Trust-Region Sequential Quadratic Programming for Stochastic Optimization with Random Models: First-Order Stationarity

Yuchen Fang YCFANG@UCHICAGO.EDU
Committee on Computational and Applied Mathematics, The University of Chicago

Sen Na SENNA@BERKELEY.EDU
ICSI and Department of Statistics, University of California, Berkeley

Mladen Kolar MLADEN.KOLAR@CHICAGOBOOTH.EDU
Booth School of Business, The University of Chicago

Abstract

We consider optimization problems with a stochastic objective and deterministic constraints, and design a trust-region sequential quadratic programming (TR-SQP) method to solve them. We name our method TR-SQP for STochastic Optimization with Random Models (TR-SQP-STORM). In each iteration, our algorithm constructs a random model for the objective that satisfies suitable accuracy conditions with a high but fixed probability. The algorithm decides whether a trial step is successful or not based on two ratios: the ratio between the estimated actual reduction and predicted reduction on the ℓ_2 merit function, and the ratio between the estimated KKT residual and trust-region radius. For each successful step, the algorithm increases the trust-region radius, and further decides whether the step is reliable or not based on the amount of the predicted reduction. If the step is reliable, then the algorithm relaxes the accuracy conditions for the next iteration. To resolve the infeasibility issue of trust-region methods for constrained problems, we employ an *adaptive relaxation technique* proposed by a companion paper. Under reasonable assumptions, we establish the global first-order convergence guarantee: the KKT residual converges to zero almost surely. We apply our method on a subset of problems in CUTEst set to demonstrate its empirical performance.

1. Introduction

We consider solving the following constrained stochastic optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \; f(\boldsymbol{x}) = \mathbb{E}[F(\boldsymbol{x};\xi)], \qquad \text{s.t.} \;\; c(\boldsymbol{x}) = \boldsymbol{0}, \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is a stochastic objective, $F(\cdot;\xi):\mathbb{R}^d\to\mathbb{R}$ is a realization, $c:\mathbb{R}^d\to\mathbb{R}^m$ are deterministic equality constraints, and $\xi\sim\mathcal{P}$ is a random variable following the distribution $\mathcal{P}$. Problems of this form are popular in various scientific and engineering fields, including optimal control [5], constrained deep neural networks [7], and reinforcement learning [1].

Numerous methods have been proposed for solving constrained optimization problems, such as penalty methods, augmented Lagrangian methods, and sequential quadratic programming (SQP) methods. All of these methods have advantages that make them suitable for different settings. Our paper focuses on solving (1) with a stochastic SQP (StoSQP) method. [3] proposed the first StoSQP algorithm, in which a single sample is obtained in each step to estimate the objective model and a projection procedure is used to select the stepsize. Based on the ℓ_1 merit function, that projection procedure uses a deterministic user-specified sequence $\{\beta_k\}$ to construct a projection interval in each step and determine the stepsize. The projection procedure adds adaptivity to the method compared to algorithms that require users to directly specify stepsizes. However, since the projection interval boundaries scale as β_k with the length scaling as β_k^2, the prespecified sequence $\{\beta_k\}$ still highly affects the magnitude of the selected stepsize, and thus affects the performance of the algorithm. To address this issue, [12] considered a *random model setup* and incorporated a stochastic line search procedure with StoSQP. In each step, a model of the objective is constructed that satisfies suitable accuracy conditions with a high but fixed probability. This model is used to compute an augmented Lagrangian merit function, and the stepsize is selected by checking the Armijo condition. Several papers built upon [3] and [12], and have further improved the design from different points of view. [2] designed a scheme to deal with rank-deficient Jacobians; [9] allowed approximately solving the SQP subproblems; [4] applied the SVRG technique inside StoSQP for finite-sum problems; and [11] generalized the study in [12] by developing an active-set method to enable inequality constraints.

