\rho K}\sum_{i=0}^{K-1}\eta_i} \triangleq \sqrt{\tfrac{\tilde{C}_d}{K}}
$$

where $\tilde{C}_d$ is a constant.

Proof. We begin by noting that $g_{\eta_k}(\mathbf{x}_{k+1})$ can be expressed as

$$
g_{\eta_k}(\mathbf{x}_{k+1}) = \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k) + \left(\Pi_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)
$$

Recall that $d_-(u+v) \leq d\_(u) + \|v\|$ for any $u, v \in \mathbb{R}^m$. Consequently, we have that

$$
d\_(g_{\eta_k}(\mathbf{x}_{k+1})) \leq \|\nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\| + \underbrace{d\_\left(\Pi_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)}_{=0} = \|\nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\|. \quad (5)
$$

By definition of $d\_(\bullet)$, convexity of $\max\{g_j(\bullet),0\}$, and $\|u\|_2 \leq \|u\|_1 \leq \sqrt{m}\|u\|_2$,

$$d\_\left(g(\bar{\mathbf{x}}_K)\right) = \inf_{u\in\mathbb{R}^m_-} \|g(\bar{\mathbf{x}}_K) - u\|_1 \leq \sum_{j=1}^{m} \inf_{u_j\leq 0} |g_j(\bar{\mathbf{x}}_K) - u_j| = \sum_{j=1}^{m} \max\{g_j(\bar{\mathbf{x}}_K),0\}$$

$$\leq \tfrac{1}{K}\sum_{i=1}^{K-1}\sum_{j=1}^{m} \max\{g_j(\mathbf{x}_{i+1}),0\} \leq \tfrac{1}{K}\sum_{i=1}^{K-1}\sum_{j=1}^{m} \max\{g_{j,\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta, 0\} = \tfrac{1}{K}\sum_{k=0}^{K-1} \inf_{u\in\mathbb{R}^m_-} \|g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1} - u\|_1$$

$$\leq \tfrac{1}{K}\sum_{i=0}^{K-1} \inf_{u\in\mathbb{R}^m_-} \sqrt{m}\|g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1} - u\|_2 = \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1} d\_\left(g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1}\right)$$

$$\leq \tfrac{\sqrt{m}}{K}\sum_{i=1}^{K-1} \left(\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1},\lambda)\| + m\eta_i\beta\right)$$

$$\leq \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1} \left(\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1},\lambda) - \nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + \|\nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + m\eta_i\beta\right).$$

By squaring both sides and taking unconditional expectations, we obtain the following.

$$\mathbb{E}\left[\left(d\_\left(g(\bar{\mathbf{x}}_K)\right)\right)^2\right] \leq \mathbb{E}\left[\left(\tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1} \left(\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1},\lambda) - \nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + \|\nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + m\eta_i\beta\right)\right)^2\right]$$

$$\leq \tfrac{m}{K^2}\sum_{i=0}^{K-1} K\mathbb{E}\left[\left(\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1},\lambda) - \nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + \|\nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + m\eta_i\beta\right)^2\right]$$

$$\leq \tfrac{m}{K}\sum_{i=0}^{K-1} \mathbb{E}\left[3\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1},\lambda) - \nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\|^2 + 3\|\nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\|^2 + 3m^2\eta_i^2\beta^2\right]$$

$$\leq \tfrac{m}{K}\sum_{i=0}^{K-1} \left(\tfrac{6\epsilon_i\eta_i^b}{\rho} + 3m^2\eta_i^2\beta^2\right) + \tfrac{m}{K}\sum_{i=0}^{K-1} \mathbb{E}\left[3\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_i)\|^2\right].$$

