A ZEROTH-ORDER PRIMAL-DUAL METHOD FOR CONSTRAINED BLACK-BOX PROBLEM WITH BLACK-BOX CONVEX CONSTRAINTS

Hongcheng Dong[⋆†] *Licheng Zhao*[⋆] *Wenqiang Pu*[⋆†] *Rui Zhou*[⋆†] *Feng Yin*[†]

[⋆] Shenzhen Research Institute of Big Data, Shenzhen, China
[†] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

We investigate zeroth-order constrained optimization (ZOCO) problems with a nonconvex objective and convex constraints. Both objective and constraint functions are black-box and only function values are available. Existing methods are often limited by high iteration complexity or computationally intensive operations. To address these limitations, we adopt a mini-batch two-point gradient estimator and perform gradient descent on a proximal Lagrangian function. We establish convergence to a stochastic ϵ-stationary point with $\mathcal{O}(\epsilon^{-2})$ iteration complexity. Numerical experiments on non-convex QCQP demonstrate that our proposed method achieves better computational efficiency in terms of fewer iterations and less running time compared to state-of-the-art baselines.

Index Terms— Nonconvex Constrained Optimization, Zeroth-Order Optimization, Primal-Dual Method

1. INTRODUCTION

Zeroth-order (ZO) optimization [1, 2] is a gradient-free optimization framework. It solely relies on function-value queries because the gradient is unavailable or expensive to compute. In unconstrained settings, ZO gradient estimators using Gaussian or uniform spherical distribution [3, 4, 5] have been proposed. When the problem involves black-box constraints, it falls under the category of ZO constrained optimization (ZOCO). The significance of ZOCO [6, 7] has been recognized in many engineering applications, including deep neural network training [8], online sensor management [9], and hyperparameter tuning [10]. The formal formulation of ZOCO can be expressed as follows:

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) \quad \text{s.t.} \quad h_i(\boldsymbol{x}) \le 0, \quad i = 1, \dots, m, \tag{1}$$

where the feasible set $\mathcal{X} \subseteq \mathbb{R}^n$ is convex and compact[1], the objective $f(\boldsymbol{x})$ is a black-box function and possibly nonconvex, and each constraint $h_i(\boldsymbol{x})$ is a black-box convex function.

However, classical ZO methods such as [11, 12, 13] are inapplicable to ZOCO because they cannot maintain feasibility under black-box constraints. The authors of [10, 14] tackle ZOCO using conditional gradient (Frank-Wolfe) descent and maintain feasibility via implicit projections, but these methods only address linear black-box constraints. Conversely, for problems with nonlinear black-box constraints, several primal-dual-based approaches have been developed. Following the alternate primal-dual update scheme, ZO-ALM [15] handles nonlinear equality constraints via an augmented Lagrangian method with $\mathcal{O}(\epsilon^{-2.5})$ iteration complexity, while DSZOG [16] tackles nonlinear inequality constraints using a min-max penalty formulation that attains $\mathcal{O}(\epsilon^{-4})$ iteration complexity. In contrast, ZO-ConEX [17] represents a specific primal-dual method where the inequality constraints are dealt with by a constraint-extrapolation step, leading to an $\mathcal{O}(\epsilon^{-3})$ iteration complexity. To perform dual variable update, ZO-ConEX calculates an auxiliary variable by constructing a surrogate function, which is a computationally intensive operation. A detailed comparison of these approaches is provided in Table 1.

A clear gap remains in the ZOCO literature: whether there exists a primal-dual method that can achieve low iteration complexity and low per-iteration computational cost simultaneously. To overcome this limitation, we introduce a proximal primal-dual method. It performs gradient descent over a proximal Lagrangian function, in which a mini-batch two-point ZO gradient estimator can be directly applied. Our main contributions are: 1) we propose an algorithm named ZO-SPLM that is broadly applicable to general nonlinear convex constraints under the black-box setting, and 2) under standard smoothness and regularity assumptions, ZO-SPLM reaches an ϵ-stationary point with a reduced iteration complexity of $\mathcal{O}(\epsilon^{-2})$ compared withexisting ZOCO methods.

2. THE ZO-SPLM ALGORITHM

Recall the nonlinear constrained problem in (1). The proximal Lagrangian function for the problem in (1) is defined as:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{y}) := f(\boldsymbol{x}) + \langle \boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})\rangle + \frac{p}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2, \tag{2}$$

where $\boldsymbol{h}(\boldsymbol{x}) := (h_1(\boldsymbol{x}), \dots, h_m(\boldsymbol{x}))^\top$, $\boldsymbol{y} \in \mathbb{R}^m_+$ is the dual variable, $\boldsymbol{z} \in \mathbb{R}^n$ is the auxiliary proximal variable, and $p > 0$ is the proximal parameter.

Since the proximal Lagrangian function $\mathcal{L}$ consists of the black-box objective $f(\boldsymbol{x})$ and constraint functions $\boldsymbol{h}(\boldsymbol{x})$, the partial derivative $\nabla_{\boldsymbol{x}}\mathcal{L}$ is unavailable. To obtain a descent direction, we approximate the gradient using a mini-batch two-point estimator. Specifically, given a mini-batch size k and a smoothing radius $r > 0$, we draw independent directions $\{\boldsymbol{\Delta}_i\}_{i=1}^k$, each from one of the following distributions:

- `Uniform`: $\boldsymbol{\Delta}_i \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, where $\mathrm{Unif}(\mathbb{S}^{n-1})$ is the uniform distribution on the unit sphere,
- `Gaussian`: $\boldsymbol{\Delta}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n/n)$, where $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n/n)$ is the Gaussian distribution.

With k independent and identically distributed samples, the mini-batch two-point estimator is given as:

$$\widehat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{y}) = \frac{n}{rk}\sum_{i=1}^{k}\big[\mathcal{L}(\boldsymbol{x}+r\boldsymbol{\Delta}_i, \boldsymbol{z}; \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{y})\big]\boldsymbol{\Delta}_i. \tag{3}$$

[1] In this work, we consider $\mathcal{X}$ is known and easy to project onto.

Table 1: Comparison of Frameworks for ZO Nonlinear Constrained Optimization

Algorithm	**Constraint**	**Framework**	**Scheme Type**	**Iteration Complexity**
ZO-ALM	Nonlinear Equality	Augmented Lagrangian	Primal-Dual	$\mathcal{O}(\epsilon^{-2.5})$
DSZOG	Nonlinear Inequality	Min-Max Penalty	Primal-Dual	$\mathcal{O}(\epsilon^{-4})$
ZO-ConEX	Nonlinear Inequality	Constraint Extrapolation	Specific Primal-Dual	$\mathcal{O}(\epsilon^{-3})$

Note that the stochastic gradient estimator is biased, and the bias and variance are bounded above [6]:

$$\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y}) - \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y})\|\big] \le L_{\mathcal{L}}r, \tag{4}$$

$$\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y}) - \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y})\|^2\big] \le C_v L_{\mathcal{L}}^2 r^2/k, \tag{5}$$

where function $\mathcal{L}$ has an $L_{\mathcal{L}}$-Lipschitz continuous gradient with respect to $\boldsymbol{x}$. Here, $C_v = n^2/4$ for $\boldsymbol{\Delta}_i \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ and $C_v = (n+2)(n+4)/4$ for $\boldsymbol{\Delta}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n/n)$.

By integrating this stochastic gradient estimator into the primal-update step, we propose an algorithm named ZO-SPLM, outlined in Algorithm 1. ZO-SPLM performs primal, dual and proximal updates sequentially. Numerical stability of dual variable update is guaranteed by projection onto a bounded box $[0, B]^m$ with a fixed constant $B > 0$.

Algorithm 1 Zeroth-Order Smoothed Proximal Lagrangian Method (ZO-SPLM)

```
Input: step sizes $(c, \alpha, \beta)$, proximal parameter $p$, constant $B$, mini-batch size $k$, smoothing radius $r$
Initialize: $\boldsymbol{x}^0 \in \mathcal{X}$, $\boldsymbol{z}^0 = \boldsymbol{x}^0$, $\boldsymbol{y}^0 \in [0, B]^m$
for $t = 0, 1, 2, \dots$ do
    Draw $\boldsymbol{\Delta}_1, \dots, \boldsymbol{\Delta}_k$ from $\mathrm{Unif}(\mathbb{S}^{n-1})$ or $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n/n)$
    Compute the ZO gradient $\widehat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t, \boldsymbol{z}^t; \boldsymbol{y}^t)$ using (3)
    $\boldsymbol{x}^{t+1} = \Pi_{\mathcal{X}}\big(\boldsymbol{x}^t - c\widehat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t, \boldsymbol{z}^t; \boldsymbol{y}^t)\big)$    ▷ Primal update
    $\boldsymbol{y}^{t+1} = \Pi_{[0,B]^m}\big(\boldsymbol{y}^t + \alpha\, h(\boldsymbol{x}^{t+1})\big)$    ▷ Dual update
    $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t + \beta\big(\boldsymbol{x}^{t+1} - \boldsymbol{z}^t\big)$    ▷ Proximal update
end for
```

3. CONVERGENCE ANALYSIS

3.1. Assumptions and Stochastic ϵ-Stationarity

Convergence analysis is presented in this section. We first state the assumptions and the definition of stochastic ϵ-stationarity.

Assumption 1. *We make the following smoothness assumptions.*

- *The objective function $f(\boldsymbol{x})$ is continuously differentiable with an L_f-Lipschitz continuous gradient.*
- *Each constraint function $h_i(\boldsymbol{x})$ is convex and continuously differentiable with an L_{h_i}-Lipschitz continuous gradient.*

In the zeroth-order setting, the stochasticity in gradient estimation makes the exact Karush-Kuhn-Tucker (KKT) conditions unsuitable for convergence assessment. Therefore, the following definition of stochastic ϵ-stationarity is introduced.

Definition 1 (Stochastic ϵ-stationarity)**.** *A random point $\boldsymbol{x} \in \mathcal{X}$ is a stochastic ϵ-stationary point of problem* (1) *if there exists a dual variable $\boldsymbol{y} \in \mathbb{R}^m_+$ such that:*

$$\mathbb{E}_{\boldsymbol{x}}\Big[\mathrm{dist}\big(\boldsymbol{0},\ \nabla f(\boldsymbol{x}) + \nabla h(\boldsymbol{x})^\top \boldsymbol{y} + N_{\mathcal{X}}(\boldsymbol{x})\big)\Big] \ \le\ \epsilon, \tag{6}$$

$$\mathbb{E}_{\boldsymbol{x}}\big[\|\Pi_{\mathbb{R}^m_+}(h(\boldsymbol{x}))\|\big] \ \le\ \epsilon, \tag{7}$$

$$\mathbb{E}_{\boldsymbol{x}}\big[|\langle \boldsymbol{y}, h(\boldsymbol{x})\rangle|\big] \ \le\ \epsilon. \tag{8}$$

Here, $N_{\mathcal{X}}(\boldsymbol{x})$ is the normal cone of $\mathcal{X}$ at $\boldsymbol{x}$, $\mathrm{dist}(\boldsymbol{0}, \cdot)$ denotes the Euclidean distance from the origin to a set, and the expectation $\mathbb{E}_{\boldsymbol{x}}[\cdot]$ is over the randomness of $\boldsymbol{x}$. We define $\boldsymbol{x}$ to be a stationary point when $\epsilon = 0$, and let the set of all stationary points by $\mathcal{X}^*$.