We contribute to the above literature by proposing the first trust-region StoSQP method for stochastic optimization with random models, which we term TR-SQP-STORM.[1] Under the random model setup [8, 11, 12], we assume in each iteration that the objective model is accurately estimated with a high but fixed probability, which can be achieved by using a batch of samples with adaptively selected batch sizes. In each iteration of TR-SQP-STORM, we decide the trial step as successful or not based on (1) the ratio between the estimated actual reduction and predicted reduction on the ℓ_2 merit function, and (2) the ratio between the estimated KKT residual and the trust-region radius. The former ratio is used for deterministic trust-region methods [13, Chapter 4], while the latter ratio is inspired by [8] and is particularly used for stochastic optimization. If the trial step is successful, then the trust-region radius is increased, otherwise, it is decreased. Unlike [8], for each successful trial step, the algorithm further decides whether the step is reliable or not based on the amount of the predicted reduction. When the predicted reduction is large, we relax some model accuracy conditions (e.g., the variance of the objective function estimate) for the next iteration. In addition, when designing trust-region methods for constrained optimization, we have to resolve an infeasibility issue—the intersection of linearized constraints and trust-region constraint may lead to an empty feasible set for SQP subproblems. In this regard, we employ an adaptive relaxation technique proposed in a companion paper. Specifically, we relax the linearized constraints by controlling the sizes of the normal step and the tangential step separately, and the control radii are computed by splitting the full trust-region radius into two parts. With the above extensions of the STORM method for unconstrained optimization [8], our TR-SQP-STORM method enjoys three advantages over existing line-search-based StoSQP methods [11, 12].

First, our method computes the search direction and selects the stepsize jointly, which can lead to a better performance in some cases [13, Chapter 4]. This property is in contrast to the line-search-based procedure, where a search direction is computed first and then the setpsize is decided based on either line search procedure or projection techniques. Second, due to the trust-region constraint, the SQP subproblems are well defined even with an indefinite Lagrangian Hessian approximate. Thus, we can employ the unperturbed Hessian estimate at the current iterate to form the SQP subproblem. On the contrary, existing StoSQP schemes all require a positive definite Hessian approximate in the null space of linearized constraints, so a Hessian perturbation procedure is generally needed. Third, in addition to ensuring that the search direction is a descent direction, [11, 12] imposed an extra fea-

[1] The name STORM is borrowed from [8], where the authors designed a trust-region method for unconstrained stochastic optimization with random models (STORM).

sibility error condition when selecting the merit parameters. This condition is completely eliminated in our study.

With the above differences to existing StoSQP methods, we establish the first-order convergence guarantee of TR-SQP-STORM by showing that the KKT residual converges to zero almost surely. We note that a recent work [6] designed a second-order STORM method for unconstrained problems that converges to a strict local minimum. Motivated by that, we will also report on our second-order stationarity results for constrained problems shortly. We apply the method on a subset of problems in CUTEst test set to demonstrate its empirical performance.

Notation. We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the operator norm for matrices. We use I to denote the identity matrix and $\mathbf{0}$ to denote the zero matrix, whose dimensions can be inferred from the context. We let $G(\boldsymbol{x}) = \nabla^T c(\boldsymbol{x}) \in \mathbb{R}^{m\times d}$ be the Jacobian matrix of the constraints and let $P(\boldsymbol{x}) = I - G^T(\boldsymbol{x})[G(\boldsymbol{x})G^T(\boldsymbol{x})]^{-1}G(\boldsymbol{x})$ be a projection matrix. We let $\bar{g}(\boldsymbol{x}) = \nabla F(\boldsymbol{x};\xi)$ be a realized objective gradient, and by $\bar{(\cdot)}$ we denote all stochastic estimates. The Lagrangian of (1) is defined as $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T c(\boldsymbol{x})$. At the k-th iteration, we let $\bar{g}_k = \bar{g}(\boldsymbol{x}_k)$ and $G_k = G(\boldsymbol{x}_k)$ (similar notation is used for $c_k, P_k, \nabla\mathcal{L}_k, \bar{\nabla}\mathcal{L}_k$, etc.).

2. Adaptive Relaxation Technique

We describe an adaptive relaxation technique proposed in a companion paper. This technique is used to address the infeasibility issue of trust-region methods for constrained optimization: the intersection of $\{\Delta\boldsymbol{x}_k \in \mathbb{R}^d : c_k + G_k\Delta\boldsymbol{x}_k = \mathbf{0}\}$ and $\{\Delta\boldsymbol{x}_k \in \mathbb{R}^d : \|\Delta\boldsymbol{x}_k\| \le \Delta_k\}$ can be empty, where Δ_k is the given trust-region radius. To resolve this problem, we decompose the trial step $\Delta\boldsymbol{x}_k$ into a normal step and a tangential step as $\Delta\boldsymbol{x}_k = \boldsymbol{w}_k + \boldsymbol{t}_k$, where $\boldsymbol{w}_k \in \mathrm{im}(G_k^T)$ is the normal step and $\boldsymbol{t}_k \in \ker(G_k)$ is the tangential step. We further write $\boldsymbol{t}_k = P_k\boldsymbol{u}_k$ for some vector $\boldsymbol{u}_k \in \mathbb{R}^d$. Without the trust-region constraint $\|\Delta\boldsymbol{x}_k\| \le \Delta_k$, we know that $\boldsymbol{w}_k = -G_k^T[G_kG_k^T]^{-1}c_k =: \boldsymbol{v}_k$ (suppose G_k has full row rank). However, the trust-region constraint may prevent us from taking the full length of $\boldsymbol{v}_k$. Thus, we let $\boldsymbol{w}_k = \bar{\gamma}_k\boldsymbol{v}_k$ for a scalar $\bar{\gamma}_k \in (0,1]$. Finally, we have $\Delta\boldsymbol{x}_k = \bar{\gamma}_k\boldsymbol{v}_k + P_k\boldsymbol{u}_k$. To select proper $\bar{\gamma}_k$ and $\boldsymbol{u}_k$, we decompose the trust-region radius into two segments