Recall that if $\mathcal{D}_{\eta_k,\rho}$ is a $1/\rho-$smooth concave function, then for any $\lambda \in \mathbb{R}^m$, we have that $\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda)\|^2 \leq \frac{2}{\rho}(\mathcal{D}_{\eta_k,\rho}(\lambda^*) - \mathcal{D}_{\eta_k,\rho}(\lambda))$, where λ^* is a maximizer of $\mathcal{D}_\rho$. By leveraging the concavity of the $\sqrt{\bullet}$ function,

$$\mathbb{E}\left[\left(d\_(g(\bar{\mathbf{x}}_K))\right)^2\right] \leq \tfrac{m}{K}\sum_{i=0}^{K-1} \left(\tfrac{6\epsilon_i\eta_i^b}{\rho} + 3m^2\eta_i^2\beta^2\right) + \tfrac{m}{K}\sum_{i=0}^{K-1} \mathbb{E}\left[\tfrac{6}{\rho}\left(\mathcal{D}_{\eta_i,\rho}(\lambda^*_{\eta_i}) - \mathcal{D}_{\eta_k,\rho}(\lambda_i)\right)\right]$$

$$\leq \tfrac{m}{K}\sum_{i=0}^{K-1} \left(\tfrac{6\epsilon_i\eta_i^b}{\rho} + 3m^2\eta_i^2\beta^2\right) + \mathbb{E}\left[\tfrac{6m}{\rho K}\sum_{i=0}^{K-1} \left(\mathcal{D}_\rho(\lambda^*) - \mathcal{D}_\rho(\lambda_i)\right)\right] + \tfrac{6m}{\rho K}\sum_{i=0}^{K-1} \eta_i\beta\left((B_\lambda + 2b_{\lambda,\eta})m + 2\right)$$

$$\leq \tfrac{m}{K}\sum_{i=0}^{K-1} \left(\tfrac{6\epsilon_i\eta_i^b}{\rho} + 3m^2\eta_i^2\beta^2\right) + \tfrac{6m}{\rho}\mathbb{E}\left[\left(\mathcal{D}_\rho(\lambda^*) - \mathcal{D}_\rho\left(\tfrac{1}{K}\sum_{i=0}^{K-1}\lambda_i\right)\right)\right]\Bigg) + \tfrac{C_B}{\rho K}\sum_{i=0}^{K-1}\eta_i$$

$$\leq \tfrac{m}{K}\sum_{i=0}^{K-1} \left(\tfrac{6\epsilon_i\eta_i^b}{\rho} + 3m^2\eta_i^2\beta^2\right) + \tfrac{6mC_D}{\rho K} + \tfrac{C_B}{\rho K}\sum_{i=0}^{K-1}\eta_i \triangleq \tfrac{\tilde{C}_d}{K},$$

where $\tilde{C}_d$ and C_B are constants. By Jensen's inequality, $\mathbb{E}\left[(d\_(g(\bar{\mathbf{x}}_K)))\right] \leq \sqrt{\mathbb{E}\left[(d\_(g(\bar{\mathbf{x}}_K)))^2\right]} \triangleq \sqrt{\frac{\tilde{C}_d}{K}}$.

□

Proposition 11 (Rate on primal sub-optimality) Consider the sequence $\{(\mathbf{x}_k, \lambda_k)\}$ generated by (**VR-AL**). Suppose $\rho_k = \rho$ and $M_k = \frac{1}{k^{2+\delta}}$ for $k \geq 0$ and $\delta > 0$. If $\bar{\mathbf{x}}_K = \frac{\sum_{i=0}^{K-1} \mathbf{x}_i}{K}$, for any $K > 0$,

$$f(\mathbf{x}^*) - \mathbb{E}[f(\bar{\mathbf{x}}_K)] \leq \eta_K \beta + \tfrac{\rho \tilde{C}_d}{2K} + \tfrac{b_{\lambda,\eta}\sqrt{\tilde{C}_d}}{\sqrt{K}} \leq \tfrac{C_f}{\sqrt{K}}$$
$$f(\mathbf{x}^*) - \mathbb{E}[f(\bar{\mathbf{x}}_K)] \geq -\tfrac{\|\lambda_0\|^2}{2\rho K} - \tfrac{1}{K}\sum_{k=0}^{K-1}\left((B_\lambda + b_\lambda)\tfrac{\nu_G}{\sqrt{M_k}} + \tfrac{(\rho+1)\nu_G^2}{M_k} + \epsilon_k \eta_k^b + \eta_k \beta\right) \geq -\tfrac{\tilde{C}_f}{K}$$

where $C_f, \tilde{C}_f$ are non-negative constants.