Remark 1. *These three conditions are stochastic relaxations of the original KKT conditions, where* (6), (7), *and* (8) *correspond to stationarity, primal feasibility, and complementary slackness, respectively.*

In addition to Assumption 1, the convergence analysis relies on regularity conditions. We assume $\mathcal{X} := \{\boldsymbol{x} \in \mathbb{R}^n : g_i(\boldsymbol{x}) \le 0, i = 1, \dots, l\}$. For any point $\boldsymbol{x} \in \mathcal{X}$, define the active set as $A[\boldsymbol{x}] := \{i \in \{1, \dots, m\} : h_i(\boldsymbol{x}) = 0\} \cup \{j \in \{1, \dots, l\} : g_j(\boldsymbol{x}) = 0\}$. Then, define the Jacobian of all constraints by $J(\boldsymbol{x}) \in \mathbb{R}^{(m+l)\times n}$ and use $J_{A[\boldsymbol{x}]}(\boldsymbol{x})$ to represent the submatrix of $J(\boldsymbol{x})$ whose rows are indexed by the active set $A[\boldsymbol{x}]$.

Assumption 2. *We make the following regularity assumptions.*

- *There exists $\hat{\boldsymbol{x}} \in \mathcal{X}$ and $\Delta_0 > 0$ such that $h_i(\hat{\boldsymbol{x}}) \le -\Delta_0$ for all $i \in \{1, \dots, m\}$.*
- *The gradients of all subsets of the active constraints are uniformly linearly independent over the set of $\mathcal{X}^*$. That is, there exists a constant $\sigma_{\mathcal{X}^*} > 0$ given by*

  $$\sigma_{\mathcal{X}^*} := \inf_{\boldsymbol{x}^* \in \mathcal{X}^*,\, \mathcal{S}_A \subseteq A[\boldsymbol{x}^*]} \sigma_{\min}\big(J_{\mathcal{S}_A}(\boldsymbol{x}^*)^\top\big) > 0,$$

  where $\sigma_{\min}(\cdot)$ denotes the minimum singular value of a matrix.

3.2. Main Convergence Result

Our convergence analysis for Algorithm 1 is based on a potential function Φ^t, which is defined as:

$$\Phi^t := \mathbb{E}_{\boldsymbol{x}}\big[\mathcal{L}(\boldsymbol{x}^t, \boldsymbol{z}^t; \boldsymbol{y}^t) - 2d(\boldsymbol{y}^t, \boldsymbol{z}^t) + 2P(\boldsymbol{z}^t) \mid \mathcal{F}_t\big], \tag{9}$$

where $\mathcal{F}_t := \{\boldsymbol{x}^s, \boldsymbol{y}^s, \boldsymbol{z}^s, \forall s < t\}$,

$$d(\boldsymbol{y}^t, \boldsymbol{z}^t) := \min_{\boldsymbol{x}^t \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{z}^t; \boldsymbol{y}^t),$$

and

$$P(\boldsymbol{z}^t) := \max_{\boldsymbol{y}^t \in [0,B]^m} d(\boldsymbol{y}^t, \boldsymbol{z}^t).$$

The fundamental aspect of the analysis is establishing a descent property for Φ^t. Specifically, we show that it decreases in expectation at each step, except for an additional error term caused by the stochastic ZO gradient estimation. The following lemma formalizes this crucial result, which lays a solid foundation for our main convergence theorem.

Lemma 1 (Potential Descent)**.** *Suppose the Assumptions 1 and 2 hold, and the algorithm parameters are chosen as $B > \bar{B}$, where $\bar{B}, c, \alpha, \beta$ are some constants, define $\boldsymbol{y}_+^t(\boldsymbol{z}^t) := \Pi_{[0,B]^m}(\boldsymbol{y}^t + \alpha h(\boldsymbol{x}^t(\boldsymbol{y}^t, \boldsymbol{z}^t)))$ and $\boldsymbol{x}^t(\boldsymbol{y}^t, \boldsymbol{z}^t) = \arg\min_{\boldsymbol{x}^t \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{z}^t; \boldsymbol{y}^t)$, then for any $t \in \mathbb{N}$:*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\Phi^t - \Phi^{t+1} \mid \mathcal{F}_t] \geq& \mathbb{E}_{\boldsymbol{x}}\left[\frac{\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2}{32c} + \frac{\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2}{16\alpha}\right.\\ &\left. + \frac{p\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2}{16\beta} \,\middle|\, \mathcal{F}_t\right] - \frac{cC_v L_{\mathcal{L}}^2 r^2}{k},\end{aligned}$$

where $L_{\mathcal{L}} = L_f + B\sqrt{m \sum_{i=1}^m L_{h_i}^2} + p$.

Proof. We provide the sketch of the proof.

1. **Lagrangian Descent**: Using the $L_{\mathcal{L}}$-smoothness of $\mathcal{L}(\cdot, \boldsymbol{z}^t; \boldsymbol{y}^t)$ and the projected step, we obtain

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}}\Big[\mathcal{L}(\boldsymbol{x}^t, \boldsymbol{z}^t; \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^{t+1}, \boldsymbol{z}^{t+1}; \boldsymbol{y}^{t+1}) \mid \mathcal{F}_t\Big]\\ &\geq \mathbb{E}_{\boldsymbol{x}}\Big[\frac{1}{4c}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 + \langle h(\boldsymbol{x}^{t+1}), \boldsymbol{y}^t - \boldsymbol{y}^{t+1}\rangle\\ &\quad + \frac{p}{2\beta}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 \,\Big|\, \mathcal{F}_t\Big] - \frac{cC_v L_{\mathcal{L}}^2 r^2}{k}.\end{aligned}$$

2. **Dual Ascent**: The dual part $d(\boldsymbol{y}, \boldsymbol{z})$ (Lemma 10 in [18]) yields

$$d(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t) - d(\boldsymbol{y}^t, \boldsymbol{z}^t) \geq \frac{1}{8\alpha}\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2.$$

3. **Proximal Descent**: The strong convexity of $\mathcal{L}$ in $\boldsymbol{z}$ and the averaging step $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t + \beta(\boldsymbol{x}^{t+1} - \boldsymbol{z}^t)$ (Lemma 11 in [18]) imply

$$P(\boldsymbol{z}^t) - P(\boldsymbol{z}^{t+1}) \geq \frac{p}{8\beta}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2.$$

4. **Final Descent Inequality**: Summing the three bounds and choosing parameters (c, α, β) appropriately, we obtain the final descent inequality:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\Phi^t - \Phi^{t+1} \mid \mathcal{F}_t] \geq \mathbb{E}_{\boldsymbol{x}}\Big[&\frac{\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2}{32c} + \frac{\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2}{16\alpha}\\ &+ \frac{p\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2}{16\beta} \,\Big|\, \mathcal{F}_t\Big] - \frac{cC_v L_{\mathcal{L}}^2 r^2}{k}.\end{aligned}$$

Further details can be found in the supplementary material. □

Lemma 1 provides the key descent property for our potential function Φ^t. When the smoothing radius r is chosen sufficiently small or the mini-batch size k sufficiently large, a strict descent property of Φ^t can be guaranteed. By summing this descent inequality over all iterations, we can derive the iteration complexity stated in the following theorem.

Theorem 1. *Suppose the Assumptions 1 and 2 hold, and the algorithm parameters are chosen as $B > \bar{B}$, where $\bar{B}, c, \alpha, \beta$ are some constants. By selecting the smoothing radius $r = \epsilon/(2L_{\mathcal{L}})$ and the mini-batch size $k \geq c\,n^2 C_0$, Algorithm 1 finds an ϵ-stationary point within at most $\mathcal{O}(\epsilon^{-2})$ iterations.*

Proof. We provide a sketch of the proof.

1. **Potential Descent Bound**: Summing the descent inequality in Lemma 1 from $t = 0$ to $T-1$ and applying the lower bound $\Phi^t \geq \underline{f}$ (detailed in supplementary material), we bound the expected progress at the best iterate $t^* < T$:

$$\begin{aligned}\varrho_{t^*+1} :=& \max\{\mathbb{E}_{\boldsymbol{x}}\big[\|\boldsymbol{x}^{t^*} - \boldsymbol{x}^{t^*+1}\|\big], \mathbb{E}_{\boldsymbol{x}}\big[\|\boldsymbol{y}^{t^*} - \boldsymbol{y}_+^{t^*}(\boldsymbol{z}^{t^*})\|\big],\\ &\mathbb{E}_{\boldsymbol{x}}\big[\|\boldsymbol{z}^{t^*} - \boldsymbol{z}^{t^*+1}\|\big]\}\\ \leq& \sqrt{\frac{\Phi^0 - \underline{f}}{T\lambda_0} + \frac{cC_v L_{\mathcal{L}}^2 r^2}{k\lambda_0}},\end{aligned}$$

where $\lambda_0 = \min\{1/(32c), 1/(16\alpha), p/(16\beta)\}$.

2. **Stationarity Bound**: Combining Definition 1 with the gradient estimator error bounds (4) and (5), we obtain:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}}\left[\operatorname{dist}\Big(\boldsymbol{0}, \nabla f(\boldsymbol{x}^{t^*+1}) + \nabla h(\boldsymbol{x}^{t^*+1})^\top \boldsymbol{y}^{t^*} + N_{\mathcal{X}}(\boldsymbol{x}^{t^*+1})\Big)\right]\\ &\leq \lambda_1 \varrho_{t^*+1} + L_{\mathcal{L}} r.\end{aligned}$$

3. **Primal Feasibility and Complementary Slackness Bounds**: Following Lemma 5 of [18], we bound the primal feasibility and complementary slackness as:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\Big[\|\Pi_{\mathbb{R}_+^m}(h(\boldsymbol{x}^{t^*+1}))\|\Big] &\leq \lambda_2 \varrho_{t^*+1},\\ \mathbb{E}_{\boldsymbol{x}}\Big[|\langle \boldsymbol{y}^{t^*}, h(\boldsymbol{x}^{t^*+1})\rangle|\Big] &\leq \lambda_3 \varrho_{t^*+1}.\end{aligned}$$

4. **Iteration Complexity**: Setting $r = \epsilon/(2L_{\mathcal{L}})$ and $k \geq cn^2 C_0$, and combining all bounds above, we derive the required number of iterations:

$$T \geq \left\lceil \frac{C_1(\Phi^0 - \underline{f})}{\epsilon^2(C_2 - C_3/k)} \right\rceil = \mathcal{O}(\epsilon^{-2}).$$

The complete proof is provided in the supplementary material. □

Theorem 1 shows that our proposed ZO-SPLM achieves an iteration complexity of $\mathcal{O}(\epsilon^{-2})$, consistent with the lower bound for finding ϵ-stationary points [19]. Moreover, the gradient estimation error can be properly suppressed through an appropriate choice of smoothing radius r and mini-batch size k, and thus the optimal iteration complexity can be achieved.

4. NUMERICAL EXPERIMENTS

4.1. Experimental Setup

We evaluate the performance of the proposed ZO-SPLM based on the following on nonconvex quadratically constrained quadratic programming (NC-QCQP) problems:

$$\begin{aligned}\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{d}^\top \boldsymbol{x}\\ \text{s.t.} \quad & \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c \;\leq\; 0,\\ & \boldsymbol{\ell}_i \;\leq\; \boldsymbol{x}_i \;\leq\; \boldsymbol{u}_i, \quad j = 1, \ldots, n,\end{aligned} \tag{10}$$

where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is a symmetric matrix (not necessarily positive semidefinite), $\boldsymbol{A}$ is a positive semidefinite matrix, $\boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^n$, and $c < 0$. The box constraints are set to $\boldsymbol{\ell}_i = -10$ and $\boldsymbol{u}_i = 10$ for all $i = 1, \ldots, n$.

We evaluate the numerical performance by comparing ZO-SPLM against two state-of-the-art ZOCO methods:

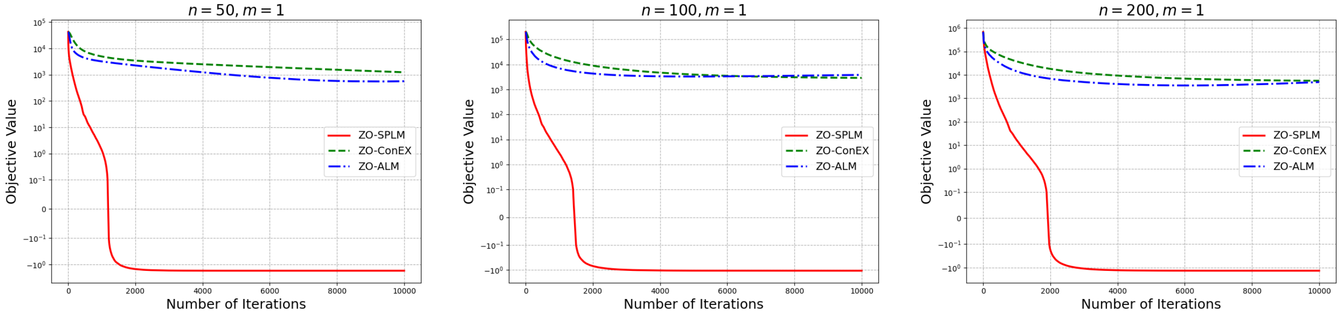

Fig. 1: Convergence of the objective value for trials with increasing problem dimension $n = 50, 100, 200$.

Table 2: Iterations (Iters), total running time (Total), and per-iteration running time (Per-iter) (averaged over 20 trials) to achieve stochastic ϵ-stationarity with $\epsilon < 1.0$ across growing dimensions n with mini-batch size $k = 10000$.

Method	$n = 50$			$n = 100$			$n = 200$		
	Iters	Total (s)	Per-iter (s)	Iters	Total (s)	Per-iter (s)	Iters	Total (s)	Per-iter (s)
ZO-SPLM	**496**	**204.35**	**0.412**	**1309**	**627.01**	**0.479**	**1962**	**1589.22**	**0.810**
ZO-ConEX	1374	971.42	0.707	9620	10697.44	1.112	14283	21510.06	1.506
ZO-ALM	2066	876.21	0.424	10850	5403.13	0.498	29073	26630.23	0.916

- **ZO-ConEX** [17]: A zeroth-order primal-dual method that employs a constraint extrapolation strategy to handle nonlinear inequality constraints.
- **ZO-ALM** [15]: A zeroth-order primal-dual method based on the augmented Lagrangian framework[2].

The NC-QCQP instances for our experiments are constructed as follows: The symmetric matrices are set as $\boldsymbol{Q} = \boldsymbol{M}^\top \boldsymbol{M} - \delta \boldsymbol{I}$ and $\boldsymbol{A} = \boldsymbol{M}_1^\top \boldsymbol{M}_1$, where $\boldsymbol{M}$ and $\boldsymbol{M}_1$ are sampled from a standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\delta > 0$ is chosen so that the minimum eigenvalue $\lambda_{\min}(\boldsymbol{Q}) = -0.1$. Meanwhile, the vectors $\boldsymbol{d}$ and $\boldsymbol{b}$ are sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and c is set to -1. The ZO gradient estimators for all methods use an identical setting: random directions are drawn from $\mathrm{Unif}(\mathbb{S}^{n-1})$, and the smoothing radius is set to $r = 10^{-5}$.

4.2. Results

Firstly, we compare the convergence curves of one realization for different problem sizes n in Figure 1, where the mini-batch size of $k = 50$ is used. The figures clearly demonstrate that ZO-SPLM converges faster than the baselines in all tested dimensions.

Then we record the iterations and running time required to reach stochastic ϵ-stationarity (cf. Definition 1), and assign a sufficiently large mini-batch size $k = 10,000$ to approximate the true mean value. The results in Table 2 confirm that ZO-SPLM converges with significantly fewer iterations than the baselines, and this performance improvement expands as the problem dimension increases.

Moreover, the per-iteration running time of ZO-SPLM is less than ZO-ConEX, which has verified that the proposed method avoids the intensive calculation of the auxiliary variable. Although ZO-ConEX takes fewer iterations before termination, the total running time is longer than ZO-ALM.

5. CONCLUSION

In this work, we have studied the ZOCO problem with a non-convex objective and convex constraints. We observe that existing methods often suffer from high iteration complexity or computationally intensive operations. To address these drawbacks, we have utilized a mini-batch two-point gradient estimator and performed gradient descent on a proximal Lagrangian function. ZO-SPLM is provably convergent to a stochastic ϵ-stationary point with an iteration complexity of $\mathcal{O}(\epsilon^{-2})$ under standard smoothness and regularity assumption. Numerical experiments on NC-QCQP have shown that ZO-SPLM attains the stochastic ϵ-stationarity consuming fewer iterations and less running time compared to the representative baselines, and the performance improvement scales with the problem dimension.

6. REFERENCES

[1] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C.J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[2] S. Liu, B. Kailkhura, P. Chen, P. Ting, S. Chang, and L. Amini, "Zeroth-order stochastic variance reduction for nonconvex optimization," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[3] J.C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," in *IEEE Transactions on Automatic Control*. IEEE, 2002, vol. 37, pp. 332–341.

[4] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," in *Foundations of Computational Mathematics*. Springer, 2017, vol. 17, pp. 527–566.

[2] Although ZO-ALM was designed for ZOCO with nonlinear equality constraints, we modify it by relaxing the inequality constraints for fair comparison.

[5] A.D. Flaxman, A.T. Kalai, and H.B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Vancouver, British Columbia, 2005, ACM-SIAM, SODA '05, pp. 385–394.

[6] S. Liu, P. Chen, B. Kailkhura, G. Zhang, A.O. Hero III, and P.K. Varshney, "A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications," in *IEEE Signal Processing Magazine*. IEEE, 2020, vol. 37, pp. 43–54.

[7] Z. Liu, C. Chen, L. Luo, and B.K.H. Low, "Zeroth-order methods for constrained nonconvex nonsmooth stochastic optimization," in *Forty-first International Conference on Machine Learning*, 2024.

[8] A. Chen, Y. Zhang, J. Jia, J. Diffenderfer, K. Parasyris, J. Liu, Y. Zhang, Z. Zhang, B. Kailkhura, and S. Liu, "Deepzero: Scaling up zeroth-order optimization for deep model training," in *ICLR*, 2024.

[9] S. Liu, J. Chen, P. Chen, and A. Hero, "Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 288–297.

[10] K. Balasubramanian and S. Ghadimi, "Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points," in *Foundations of Computational Mathematics*. Springer, 2022, vol. 22, pp. 35–76.

[11] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," in *Mathematical Programming*. Springer, 2016, vol. 155, pp. 267–305.

[12] J.C. Duchi, M.I. Jordan, M.J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," in *IEEE Transactions on Information Theory*. IEEE, 2015, vol. 61, pp. 2788–2806.

[13] X. Chen, S. Liu, K. Xu, X. Li, X. Lin, M. Hong, and D. Cox, "Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.

[14] K. Balasubramanian and S. Ghadimi, "Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[15] Z. Li, P. Chen, S. Liu, S. Lu, and Y. Xu, "Zeroth-order optimization for composite problems with functional constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 7453–7461.

[16] W. Shi, H. Gao, and B. Gu, "Gradient-free method for heavily constrained nonconvex optimization," in *International Conference on Machine Learning*, 2022, pp. 19935–19955.

[17] A. Nguyen and K. Balasubramanian, "Stochastic zeroth-order functional constrained optimization: Oracle complexity and applications," in *INFORMS Journal on Optimization*. INFORMS, 2023, vol. 5, pp. 256–272.

[18] W. Pu, K. Sun, and J. Zhang, "Smoothed proximal lagrangian method for nonlinear constrained programs," in *arXiv preprint arXiv:2408.15047*, 2024.

[19] C. Cartis, N.I.M. Gould, and P.L. Toint, "On the complexity of finding first-order critical points in constrained nonlinear optimization," in *Mathematical Programming*, Germany, 2014, vol. 144 of *Ser. A*, pp. 93–106, Springer.

[20] J. Zhang, P. Xiao, R. Sun, and Z. Luo, "A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave minmax problems," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 7377–7389.

A. PROOF OF THE ZEROTH-ORDER ESTIMATOR BOUNDS

Throughout let

$$\widehat{\nabla}\phi(\boldsymbol{x}) \;=\; \frac{n}{rk}\sum_{i=1}^{k}\big[\phi(\boldsymbol{x}+r\boldsymbol{\Delta}_i)-\phi(\boldsymbol{x})\big]\boldsymbol{\Delta}_i, \qquad \boldsymbol{\Delta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{D},$$

where the perturbation distribution $\mathcal{D} = \mathrm{Unif}(\mathbb{S}^{n-1})$ or $\mathcal{N}(\mathbf{0}, \frac{1}{n}\boldsymbol{I}_n)$ satisfies $\mathbb{E}_{\boldsymbol{\Delta}}[\boldsymbol{\Delta}] = \mathbf{0}$ and $\mathbb{E}_{\boldsymbol{\Delta}}[\boldsymbol{\Delta}\boldsymbol{\Delta}^\top] = \frac{1}{n}\boldsymbol{I}_n$.