Proof. Recall that since $\mathbf{x}_k$ is not necessarily feasible with respect to the constraints, we derive upper and lower bounds on the sub-optimality. Let $\mathbf{x}^* \in \arg\min \mathcal{L}_\rho(\mathbf{x}, \lambda^*)$ and $\mathbf{x}^*_{\eta_k} \in \arg\min \mathcal{L}_{\eta_k,\rho}(\mathbf{x}, \lambda^*_{\eta_k})$.
(i) *Lower bound.* A rate statement for the lower bound can be constructed as follows. Since $\max_\lambda \mathcal{D}_\rho(\lambda) = \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_\rho(\mathbf{x}, \lambda^*) = f^*$, we have the following sequence of inequalities.

$$\begin{aligned} f_{\eta_K}(\mathbf{x}^*_{\eta_K}) &\leq \mathcal{L}_{\eta_K,\rho}(\bar{\mathbf{x}}_K, \lambda^*_{\eta_K}) = f_{\eta_K}(\bar{\mathbf{x}}_K) + \tfrac{\rho}{2}\left(d_-\left(\tfrac{\lambda^*_{\eta_K}}{\rho} + g_{\eta_K}(\bar{\mathbf{x}}_K)\right)\right)^2 - \tfrac{1}{2\rho}\|\lambda^*_{\eta_K}\|^2 \\ &\leq f_{\eta_K}(\bar{\mathbf{x}}_K) + \tfrac{\rho}{2}\left(d_-(g_{\eta_K}(\bar{\mathbf{x}}_K))^2 + 2\left\|\tfrac{\lambda^*_{\eta_K}}{\rho}\right\| d_-(g_{\eta_K}(\bar{\mathbf{x}}_K))\right). \end{aligned}$$

Taking expectations on both sides, we obtain

$$\begin{aligned} f_{\eta_K}(\mathbf{x}^*_{\eta_K}) - \mathbb{E}[f_{\eta_K}(\bar{\mathbf{x}}_K)] &= \tfrac{\rho}{2}\mathbb{E}\left[(d_-(g_{\eta_K}(\bar{\mathbf{x}}_K)))^2\right] + \mathbb{E}\left[\|\lambda^*_{\eta_K}\| d_-(g_{\eta_K}(\bar{\mathbf{x}}_K))\right] \\ &\leq \tfrac{\rho}{2}\mathbb{E}\left[(d_-(g_{\eta_K}(\bar{\mathbf{x}}_K)))^2\right] + b_{\lambda,\eta}\mathbb{E}[d_-(g_{\eta_K}(\bar{\mathbf{x}}_K))] \leq \tfrac{\rho\tilde{C}_d}{2K} + \tfrac{b_{\lambda,\eta}\sqrt{\tilde{C}_d}}{\sqrt{K}} \\ \implies f(\mathbf{x}^*) - \mathbb{E}[f(\bar{\mathbf{x}}_K)] &= \mathbb{E}\Big[\underbrace{f(\mathbf{x}^*) - f(\mathbf{x}^*_{\eta_K})}_{\leq 0} + \underbrace{f(\mathbf{x}^*_{\eta_K}) - f_{\eta_K}(\mathbf{x}^*_{\eta_K})}_{\leq \eta_K\beta} + \left(f_{\eta_K}(\mathbf{x}^*_{\eta_K}) - f_{\eta_K}(\bar{\mathbf{x}}_K)\right) \\ &\quad + \underbrace{f_{\eta_K}(\bar{\mathbf{x}}_K) - f(\bar{\mathbf{x}}_K)}_{\leq 0}\Big] \leq \eta_K\beta + \tfrac{\rho\tilde{C}_d}{2K} + \tfrac{b_{\lambda,\eta}\sqrt{\tilde{C}_d}}{\sqrt{K}} \leq \tfrac{C_f}{\sqrt{K}} \end{aligned}$$