Then we can get

$$\mathbb{E}\big[\widehat{\nabla}\phi(\boldsymbol{x})\big] = \frac{n}{r}\,\mathbb{E}_{\boldsymbol{\Delta}}\big[\big(\phi(\boldsymbol{x}+r\boldsymbol{\Delta})-\phi(\boldsymbol{x})\big)\boldsymbol{\Delta}\big] = \nabla\phi_r(\boldsymbol{x}),$$
$$\phi_r(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\Delta}}\big[\phi(\boldsymbol{x}+r\boldsymbol{\Delta})\big].$$

If $\nabla\phi$ is L_ϕ-Lipschitz, then by a first-order expansion of $\nabla\phi$ along $\{\boldsymbol{x}+s\boldsymbol{\Delta} : s\in[0,r]\}$,

$$\big\|\nabla\phi_r(\boldsymbol{x})-\nabla\phi(\boldsymbol{x})\big\| \le \mathbb{E}_{\boldsymbol{\Delta}}\int_0^r \big\|\nabla\phi(\boldsymbol{x}+s\boldsymbol{\Delta})-\nabla\phi(\boldsymbol{x})\big\|\,\mathrm{d}s \le L_\phi r,$$

and hence

$$\mathbb{E}\big[\|\widehat{\nabla}\phi(\boldsymbol{x})-\nabla\phi(\boldsymbol{x})\|\big] \le L_\phi r,$$

which is (4).

Define $g_i := \dfrac{n}{r}\big[\phi(\boldsymbol{x}+r\boldsymbol{\Delta}_i)-\phi(\boldsymbol{x})\big]\boldsymbol{\Delta}_i$ so that $\widehat{\nabla}\phi(\boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^{k} g_i$ and $\mathbb{E}[g_i] = \nabla\phi_r(\boldsymbol{x})$.

Independence implies

$$\mathbb{E}\big[\|\widehat{\nabla}\phi(\boldsymbol{x})-\nabla\phi_r(\boldsymbol{x})\|^2\big] = \frac{1}{k}\,\mathbb{E}\big[\|g_1-\nabla\phi_r(\boldsymbol{x})\|^2\big] \le \frac{1}{k}\,\mathbb{E}\big[\|g_1\|^2\big].$$

Using the mean-value theorem and the L_ϕ-smoothness upper quadratic bound,

$$\big|\phi(\boldsymbol{x}+r\boldsymbol{\Delta})-\phi(\boldsymbol{x})\big| \le \frac{L_\phi}{2}r^2\|\boldsymbol{\Delta}\|^2, \Longrightarrow \|g_1\| \le \frac{n}{2}L_\phi r\,\|\boldsymbol{\Delta}\|^3.$$

Therefore

$$\begin{aligned}\mathbb{E}\big[\|g_1\|^2\big] &\le \frac{n^2}{4}L_\phi^2 r^2\,\mathbb{E}\big[\|\boldsymbol{\Delta}\|^6\big],\\ \mathbb{E}\big[\|\widehat{\nabla}\phi(\boldsymbol{x})-\nabla\phi_r(\boldsymbol{x})\|^2\big] &\le \frac{n^2}{4k}L_\phi^2 r^2\,\mathbb{E}\big[\|\boldsymbol{\Delta}\|^6\big].\end{aligned} \tag{11}$$

If $\boldsymbol{\Delta} \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, then $\|\boldsymbol{\Delta}\| \equiv 1$ and thus $\mathbb{E}\|\boldsymbol{\Delta}\|^6 = 1$. From (11),

$$\mathbb{E}\big[\|\widehat{\nabla}\phi(\boldsymbol{x})-\nabla\phi(\boldsymbol{x})\|^2\big] \;\le\; \frac{n^2L_\phi^2r^2}{4k},$$

which is exactly (5) with $C_v = n^2/4$.

If $\boldsymbol{\Delta} \sim \mathcal{N}(\mathbf{0}, \frac{1}{n}\boldsymbol{I}_n)$, then $\|\boldsymbol{\Delta}\|^2 = \frac{1}{n}\chi_n^2$ and the χ^2-moment formula $\mathbb{E}[(\chi_n^2)^3] = n(n+2)(n+4)$ yields

$$\mathbb{E}\|\boldsymbol{\Delta}\|^6 = \mathbb{E}\Big[\Big(\tfrac{1}{n}\chi_n^2\Big)^3\Big] = \frac{(n+2)(n+4)}{n^2}.$$

Plugging into (11) gives

$$\mathbb{E}\big[\|\widehat{\nabla}\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x})\|^2\big] \;\le\; \frac{(n+2)(n+4)}{4k}\, L_\phi^2 r^2,$$

i.e., (5) with $C_v = (n+2)(n+4)/4$.

B. PROOF OF LEMMA 1

For brevity, we denote $\mathbb{E}_{\boldsymbol{x}}[\cdot]$ simply as $\mathbb{E}[\cdot]$ throughout the paper.

First, under Assumption 1, we define constants capturing problem structure: $\nabla_f := \max_{\boldsymbol{x}\in\mathcal{X}} |\nabla f(\boldsymbol{x})|$; $D_{\mathcal{X}} := \max_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{X}} |\boldsymbol{u} - \boldsymbol{v}|$; $M_h := \max_{\boldsymbol{x}\in\mathcal{X}} |h(\boldsymbol{x})|$; and $K_h := (\sum_{i=1}^m K_{h_i}^2)^{1/2}$, where each K_{h_i} denotes the upper bound on the gradient norm of the constraint h_i. let $h(\boldsymbol{x}) := (h_1(\boldsymbol{x}),\ldots,h_m(\boldsymbol{x}))^\top \in \mathbb{R}^m$, $\nabla h(\boldsymbol{x}) := [\nabla h_1(\boldsymbol{x}),\ldots,\nabla h_m(\boldsymbol{x})] \in \mathbb{R}^{n\times m}$, and $L_h := \sqrt{\sum_{i=1}^m L_{h_i}^2}$.

The following Lemmas will be useful in our analysis.

Lemma 2 (Lemma 1 of [18]). *For any $\boldsymbol{z} \in \mathbb{R}^n$, it holds that*

$$\begin{aligned} P(\boldsymbol{z}) &= \min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y}) = \max_{\boldsymbol{y}\in\mathcal{Y}} \min_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y}) \\ &= \max_{\boldsymbol{y}\in\mathcal{Y}} d(\boldsymbol{y},\boldsymbol{z}), \end{aligned} \tag{12}$$

$$\boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}),\boldsymbol{z}) = \boldsymbol{x}(\boldsymbol{z}) \quad \textit{for any } \boldsymbol{y}(\boldsymbol{z}) \in \mathcal{Y}(\boldsymbol{z}), \tag{13}$$

where $\mathcal{Y}(\boldsymbol{z}) := \arg\max_{\boldsymbol{y}\in\mathcal{Y}} d(\boldsymbol{y},\boldsymbol{z})$

Lemma 3 (Lemma 7 of [18]). *Suppose Assumption 1 holds. Define constants*

$$\kappa_1 := \frac{p}{\mu_{\mathcal{L}}},\ \kappa_2 := \frac{K_h}{\mu_{\mathcal{L}}}, \kappa_3 := 1 + \frac{1}{c\mu_{\mathcal{L}}} \textit{ and } \mu_{\mathcal{L}} := p - L_f.$$

Then for any $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$ and $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{X}$, it holds that

$$\|\boldsymbol{x}(\boldsymbol{y},\boldsymbol{z}) - \boldsymbol{x}(\boldsymbol{y},\boldsymbol{z}')\| \le \kappa_1 \|\boldsymbol{z} - \boldsymbol{z}'\|, \tag{14}$$

$$\|\boldsymbol{x}(\boldsymbol{z}) - \boldsymbol{x}(\boldsymbol{z}')\| \le \kappa_1 \|\boldsymbol{z} - \boldsymbol{z}'\|, \tag{15}$$

$$\|\boldsymbol{x}(\boldsymbol{y},\boldsymbol{z}) - \boldsymbol{x}(\boldsymbol{y}',\boldsymbol{z})\| \le \kappa_2 \|\boldsymbol{y} - \boldsymbol{y}'\|, \tag{16}$$

where $\boldsymbol{x}(\boldsymbol{z}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}}(\max_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y}))$.

In addition, for any $t \in \mathbb{N}$, we have

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}(\boldsymbol{y}^t,\boldsymbol{z}^t)\| \le \kappa_3 \|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|, \tag{17}$$

$$\|\boldsymbol{y}^{t+1} - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\| \le \alpha K_h \kappa_3 \|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|. \tag{18}$$

Lemma 4 (Lemma 8 of [18]). *Suppose Assumption 1 holds. The dual function $d(\boldsymbol{y},\boldsymbol{z})$ is a Lipschitz differentiable in $\boldsymbol{y}$ with modulus $K_h\kappa_2$, i.e., $\|\nabla_{\boldsymbol{y}} d(\boldsymbol{y},\boldsymbol{z}) - \nabla_{\boldsymbol{y}} d(\boldsymbol{y}',\boldsymbol{z})\| \le K_h\kappa_2\|\boldsymbol{y} - \boldsymbol{y}'\|$, $\forall \boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$.*

In the next three lemmas, we establish primal descent, dual ascent, and proximal descent properties of the proposed method.

Lemma 5 (Primal descent). *Fix $t \in \mathbb{N}$ and condition on the history $\mathcal{F}_t := \sigma\{\boldsymbol{x}^s, \boldsymbol{y}^s, \boldsymbol{z}^s, \forall s < t\}$. Let*

$$L_{\mathcal{L}} := L_f + \sqrt{m}L_h B + p, \qquad 0 < c \;\le\; \frac{1}{2L_{\mathcal{L}}}.$$

Then, under Assumption 1,

$$\begin{aligned} &\mathbb{E}\Big[\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t;\boldsymbol{y}^t) \;-\; \mathcal{L}(\boldsymbol{x}^{t+1},\boldsymbol{z}^{t+1};\boldsymbol{y}^{t+1}) \;\Big|\; \mathcal{F}_t\Big] \\ \ge&\mathbb{E}\Big[\frac{1}{4c}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 + \langle h(\boldsymbol{x}^{t+1}),\, \boldsymbol{y}^t - \boldsymbol{y}^{t+1}\rangle \\ &+ \frac{p}{2\beta}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 \Big|\, \mathcal{F}_t\Big] - \frac{cC_vL_{\mathcal{L}}^2r^2}{k}. \end{aligned} \tag{19}$$

Proof. Notice that the step of updating x uses a gradient approximation estimator. Specifically, $\boldsymbol{x}^{t+1}$ is obtained by solving:

$$\boldsymbol{x}^{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{x}^t - c\hat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t))$$

where $\hat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t)$ is calculated by 3.

So we can get

$$\langle \hat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t), \boldsymbol{x}^t - \boldsymbol{x}^{t+1}\rangle = \frac{1}{c}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 \tag{20}$$

Then

$$\begin{aligned} &\mathbb{E}\Big[\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t;\boldsymbol{y}^t) \;-\; \mathcal{L}(\boldsymbol{x}^{t+1},\boldsymbol{z}^t;\boldsymbol{y}^t) \;\Big|\; \mathcal{F}_t\Big] \\ \ge&\mathbb{E}\Big[\langle \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t,y^t), \boldsymbol{x}^t - \boldsymbol{x}^{t+1}\rangle - \frac{L_{\mathcal{L}}}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2\Big|\mathcal{F}_t\Big] \\ =&\mathbb{E}\Big[\langle \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t) - \hat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t), \boldsymbol{x}^t - \boldsymbol{x}^{t+1}\rangle \\ &+ \langle \hat{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t), \boldsymbol{x}^t - \boldsymbol{x}^{t+1}\rangle - \frac{L_{\mathcal{L}}}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2\Big|\mathcal{F}_t\Big] \\ \ge&\mathbb{E}\Big[(\frac{1}{c} - \frac{L_{\mathcal{L}}}{2})\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 - \frac{\sqrt{C_v}L_{\mathcal{L}}r}{\sqrt{k}}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|\Big|\mathcal{F}_t\Big] \\ \ge&\mathbb{E}\Big[(\frac{1}{c} - \frac{L_{\mathcal{L}}}{2})\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 - \frac{1}{4c}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 - \frac{cC_vL_{\mathcal{L}}^2r^2}{k}\Big|\mathcal{F}_t\Big] \\ \ge&\mathbb{E}\Big[\frac{1}{4c}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2\Big|\mathcal{F}_t\Big] - \frac{cC_vL_{\mathcal{L}}^2r^2}{k}, \end{aligned}$$

where the first inequality because $\mathcal{L}(\cdot,\boldsymbol{z}^t;\boldsymbol{y}^t)$ is $L_{\mathcal{L}}$-smooth, the second use 5 and Cauchy-Schwarz, the third use Young inequality.

Next, by definition of $\nabla_{\boldsymbol{y}}\mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y})$, we have

$$\begin{aligned} &\mathbb{E}\Big[\mathcal{L}(\boldsymbol{x}^{t+1},\boldsymbol{z}^t;\boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^{t+1},\boldsymbol{z}^t;\boldsymbol{y}^{t+1})\Big|\, \mathcal{F}_t\Big] \\ =&\mathbb{E}\Big[\langle h(\boldsymbol{x}^{t+1}), \boldsymbol{y}^t - \boldsymbol{y}^{t+1}\rangle\Big|\, \mathcal{F}_t\Big]. \end{aligned} \tag{21}$$

Based on the update of variable $\boldsymbol{z}^{t+1}$, i.e. $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t + \beta(\boldsymbol{x}^{t+1} - \boldsymbol{z}^t)$, it is easy to show that

$$\begin{aligned} &\mathbb{E}\Big[\mathcal{L}(\boldsymbol{x}^{t+1},\boldsymbol{z}^t;\boldsymbol{y}^{t+1}) - \mathcal{L}(\boldsymbol{x}^{t+1},\boldsymbol{z}^{t+1};\boldsymbol{y}^{t+1})\Big|\, \mathcal{F}_t\Big] \\ \ge&\mathbb{E}\Big[\frac{p}{2\beta}\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2\Big|\, \mathcal{F}_t\Big]. \end{aligned}$$

Combining the above three inequalities completes the proof. □

Lemma 6 (Dual Ascent, Lemma 10 of [18])**.** *Suppose Assumption 1 holds. For any $t \in \mathbb{N}$, we have*

$$\begin{aligned}&d(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}) - d(\boldsymbol{y}^t, \boldsymbol{z}^t)\\ \geq&\langle \nabla_{\boldsymbol{y}} d(\boldsymbol{y}^t, \boldsymbol{z}^t), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle - \frac{K_h \kappa_2}{2}\|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|^2\\ &+ \frac{p}{2}\langle \boldsymbol{z}^{t+1} - \boldsymbol{z}^t, \boldsymbol{z}^{t+1} + \boldsymbol{z}^t - 2\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\rangle.\end{aligned}$$

Lemma 7 (Proximal Descent, Lemma 11 of [18])**.** *Suppose Assumption 1 holds. For any $t \in \mathbb{N}$, we have*

$$P(\boldsymbol{z}^t) - P(\boldsymbol{z}^{t+1}) \geq \frac{p}{2}\langle \boldsymbol{z}^t - \boldsymbol{z}^{t+1}, \boldsymbol{z}^t + \boldsymbol{z}^{t+1} - 2\boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}^{t+1}), \boldsymbol{z}^t)\rangle$$

for any $\boldsymbol{y}(z^{t+1}) \in \mathcal{Y}(z^{t+1})$.

Lemma 8 (Basic Potential descent)**.** *Let Assumption 1 hold. Define constants $(\kappa_1, \kappa_2, \kappa_3, \mu_{\mathcal{L}}, L_{\mathcal{L}})$ as Lemma 3 and 5. If step sizes satisfy $0 < c \leq \min\left\{\frac{1}{8K_h}, \ \frac{\kappa_1}{32p\kappa_2^2}, \ \frac{1}{2L_{\mathcal{L}}}\right\}$ and $0 < \alpha \leq \min\left\{\frac{3}{4K_h(\kappa_3^2 - \kappa_2)}, \ \frac{1}{8cK_h^2\kappa_3^2}\right\}$, $0 < \beta \leq \frac{1}{24\kappa_1}$, then conditioning on $\mathcal{F}_t := \sigma\{\boldsymbol{x}^s, \boldsymbol{y}^s, \boldsymbol{z}^s, \forall s < t\}$, it holds:*

$$\begin{aligned}&\mathbb{E}[\Phi^t - \Phi^{t+1} \mid \mathcal{F}_t]\\ \geq&\frac{1}{16c}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 + \frac{1}{8\alpha}\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2 + \frac{p}{8\beta}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2\\ &- 24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)\|^2 - \frac{cC_v L_{\mathcal{L}}^2 r^2}{k}.\end{aligned}$$

Proof. Combining Lemmas 5-7, we have

$$\begin{aligned}&\mathbb{E}\big[\Phi^t - \Phi^{t+1} \mid \mathcal{F}_t\big] + \frac{cC_v L_{\mathcal{L}}^2 r^2}{k}\\ \geq&\mathbb{E}[\frac{1}{4c}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 + \langle h(\boldsymbol{x}^{t+1}), \boldsymbol{y}^t - \boldsymbol{y}^{t+1}\rangle\\ &+ \frac{p}{2\beta}\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 + 2\langle \nabla_{\boldsymbol{y}} d(\boldsymbol{y}^t, \boldsymbol{z}^t), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle\\ &- K_h\kappa_2\|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|^2\\ &+ p\langle \boldsymbol{z}^{t+1} - \boldsymbol{z}^t, \boldsymbol{z}^{t+1} + \boldsymbol{z}^t - 2\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\rangle\\ &- p\langle \boldsymbol{z}^{t+1} - \boldsymbol{z}^t, \boldsymbol{z}^{t+1} + \boldsymbol{z}^t - 2\boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}^{t+1}), \boldsymbol{z}^t)\rangle \mid \mathcal{F}_t]\\ =&\mathbb{E}[\frac{1}{4c}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 + \langle h(\boldsymbol{x}^{t+1}), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle\\ &+ \frac{p}{2\beta}\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 - K_h\kappa_2\|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|^2\\ &+ 2\langle \nabla_{\boldsymbol{y}} d(\boldsymbol{y}^t, \boldsymbol{z}^t) - h(\boldsymbol{x}^{t+1}), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle\\ &+ 2p\langle \boldsymbol{z}^{t+1} - \boldsymbol{z}^t, \boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}^{t+1}), \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\rangle \mid \mathcal{F}_t].\end{aligned}$$

Next we bound different terms above separately.

First, since proximal optimal and $\boldsymbol{y}^{t+1} = \Pi_{\mathcal{Y}}(\boldsymbol{y}^t + \alpha h(\boldsymbol{x}^{t+1}))$, we have

$$\langle \boldsymbol{y}^{t+1} - (\boldsymbol{y}^t + \alpha h(\boldsymbol{x}^{t+1})), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle \leq 0,$$

rearrange it gives

$$\langle h(\boldsymbol{x}^{t+1}), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle \geq \frac{1}{\alpha}\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2.$$

Next, it holds that

$$\begin{aligned}&2\langle \nabla_{\boldsymbol{y}} d(\boldsymbol{y}^t, \boldsymbol{z}^t) - h(\boldsymbol{x}^{t+1}), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle\\ \geq&-2\|h(\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t)) - h(\boldsymbol{x}^{t+1})\|\|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|\\ \geq&-2K_h\|\boldsymbol{x}^{t+1} - \boldsymbol{x}(y^t, z^t)\|\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|\\ \geq&-2K_h\left(\frac{1}{2\kappa_3^2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t)\|^2 + \frac{\kappa_3^2}{2}\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2\right)\\ \geq&-K_h\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 - K_h\kappa_3^2\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2,\end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality, the second inequality is due to h being K_h-Lipschitz, the third inequality is due to the AM-GM inequality, and the last inequality is due to (17). Also notice that

$$\begin{aligned}&2p\langle \boldsymbol{z}^{t+1} - \boldsymbol{z}^t, \boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}^{t+1}), \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\rangle\\ =&2p\langle \boldsymbol{z}^{t+1} - \boldsymbol{z}^t, \boldsymbol{x}(y(\boldsymbol{z}^{t+1}), \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}^{t+1}), \boldsymbol{z}^{t+1})\rangle\\ &+ 2p\langle \boldsymbol{z}^{t+1} - \boldsymbol{z}^t, \boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}^{t+1}), \boldsymbol{z}^{t+1}) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\rangle\\ \geq&-2p\kappa_1\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + 2p\langle \boldsymbol{z}^{t+1} - \boldsymbol{z}^t, \boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}^{t+1}), \boldsymbol{z}^{t+1})\\ &- \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\rangle\\ \geq&-2p\kappa_1\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 - \frac{p}{6\beta}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2\\ &- 6p\beta\|\boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}^{t+1}), \boldsymbol{z}^{t+1}) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\|^2,\\ =&-2p\kappa_1\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 - \frac{p}{6\beta}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2\\ &- 6p\beta\|\boldsymbol{x}(\boldsymbol{z}^{t+1}) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\|^2,\end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality and (14), the second inequality is due to the Chachy-Schwarz inequality, and the last equality is due to (13) in Lemma 2.