(ii) *Upper bound.* Let $\mathbf{x}^*_{\eta_k,\lambda_k} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\eta_k,\rho}(\mathbf{x}, \lambda_k)$. Based on the definition of $\mathbf{x}^*_{\eta_k,\lambda_k}$ and $\mathbf{x}^*_{\eta_k}$,

$$\begin{aligned} &\mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k,\lambda_k}, \lambda_k) \leq \epsilon_k \eta_k^b \text{ and } \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k,\rho_k}, \lambda_k) \leq \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k}, \lambda_k) \\ \implies &\mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k,}, \lambda_k) \leq \epsilon_k \eta_k^b. \end{aligned}$$

By Lemma 1, we have that

$$\mathbb{E}\left[f_{\eta_k}(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}^*_{\eta_k})\right] \leq \mathbb{E}\left[\tfrac{\rho}{2}\left(d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)^2 - \tfrac{\rho}{2}\left(d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2\right] + \epsilon_k \eta_k^b.$$

We observe that $d_-(u) = \|\Pi_-(u) - u\| = \|\Pi_+(u)\|$. It follows that

$$\begin{aligned} d_-\left(\tfrac{\lambda_{k+1}}{\rho}\right) = d_-\left(\Pi_+\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right) + w_k\right) &\leq d_-\left(\Pi_+\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right) + \|w_k\| \\ &\leq d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right) + \|w_k\| \\ \implies -\left(d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2 &\leq -\left\|\tfrac{\lambda_{k+1}}{\rho}\right\|^2 + 2d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\|w_k\| + \|w_k\|^2. \end{aligned}$$

Furthermore, $d_-\left(\frac{\lambda_k}{\rho}+g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right) \le \underbrace{d_-\left(g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)}_{=0,\text{ since } g_{\eta_k}(\mathbf{x}^*_{\eta_k})\le 0} + d_-\left(\frac{\lambda_k}{\rho}\right) = \left\|\frac{\lambda_k}{\rho}\right\|$. This allows us to claim that

$$\mathbb{E}\left[f_{\eta_k}(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}^*_{\eta_k})\right] \le \frac{\rho}{2}\mathbb{E}\left[\left\|\frac{\lambda_k}{\rho}\right\|^2 - \left\|\frac{\lambda_{k+1}}{\rho}\right\|^2\right] + \rho\mathbb{E}\left[d_-\left(\frac{\lambda_k}{\rho}+g(\mathbf{x}_{k+1})\right)\|w_k\| + \|w_k\|^2\right] + \epsilon_k\eta_k^b.$$

In addition, we have that

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right] &= \frac{1}{K}\sum_{k=0}^{K-1}\underbrace{\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}_{k+1})\right]}_{\le \eta_k\beta} + \mathbb{E}\left[f_{\eta_k}(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}^*_{\eta_k})\right] \\ &\quad + \frac{1}{K}\sum_{k=0}^{K-1}\underbrace{\mathbb{E}\left[f_{\eta_k}(\mathbf{x}^*_{\eta_k}) - f_{\eta_k}(\mathbf{x}^*)\right]}_{\le 0} + \underbrace{\mathbb{E}\left[f_{\eta_k}(\mathbf{x}^*) - f(\mathbf{x}^*)\right]}_{\le 0,\text{ since smoothness}}.\end{aligned}$$

Consequently, summing from $k=0$ to $K-1$, there exists a constant $\tilde{C}_f$ such that the following holds.

$$\begin{aligned}&\mathbb{E}[f(\bar{\mathbf{x}}_K) - f(\mathbf{x}^*)] \le \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right] \le \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(f_{\eta_k}(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}^*_{\eta_k})\right) + \eta_k\beta\right] \\ &\le \frac{\|\lambda_0\|^2 - \|\lambda_K\|^2}{2\rho K} + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\rho d_-\left(\frac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\|w_k\|\right] + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|w_k\|^2 + \epsilon_k\eta_k^b + \eta_k\beta\right] \\ &\le \frac{\|\lambda_0\|^2}{2\rho K} + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(\|\lambda_{k+1}\| + \rho\|w_k\|\right)\|w_k\|\right] + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|w_k\|^2 + \epsilon_k\eta_k^b + \eta_k\beta\right] \\ &\le \frac{\|\lambda_0\|^2}{2\rho K} + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\lambda_{k+1}\|\|w_k\| + (\rho+1)\|w_k\|^2 + \epsilon_k\eta_k^b + \eta_k\beta\right] \\ &\le \frac{\|\lambda_0\|^2}{2\rho K} + \frac{1}{K}\sum_{k=0}^{K-1}\left((B_\lambda + b_\lambda)\frac{\nu_G}{\sqrt{M_k}} + \frac{(\rho+1)\nu_G^2}{M_k} + \epsilon_k\eta_k^b + \eta_k\beta\right) \le \frac{\tilde{C}_f}{K}.\end{aligned}$$

□

3.3 Rate Analysis Under Increasing ρ_k

We now briefly describe the setting where $\{\rho_k\}$ is an increasing sequence. If $\eta_k = \mathcal{O}(1/\rho_k)$ and $M_k = \mathcal{O}(1/\rho_k^2)$, then it can be shown that the expected suboptimality and infeasibility diminish at the rate of $\mathcal{O}(1/\rho_k)$. **Proof Sketch.** (i) By the update rule of λ_{k+1}, $d_-\left(g_{\eta_k}(\mathbf{x}_{k+1})\right) \le d_-\left(\Pi_-\left(\frac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right) + \left\|\frac{\lambda_{k+1}-\lambda_k}{\rho_k}\right\| + \|w_k\| = \left\|\frac{\lambda_{k+1}-\lambda_k}{\rho_k}\right\| + \|w_k\|$. Therefore, $d_-\left(g(\mathbf{x}_{k+1})\right) \le d_-(g_{\eta_k}(\mathbf{x}_{k+1}) + \eta_k B\mathbf{1}) \le d_-(g_{\eta_k}(\mathbf{x}_{k+1})) + \eta_k B\|\mathbf{1}\| + \left\|\frac{\lambda_{k+1}-\lambda_k}{\rho_k}\right\| + \|w_k\|$, and then we derive the results by taking expectations on both sides.

(ii) The key steps are as follows, with the remaining proofs similar to those outlined in Proposition 11.

$$\begin{aligned}&f_{\eta_k}(\mathbf{x}^*_{\eta_k}) \le \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda^*_{\eta_k}) \le f_{\eta_k}(\mathbf{x}_{k+1}) + \frac{\rho_k}{2}\left(\frac{\|\lambda_{k+1}\|}{\rho_k} + \|w_k\| + \frac{\|\lambda^*_{\eta_k} - \lambda_k\|}{\rho_k}\right)^2 - \frac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2; \\ &f_{\eta_k}(\mathbf{x}_{k+1}) \le f_{\eta_k}(\mathbf{x}^*_{\eta_k}) + \frac{\rho_k}{2}\left(d_-\left(\left(\frac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)\right)^2 - \frac{\rho_k}{2}\left(d_-\left(\frac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2 + \epsilon_k\eta_k^b.\end{aligned}$$

4 CONCLUDING REMARKS

In this paper, we present an efficient inexact sampling-enabled AL framework for convex programs with possibly nonsmooth expectation-valued constraints. By overlaying a smoothing framework with diminishing smoothing parameters and a constant penalty parameter, we derive rate guarantees for dual suboptimality, primal suboptimality, and primal infeasibility. Future work will consider developing an overall complexity analysis with extensions to compositional constraints. The latter can be accommodated with relative ease since much of the convergence analysis persists but the overall complexity analysis is impacted by the emergence of the multi-level compositional term in the objective of the augmented Lagrangian subproblem.