Combining the previous three inequalities, we have

$$\begin{aligned}&\mathbb{E}\big[\Phi^t - \Phi^{t+1} \mid \mathcal{F}_t\big] + \frac{cC_v L_{\mathcal{L}}^2 r^2}{k}\\ \geq&\mathbb{E}\Big[\left(\frac{1}{4c} - K_h\right)\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 + \left(\frac{p}{3\beta} - 2p\kappa_1\right)\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2\\ &+ \left(\frac{1}{\alpha} - K_h\kappa_3^2 - K_h\kappa_2\right)\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2\\ &- 6p\beta\|\boldsymbol{x}(\boldsymbol{z}^{t+1}) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\|^2 \mid \mathcal{F}_t\Big].\end{aligned}$$

Since

$$\alpha \leq \frac{3}{4(K_h\kappa_3^2 - K_h\kappa_2)}, \ \beta \leq \frac{1}{24\kappa_1}, \text{ and } c \leq \frac{1}{8K_h}$$

it is straightforward to check that

$$\begin{aligned}&\mathbb{E}\big[\Phi^t - \Phi^{t+1} \mid \mathcal{F}_t\big] + \frac{cC_v L_{\mathcal{L}}^2 r^2}{k}\\ \geq&\mathbb{E}\big[\frac{1}{8c}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 + \frac{1}{4\alpha}\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2 + \frac{p}{4\beta}\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2\\ &- 6p\beta\|\boldsymbol{x}(\boldsymbol{z}^{t+1}) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\|^2 \mid \mathcal{F}_t\big].\end{aligned} \tag{22}$$

By (18), we know $\|\boldsymbol{y}^{t+1} - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\| \leq \alpha K_h\kappa_3\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|$

$$\begin{aligned}\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2 =&\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t) + \boldsymbol{y}_+^t(\boldsymbol{z}^t) - \boldsymbol{y}^{t+1}\|^2\\ \geq&\frac{1}{2}\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2 - \|\boldsymbol{y}_+^t(\boldsymbol{z}^t) - \boldsymbol{y}^{t+1}\|^2\\ \geq&\frac{1}{2}\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2 - (\alpha K_h\kappa_3)^2\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2.\end{aligned} \tag{23}$$

Using the error bounds derived in Lemma 3, we have

$$\begin{aligned}
&\|\boldsymbol{x}(\boldsymbol{z}^{t+1}) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\|^2 \\
=&\|(\boldsymbol{x}(\boldsymbol{z}^{t+1}) - \boldsymbol{x}(\boldsymbol{z}^t)) + (\boldsymbol{x}(\boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)) \\
&+ (\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t)) + (\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t) \\
&- \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}))\|^2 \\
\leq&4\|\boldsymbol{x}(\boldsymbol{z}^{t+1}) - \boldsymbol{x}(\boldsymbol{z}^t)\|^2 + 4\|\boldsymbol{x}(\boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)\|^2 \\
&+ 4\|\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t)\|^2 + 4\|\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t) \\
&- \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})\|^2 \\
\leq&4\kappa_1^2\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 + 4\|\boldsymbol{x}(\boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)\|^2 \\
&+ 4\kappa_2^2\|\boldsymbol{y}_+^t(\boldsymbol{z}^t) - \boldsymbol{y}^{t+1}\| + 4\kappa_1^2\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 \\
\leq&8\kappa_1^2\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 + 4\|\boldsymbol{x}(\boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)\|^2 \\
&+ 4\kappa_2^2\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2.
\end{aligned} \tag{24}$$

Substituting (23) and (24) into (22), we have

$$\begin{aligned}
&\mathbb{E}\big[\Phi^t - \Phi^{t+1} \mid \mathcal{F}_t\big] + \frac{cC_vL_{\mathcal{L}}^2r^2}{k} \\
\geq&\mathbb{E}\big[\left(\frac{1}{8c} - \frac{\alpha K_h^2\kappa_3^2}{4} - 24p\beta\kappa_2^2\right)\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 \\
&+ \frac{1}{8\alpha}\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(z^t)\|^2 + \left(\frac{p}{4\beta} - 48p\beta\kappa_1^2\right)\|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 \\
&- 24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)\|^2 \mid \mathcal{F}_t\big].
\end{aligned} \tag{25}$$

Since we also choose

$$\alpha \leq \frac{1}{8cK_h^2\kappa_3^2},\ \beta \leq \frac{1}{24\kappa_1},\ \text{and } c \leq \frac{\kappa_1}{32p\kappa_2^2},$$

it holds that

$$\frac{\alpha K_h^2\kappa_3^2}{4} + 24p\beta\kappa_2^2 \leq \frac{1}{32c} \text{ and } 48\beta\kappa_1^2 \leq \frac{1}{8\beta}.$$

As a result, (25) can be reduced to

$$\begin{aligned}
&\mathbb{E}[\Phi^t - \Phi^{t+1} \mid \mathcal{F}_t] \\
\geq&\frac{1}{16c}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 + \frac{1}{8\alpha}\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(z^t)\|^2 + \frac{p}{8\beta}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 \\
&- 24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)\|^2 - \frac{cC_vL_{\mathcal{L}}^2r^2}{k}.
\end{aligned}$$

This completes the proof. □

Next, we state two dual error bounds.

Lemma 9 (Lemma D.1 of [20]). *Suppose Assumption 1 holds. Define*

$$\kappa_4 := \frac{\sqrt{m}B(1 + \alpha(L_f + \sqrt{m}L_hB + p)(1 + \kappa_2))}{\alpha\mu_{\mathcal{L}}}. \tag{26}$$

Then for all $\boldsymbol{z} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$, it holds that

$$\|\boldsymbol{x}(\boldsymbol{z}) - \boldsymbol{x}(\boldsymbol{y}_+(\boldsymbol{z}), \boldsymbol{z})\|^2 \leq \kappa_4\|\boldsymbol{y} - \boldsymbol{y}_+(z)\|.$$

Lemma 10 (Proposition 1 of [18]). *Suppose Assumptions 1 and 2 hold, and assume $\mathcal{X} := \{\boldsymbol{x} \in \mathbb{R}^n : g_i(\boldsymbol{x}) \leq 0, i = 1, \ldots, l\}$, g_i has L_{g_i}-Lipschitz gradient. Denote $L_g := \sqrt{\sum_{i\in[l]} L_{g_i}^2}$ and select B as:*

$$B > \max\left\{\frac{4\nabla_f D_{\mathcal{X}} + 2pD_{\mathcal{X}}^2}{\delta_0}, \frac{2\nabla_f + 2pD_{\mathcal{X}}}{\sigma_{\mathcal{X}^*}}\right\}. \tag{27}$$

Then, there exists a constant $\delta_1 > 0$ such that, if $\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\| \leq \delta_1$, the following error bound holds:

$$\|\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{z}^t)\| \leq \kappa_5\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|$$

with

$$\kappa_5 := \frac{2\left(L_f + p + (2\nabla_f + 2pD_{\mathcal{X}})\sigma_{\mathcal{X}^*}^{-1}\sqrt{L_h^2 + L_g^2}\right)\left(\frac{1}{\alpha} + K_h\kappa_2\right)}{\mu_{\mathcal{L}}\sigma_{\mathcal{X}^*}}. \tag{28}$$

We are now ready to prove Lemma 1.

Proof of Lemma 1. Recall $(\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5)$ from Lemma 8, (26), and (28), and $\delta_1 > 0$ specified in Lemma 10. Suppose $0 < c \leq \min\left\{\frac{1}{8K_h}, \frac{\kappa_1}{32p\kappa_2^2}, \frac{1}{2L_{\mathcal{L}}}\right\}$ and $0 < \alpha \leq \min\left\{\frac{3}{4K_h(\kappa_3^2 - \kappa_2)}, \frac{1}{8cK_h^2\kappa_3^2}\right\}$, $0 < \beta \leq \frac{1}{24\kappa_1}$, and β such that

$$0 < \beta \leq \min\left\{\frac{1}{24\kappa_1}, \frac{1}{384\alpha p\kappa_5^2}, \frac{\delta_1^2}{384D_{\mathcal{X}}^2p\alpha}, \frac{\delta_1^2}{6144\kappa_2^2D_{\mathcal{X}}^2p\alpha}, \frac{\delta_1^2}{12288\kappa_3^2D_{\mathcal{X}}^2pc}, \frac{\delta_1^4}{(3080\kappa_4)^2 \times 384D_{\mathcal{X}}^2p\alpha}\right\}. \tag{29}$$

and B according to (27).

Consider the following three conditions

$$\mathbb{E}[\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 \mid \mathcal{F}_t] \leq 768D_{\mathcal{X}}^2pc\beta, \tag{30}$$

$$\mathbb{E}[\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2 \mid \mathcal{F}_t] \leq 384D_{\mathcal{X}}^2p\alpha\beta, \tag{31}$$

$$\mathbb{E}[\|\boldsymbol{z}^t - \boldsymbol{x}^{t+1}\|^2 \mid \mathcal{F}_t] \leq 384\mathbb{E}[\|\boldsymbol{x}(\boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)\|^2 \mid \mathcal{F}_t]. \tag{32}$$

We consider two cases next.