ACKNOWLEDGMENTS

We are grateful for the support provided by ONR and DOE under grants N00014-22-1-2589 and DE-SC0023303, respectively.

REFERENCES

Aybat, N. S. and G. Iyengar. 2013. "An augmented Lagrangian method for conic convex programming". *arXiv:1302.6322*.

Beck, A. and M. Teboulle. 2012. "Smoothing and first order methods: A unified framework". *SIAM Journal on Optimization* 22(2):557–580.

Hestenes, M. R. 1969. "Multiplier and gradient methods". *Journal of Optimization Theory and Applications* 4(5):303–320.

Jalilzadeh, A., U. V. Shanbhag, J. Blanchet, and P. W. Glynn. 2022. "Smoothed variable sample-size accelerated proximal methods for nonsmooth stochastic convex programs". *Stochastic Systems* 12(4):373–410.

Lan, G. and R. D. Monteiro. 2013. "Iteration-complexity of first-order penalty methods for convex programming". *Mathematical Programming* 138(1):115–139.

Lan, G. and Z. Zhou. 2020. "Algorithms for stochastic optimization with function or expectation constraints". *Computational Optimization and Applications* 76(2):461–498.

Murtagh, B. and M. Saunders. 1978. "Large-scale linearly constrained optimization". *Mathematical Programming: Series A and B* 14(1):41–72.

Necoara, I., A. Patrascu, and F. Glineur. 2019. "Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming". *Optimization Methods and Software* 34(2):305–335.

Powell, M. J. D. 1969. "A method for nonlinear constraints in minimization problems". *Optimization*:283–298.

Rockafellar, R. T. 1973. "A dual approach to solving nonlinear programming problems by unconstrained optimization". *Mathematical programming* 5(1):354–373.

Rockafellar, R. T. 1976. "Augmented Lagrangians and applications of the proximal point algorithm in convex programming". *Mathematics of Operations Research* 1(2):97–116.

Xu, Y. 2021. "Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming". *Mathematical Programming* 185(1):199–244.

Yan, Y. and Y. Xu. 2022. "Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs". *Mathematical Programming Computation* 14(2):319–363.

Zhang, L., Y. Zhang, J. Wu, and X. Xiao. 2022. "Solving stochastic optimization with expectation constraints efficiently by a stochastic augmented lagrangian-type algorithm". *INFORMS Journal on Computing* 34(6):2989–3006.

Zhang, L., Y. Zhang, X. Xiao, and J. Wu. 2023. "Stochastic approximation proximal method of multipliers for convex stochastic programming". *Mathematics of Operations Research* 48(1):177–193.

Zhang, P., U. V. Shanbhag, and E. X. Fang. 2024. "A smoothed augmented Lagrangian framework for convex optimization with nonsmooth constraints". https://peixuanz.netlify.app/publication/a-smoothed-augmented-lagrangian-framework-for-convex-optimization-with-nonsmooth-constraints/a-smoothed-augmented-lagrangian-framework-for-convex-optimization-with-nonsmooth-constraints.pdf, accessed 17^{th} August 2024.

AUTHOR BIOGRAPHIES

PEIXUAN ZHANG is a Ph.D. student in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Penn State University. She primarily works on stochastic optimization, convex optimization and her email is pqz5090@psu.edu.

UDAY V. SHANBHAG is a Professor in the Department of Industrial and Operations Engineering at University of Michigan, Ann Arbor. His email is udaybag@umich.edu and his website is https://udaybag2.github.io.