1. Conditions (30)-(32) hold. Due to the choice of β in (29) and the fact that $\kappa_3 > 1$, we have

$$\mathbb{E}[\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|\|\| \mid \mathcal{F}_t] \leq \sqrt{384D_{\mathcal{X}}^2p\alpha\beta} \leq \delta_1.$$

In addition, it holds that

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{z}^t - \boldsymbol{x}(\boldsymbol{z}^t)\|^2 \mid \mathcal{F}_t] \\
=&\mathbb{E}[\|\boldsymbol{z}^t - \boldsymbol{x}^{t+1} + \boldsymbol{x}^{t+1} - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t) + \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t) \\
&- \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t) + \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{z}^t)\|^2 \mid \mathcal{F}_t] \\
\leq&\mathbb{E}[4\|\boldsymbol{z}^t - \boldsymbol{x}^{t+1}\|^2 + 4\|\boldsymbol{x}^{t+1} - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t)\|^2 \\
&+ 4\|\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t)\|^2 \\
&+ 4\|\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{z}^t)\|^2 \mid \mathcal{F}_t] \\
\leq&\mathbb{E}[1540\|\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t), \boldsymbol{z}^t) - \boldsymbol{x}(\boldsymbol{z}^t)\|^2 + 4\kappa_3^2\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|^2 \\
&+ 4\kappa_2^2\|\boldsymbol{y}^t - \boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2 \mid \mathcal{F}_t] \\
\leq&1540\kappa_4\sqrt{384D_{\mathcal{X}}^2p\alpha\beta} + 3072\kappa_3^2D_{\mathcal{X}}^2pc\beta \\
&+ 1536\kappa_2^2D_{\mathcal{X}}^2p\alpha\beta \\
\leq&\frac{\delta_1^2}{2} + \frac{\delta_1^2}{4} + \frac{\delta_1^2}{4} = \delta_1^2,
\end{aligned}$$

where the second inequality is due to (32), (16), (17), the third inequality is due to Lemma 9, (30) and (31), and the last inequality is due to the choice of β in (29). Now we have

$$\begin{aligned}&\mathbb{E}[\frac{1}{8\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\\&-24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t)-\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t),\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]\\=&\mathbb{E}[\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2+\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\\&-24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t)-\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t),\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]\\\geq&\mathbb{E}[\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2+\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\\&-24p\beta\kappa_5^2\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]\\\geq&\mathbb{E}[\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t],\end{aligned}$$

where the first inequality is due to the dual error bound in Lemma 10, and the second inequality is due to the choice of β in (29).

2. One of (30)-(32) is violated. We consider each of the three possibilities.
   (a) Condition (30) is violated. Suppose $\mathbb{E}[\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2\mid\mathcal{F}_t]>768D_{\mathcal{X}}^2pc\beta$, then

$$\begin{aligned}&\mathbb{E}[\frac{1}{16c}\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2\\&-24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t)-\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t),\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]\\\geq&\mathbb{E}[\frac{1}{32c}\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2\mid\mathcal{F}_t]+\frac{1}{32c}768D_{\mathcal{X}}^2pc\beta\\&-24D_{\mathcal{X}}^2p\beta\\=&\mathbb{E}[\frac{1}{32c}\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2\mid\mathcal{F}_t],\end{aligned}$$

   the inequality is due to $\|\boldsymbol{x}(\boldsymbol{z}^t)-\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t),\boldsymbol{z}^t)\|\leq D_{\mathcal{X}}$.

   (b) Condition (31) is violated. Suppose $\mathbb{E}[\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]>384D_{\mathcal{X}}^2p\alpha\beta$, then

$$\begin{aligned}&\mathbb{E}[\frac{1}{8\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\\&-24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t)-\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t),\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]\\\geq&\mathbb{E}[\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]+\frac{1}{16\alpha}384D_{\mathcal{X}}^2p\alpha\beta\\&-24D_{\mathcal{X}}^2p\beta\\=&\mathbb{E}[\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t].\end{aligned}$$

   (c) Condition (32) is violated. Suppose $\mathbb{E}[\|\boldsymbol{z}^t-\boldsymbol{x}^{t+1}\|^2\mid\mathcal{F}_t]>384\mathbb{E}[\|\boldsymbol{x}(\boldsymbol{z}^t)-\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t),\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]$, then

$$\begin{aligned}&\mathbb{E}[\frac{p}{8\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|^2\\&-24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t)-\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t),z^t)\|^2\mid\mathcal{F}_t]\\=&\mathbb{E}[\frac{p}{16\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|^2+\frac{p\beta}{16}\|\boldsymbol{z}^t-\boldsymbol{x}^{t+1}\|^2\\&-24p\beta\|\boldsymbol{x}(\boldsymbol{z}^t)-\boldsymbol{x}(\boldsymbol{y}_+^t(\boldsymbol{z}^t),\boldsymbol{z}^t)\|^2\mid\mathcal{F}_t]\\\geq&\mathbb{E}[\frac{p}{16\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|^2\mid\mathcal{F}_t],\end{aligned}$$

   where the equality is due to the update $\boldsymbol{z}^{k+1}=\boldsymbol{z}^k+\beta(\boldsymbol{x}^{k+1}-\boldsymbol{z}^k)$.

In both cases, the claimed inequality holds. This completes the proof.

□

C. PROOF OF THEOREM 1

To connect the algorithm's state to the ϵ-stationary conditions, we must ensure that a small dual error implies near-feasibility of the primal constraints. The following lemma from [18] provides this crucial guarantee.

Lemma 11 (Primal Feasibility, Lemma 5 of [18])**.** *Suppose Assumptions 1 and 2 hold, and the dual bound* $B>\underline{B}:=\max\Big\{\frac{4\nabla_f D_{\mathcal{X}}+4mK_hD_{\mathcal{X}}+2pD_{\mathcal{X}}^2}{\delta_0}$ *If for some* $\boldsymbol{z}\in\mathcal{X}$ *and* $\boldsymbol{y}\in[0,B]^m$*, we have* $\|\boldsymbol{y}_+(\boldsymbol{z})-\boldsymbol{y}\|\leq\eta$ *with* $0<\eta\leq\min\{\frac{1}{4}\delta_0,1\}$*, then the primal solution* $\boldsymbol{x}(\boldsymbol{y},\boldsymbol{z})$ *is nearly feasible and satisfies:*

$$\begin{aligned}\|\Pi_{\mathbb{R}_+^m}(h(\boldsymbol{x}(\boldsymbol{y},\boldsymbol{z})))\|&\leq\frac{\sqrt{m}\eta}{\alpha},\\|\langle\boldsymbol{y}_+(\boldsymbol{z}),h(\boldsymbol{x}(\boldsymbol{y},\boldsymbol{z}))\rangle|&\leq\frac{m\underline{B}\eta}{\alpha}.\end{aligned}$$

Now we proof Theorem 1.

Proof of Theorem 1. From Lemma 1,

$$\begin{aligned}&\mathbb{E}[\frac{1}{32c}\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2+\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\\&\quad+\frac{p}{16\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|^2\mid\mathcal{F}_t]\\&\leq\mathbb{E}\big[\Phi^t-\Phi^{t+1}\mid\mathcal{F}_t\big]+\frac{cC_vL_{\mathcal{L}}^2r^2}{k}.\end{aligned}$$

Taking total expectation on both sides

$$\begin{aligned}&\mathbb{E}\big[\frac{1}{32c}\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2+\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\\&\quad+\frac{p}{16\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|^2\big]\\=&\mathbb{E}\big[\mathbb{E}[\frac{1}{32c}\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2+\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\\&+\frac{p}{16\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|^2\mid\mathcal{F}_t]\big]\\\leq&\mathbb{E}\big[\mathbb{E}[\Phi^t-\Phi^{t+1}\mid\mathcal{F}_t]\big]+\frac{cC_vL_{\mathcal{L}}^2r^2}{k}\\=&\mathbb{E}\big[\Phi^t-\Phi^{t+1}\big]+\frac{cC_vL_{\mathcal{L}}^2r^2}{k}.\end{aligned}\tag{33}$$

Summing (33) from 0 to some positive index $T-1$, we have

$$\begin{aligned}&\mathbb{E}\big[\sum_{t=0}^{T-1}\big(\frac{1}{32c}\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2+\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2\\&+\frac{p}{16\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|)\big]\\\leq&\mathbb{E}\big[\sum_{t=0}^{T-1}\Phi^t-\Phi^{t+1}+\frac{cC_vL_{\mathcal{L}}^2r^2}{k}\big]\\\leq&\Phi^0-\Phi^T+\frac{TcC_vL_{\mathcal{L}}^2r^2}{k}.\end{aligned}$$

By definitions of $d(\cdot)$ and $P(\cdot)$, we immediately have

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{y})\geq d(\boldsymbol{y},\boldsymbol{z}),\ P(\boldsymbol{z})\geq d(\boldsymbol{y},\boldsymbol{z}),\text{ and }P(\boldsymbol{z})\geq\underline{f}.$$

Thus,

$$\Phi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \geq P(\boldsymbol{z}) \geq \underline{f}.$$

So

$$\begin{aligned}&\mathbb{E}\Big[\sum_{t=0}^{T-1}\Big(\frac{1}{32c}\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|^2+\frac{1}{16\alpha}\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(z^t)\|^2\\&+\frac{p}{16\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|\Big)\Big]\\&\leq\Phi^0-\underline{f}+\frac{TcC_vL_{\mathcal{L}}^2r^2}{k},\end{aligned}$$

By summing the descent bound for $t=0,\ldots,T-1$, we obtain an index $t^*\in\{0,1,\ldots,T-1\}$. For notational convenience we drop the asterisk and denote this index simply by t, so that

$$\begin{aligned}&\mathbb{E}\big[\|\boldsymbol{x}^t-\boldsymbol{x}^t\|^2+\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|^2+\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|^2\big]\\&\leq\frac{\Phi^0-\underline{f}}{T\lambda_0}+\frac{cC_vL_{\mathcal{L}}^2r^2}{k\lambda_0},\end{aligned}$$

$\lambda_0=\min\{1/(32c),1/(16\alpha),p/(16\beta)\}$, which further implies that

$$\begin{aligned}\varrho_{t+1}:=&\max\{\mathbb{E}\big[\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|\big],\mathbb{E}\big[\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|\big],\mathbb{E}\big[\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|\big]\}\\&\leq\sqrt{\frac{\Phi^0-\underline{f}}{T\lambda_0}+\frac{cC_vL_{\mathcal{L}}^2r^2}{k\lambda_0}}.\end{aligned}\tag{34}$$

Next we verify that the iterate $(\boldsymbol{x}^{t+1},\boldsymbol{y}^t)$ satisfies the ϵ-stationarity conditions in Definition 1. From the primal update rule we have

$$\boldsymbol{x}^{t+1}=\Pi_{\mathcal{X}}\big(\boldsymbol{x}^t-c(\hat{\nabla}f(\boldsymbol{x}^t)+\hat{\nabla}h(\boldsymbol{x}^t)^\top\boldsymbol{y}^t+p(\boldsymbol{x}^t-\boldsymbol{z}^t))\big).$$

Using the optimality condition of the projection step, we know

$$\frac{1}{c}(\boldsymbol{x}^t-\boldsymbol{x}^{t+1})-\big(\hat{\nabla}f(\boldsymbol{x}^t)+\hat{\nabla}h(\boldsymbol{x}^t)^\top\boldsymbol{y}^t+p(\boldsymbol{x}^t-\boldsymbol{z}^t)\big)\in N_{\mathcal{X}}(\boldsymbol{x}^{t+1}).$$

Therefore,

$$\begin{aligned}&\operatorname{dist}\Big(\boldsymbol{0},\nabla f(\boldsymbol{x}^{t+1})+\nabla h(\boldsymbol{x}^{t+1})^\top\boldsymbol{y}^t+N_{\mathcal{X}}(\boldsymbol{x}^{t+1})\Big)\\&\leq\big\|\nabla f(\boldsymbol{x}^{t+1})-\hat{\nabla}f(\boldsymbol{x}^t)+(\nabla h(\boldsymbol{x}^{t+1})-\hat{\nabla}h(\boldsymbol{x}^t))^\top\boldsymbol{y}^t\\&\quad-p(\boldsymbol{x}^t-\boldsymbol{z}^t)-\tfrac{1}{c}(\boldsymbol{x}^{t+1}-\boldsymbol{x}^t)\big\|.\end{aligned}\tag{35}$$

By the L_f-smoothness of f, we have

$$\|\nabla f(\boldsymbol{x}^{t+1})-\nabla f(\boldsymbol{x}^t)\|\ \leq\ L_f\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|,$$

and by the bias bound of the zeroth-order estimator (4),

$$\mathbb{E}\big[\|\nabla f(\boldsymbol{x}^t)-\hat{\nabla}f(\boldsymbol{x}^t)\|\big]\ \leq\ L_fr.\tag{36}$$

Combining the above yields

$$\mathbb{E}\big[\|\nabla f(\boldsymbol{x}^{t+1})-\hat{\nabla}f(\boldsymbol{x}^t)\|\big]\ \leq\ L_f\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|+L_fr.\tag{37}$$

Similarly, by the L_h-smoothness of h,

$$\|\nabla h(\boldsymbol{x}^{t+1})-\nabla h(\boldsymbol{x}^t)\|\ \leq\ L_h\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|.$$

Using the boundedness of the multipliers $\|\boldsymbol{y}^t\|\leq B\sqrt{m}$,

$$\|(\nabla h(\boldsymbol{x}^{t+1})-\nabla h(\boldsymbol{x}^t))^\top\boldsymbol{y}^t\|\ \leq\ \sqrt{m}BL_h\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|.$$

For the estimation error we also have

$$\mathbb{E}\big[\|\nabla h(\boldsymbol{x}^t)-\hat{\nabla}h(\boldsymbol{x}^t)\|\cdot\|\boldsymbol{y}^t\|\big]\ \leq\ \sqrt{m}BL_h\,r.$$

From the update of $\boldsymbol{z}^t$ and the bound $\mathbb{E}\|\boldsymbol{x}^t-\boldsymbol{z}^t\|\leq\varrho_{t+1}+\frac{1}{\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|$, we can further control $p\|\boldsymbol{x}^t-\boldsymbol{z}^t\|$ by $\frac{p}{\beta}\varrho_{t+1}$.

The final term $\frac{1}{c}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|$ is directly bounded by $\frac{1}{c}\varrho_{t+1}$.

Substituting (37) and the above bounds into (35), and taking expectation, we obtain

$$\begin{aligned}&\mathbb{E}\Big[\operatorname{dist}\big(\boldsymbol{0},\nabla f(\boldsymbol{x}^{t+1})+\nabla h(\boldsymbol{x}^{t+1})^\top\boldsymbol{y}^t+N_{\mathcal{X}}(\boldsymbol{x}^{t+1})\big)\Big]\\&\leq\lambda_1\varrho_{t+1}+L_{\mathcal{L}}r,\end{aligned}\tag{38}$$

where $\lambda_1:=L_f+\sqrt{m}BL_h+\frac{1}{c}+p+\frac{p}{\beta}$.

Also notice that by (17) and the update of $\boldsymbol{z}^{k+1}$, we haves

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{x}(\boldsymbol{y}^t,\boldsymbol{z}^t)-\boldsymbol{z}^t\|\big]&\leq\mathbb{E}\big[\|\boldsymbol{x}(\boldsymbol{y}^t,\boldsymbol{z}^t)-\boldsymbol{x}^{t+1}\|+\|\boldsymbol{x}^{t+1}-\boldsymbol{z}^t\|\big]\\&\leq\mathbb{E}\big[\kappa_3\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|+\frac{1}{\beta}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|\big],\end{aligned}$$

and therefore we have

$$\begin{aligned}&\max\{\mathbb{E}\big[\|\boldsymbol{x}(\boldsymbol{y}^t,\boldsymbol{z}^t)-\boldsymbol{z}^t\|\big],\mathbb{E}\big[\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|\big]\}\\&\leq\eta_t:=(\kappa_3+1/\beta)\varrho_{t+1}.\end{aligned}$$

To satisfy the condition in Lemma 11, we must have $\sqrt{\frac{\Phi^0-\underline{f}}{T\lambda_0}+\frac{cC_vL_{\mathcal{L}}^2r^2}{k\lambda_0}}\leq\eta$, so we choose $T\geq\frac{k(\Phi^0-\underline{f})}{\eta k\lambda_0-cC_vL_{\mathcal{L}}^2r^2}$. Next we consider primal infeasibility at $\boldsymbol{x}^{t+1}$:

$$\begin{aligned}&\mathbb{E}\big[\|\Pi_{\mathcal{X}}(h(\boldsymbol{x}^{t+1}))\|\big]\\&\leq\mathbb{E}\big[\|\Pi_{\mathcal{X}}(h(\boldsymbol{x}(\boldsymbol{y}^t,\boldsymbol{z}^t))\|+\|\Pi_{\mathcal{X}}(h(\boldsymbol{x}(\boldsymbol{y}^t,\boldsymbol{z}^t))-\Pi_{\mathcal{X}}(h(\boldsymbol{x}^{t+1}))\|\big]\\&\leq\frac{\sqrt{m}}{\alpha}\eta_t+\mathbb{E}\big[\|h(\boldsymbol{x}(\boldsymbol{y}^t,\boldsymbol{z}^t)-h(\boldsymbol{x}^{t+1})\|\big]\\&\leq\frac{\sqrt{m}}{\alpha}\eta_t+K_h\kappa_3\mathbb{E}\big[\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|\big]\\&\leq\left(\frac{\sqrt{m}(\kappa_3+1/\beta)}{\alpha}+K_h\kappa_3\right)\varrho_{t+1}\\&:=\lambda_2\varrho_{t+1},\end{aligned}\tag{39}$$

where the second inequality is due to the non-expansiveness of the projection operator, the third inequality is due to h being K_h-Lipschitz and (17), and the last inequality is due to Lemma 11. Moreover, the violation of complementary slackness can be bounded by

$$\begin{aligned}&\max\{\mathbb{E}\big[|\langle\boldsymbol{y}^t,h(\boldsymbol{x}^{t+1})\rangle|\big]\\&\leq\max\{\mathbb{E}\big[|\langle\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t),h(\boldsymbol{x}^{t+1})\rangle|+|\langle\boldsymbol{y}_+^t(\boldsymbol{z}^t),h(\boldsymbol{x}^{t+1})\\&\quad-h(\boldsymbol{x}(\boldsymbol{z}^t,\boldsymbol{y}^t))\rangle|+|\langle\boldsymbol{y}_+^t(\boldsymbol{z}^t),h(\boldsymbol{x}(\boldsymbol{z}^t,\boldsymbol{y}^t))\rangle|\big]\\&\leq\max\{\mathbb{E}\big[M_h\|\boldsymbol{y}^t-\boldsymbol{y}_+^t(\boldsymbol{z}^t)\|+\sqrt{m}\underline{B}K_h\kappa_3\|\boldsymbol{x}^t-\boldsymbol{x}^{t+1}\|\big]\\&\quad+\frac{m\underline{B}\eta_t}{\alpha}\\&\leq\left(M_h(\kappa_3+1/\beta)+\sqrt{m}\underline{B}K_h\kappa_3+\frac{m\underline{B}(\kappa_3+1/\beta)}{\alpha}\right)\varrho_{t+1}\\&:=\lambda_3\varrho_{t+1},\end{aligned}\tag{40}$$

where the second inequality is due to h being bounded and K_h-Lipschitz over X, (17), and Lemma 11. As a result of (38)-(40), we see that

$$\begin{aligned}
&\max\Big\{\mathbb{E}\Big[\mathrm{dist}\big(\mathbf{0}, \nabla f(\boldsymbol{x}^{t+1}) + \nabla h(\boldsymbol{x}^{t+1})^\top \boldsymbol{y}^t + N_{\mathcal{X}}(\boldsymbol{x}^{t+1})\big)\Big], \\
&\qquad\qquad \mathbb{E}\big[\|\Pi_{\mathbb{R}^m_+}(h(\boldsymbol{x}^{t+1}))\|\big], \mathbb{E}\big[|\langle \boldsymbol{y}^t, h(\boldsymbol{x}^{t+1})\rangle|\big]\Big\} \\
&\quad \le \max\{\lambda_1, \lambda_2, \lambda_3\}\, \varrho_{t+1} + L_{\mathcal{L}} r \\
&\quad \le \max\{\lambda_1, \lambda_2, \lambda_3\} \sqrt{\frac{\Phi^0 - \underline{f}}{T\lambda_0} + \frac{cC_v L_{\mathcal{L}}^2 r^2}{k\lambda_0}} + L_{\mathcal{L}} r,
\end{aligned}$$

where the second inequality is due to (34).

With the smoothing radius fixed to $r = \frac{\epsilon}{2L_{\mathcal{L}}}$, the bound

$$\max\{\lambda_1, \lambda_2, \lambda_3\} \sqrt{\frac{\Phi^0 - \underline{f}}{T\lambda_0} + \frac{c\,C_v L_{\mathcal{L}}^2\, r^2}{k\lambda_0}} \;\le\; \frac{\epsilon}{2}$$

holds provided

$$4\lambda_0\, k\, L_{\mathcal{L}}^2 \;-\; c\, n^2 L_{\mathcal{L}}^2 \big(\max\{\lambda_1, \lambda_2, \lambda_3\}\big)^2 \;>\; 0, \tag{41}$$

or, equivalently,

$$k \;>\; k_{\min} := \frac{c\, n^2}{4\lambda_0}\big(\max\{\lambda_1, \lambda_2, \lambda_3\}\big)^2.$$

Under (41) we obtain

$$T \ge \frac{4\, k\, (\Phi^0 - \underline{f})\, \big(\max\{\lambda_1, \lambda_2, \lambda_3\}\big)^2}{\epsilon^2 \left(\lambda_0 k - c\, C_v\, \big(\max\{\lambda_1, \lambda_2, \lambda_3\}\big)^2\right)}. \tag{42}$$

Writing (42) in the form of Theorem 1,

$$T \ge \frac{C_1(\Phi^0 - \underline{f})}{\epsilon^2 \big(C_2 - \frac{C_3}{k}\big)},$$

gives the explicit constants

$$\begin{aligned}
C_1 &:= 4\big(\max\{\lambda_1, \lambda_2, \lambda_3\}\big)^2, \\
C_2 &:= \lambda_0 \\
C_3 &:= c\, C_v\, \big(\max\{\lambda_1, \lambda_2, \lambda_3\}\big)^2.
\end{aligned}$$

□