$$\breve{\Delta}_k = \frac{\|c_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}\Delta_k \qquad \text{and} \qquad \tilde{\Delta}_k = \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}\Delta_k. \tag{2}$$

Then, we let $\bar{\gamma}_k = \min\{\breve{\Delta}_k/\|\boldsymbol{v}_k\|, 1\}$, and let $\boldsymbol{u}_k$ be solved by a standard trust-region problem

$$\min_{\boldsymbol{u}\in\mathbb{R}^d} \quad \bar{g}_k^T P_k\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^T P_k B_k P_k\boldsymbol{u} \qquad \text{s.t.} \;\; \|\boldsymbol{u}\| \le \tilde{\Delta}_k. \tag{3}$$

Any approximate solution of the problem (3) that achieves a fraction $\kappa_{fcd} \in (0,1]$ of the objective reduction achieved by the Cauchy point is sufficient for our analysis.

3. TR-SQP-STORM

Let $\eta, p_g, p_f \in (0,1)$, $\Delta_{\max}, \kappa_g > 0, \kappa_f = \kappa_{fcd}\eta^3/16$, $\gamma > 1$ be input parameters. Given the current triple $(\boldsymbol{x}_k, \Delta_k, \bar{\epsilon}_k)$, where $\boldsymbol{x}_k$ is the (primal) iterate, Δ_k is the trust-region radius, and $\bar{\epsilon}_k$ is the reliability parameter, we perform the following four steps.

Step 1. We use a batch of samples to obtain the gradient estimate $\bar{g}_k$ that satisfies

$$\mathbb{P}(\mathcal{A}_k \mid \boldsymbol{x}_k) \geq p_g, \qquad \text{with } \; \mathcal{A}_k = \{\|\nabla f_k - \bar{g}_k\| \leq \kappa_g \Delta_k\}\,.$$

Step 2. We compute the dual iterate $\bar{\boldsymbol{\lambda}}_k = -[G_k G_k^T]^{-1} G_k \bar{g}_k$, and the gradients $\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_k$ and $\bar{\nabla} \mathcal{L}_k$. For a matrix B_k that approximates the Lagrangian Hessian $\nabla^2_{\boldsymbol{x}} \mathcal{L}_k$, we compute the step $\Delta \boldsymbol{x}_k$ as in Section 2. Based on $\Delta \boldsymbol{x}_k$, we select the merit parameter $\bar{\mu}_k$ that is large enough so that

$$\overline{\text{Pred}}_k \leq -\frac{\kappa_{fcd}}{2} \|\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_k\| \min\left\{\tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{2} \|c_k\| \min\left\{\breve{\Delta}_k, \frac{\|c_k\|}{\|G_k\|}\right\}, \tag{4}$$

where κ_{fcd} is defined in Section 2 and the estimated predicted reduction is defined as

$$\text{Pred}_k = \bar{g}_k^T \Delta \boldsymbol{x}_k + \frac{1}{2} \Delta \boldsymbol{x}_k^T B_k \Delta \boldsymbol{x}_k + \bar{\mu}_k (\|c_k + G_k \Delta \boldsymbol{x}_k\| - \|c_k\|).$$

Step 3. Let $\boldsymbol{x}_{s_k} := \boldsymbol{x}_k + \Delta \boldsymbol{x}_k$. We estimate $\bar{f}_k$, $\bar{f}_{s_k}$, the objective values at $\boldsymbol{x}_k$, $\boldsymbol{x}_{s_k}$, that satisfy

$$\mathbb{P}(\mathcal{B}_k \mid \boldsymbol{x}_k, \Delta \boldsymbol{x}_k) \geq p_f, \quad \max\left\{\mathbb{E}\left[|f_k - \bar{f}_k| \mid \boldsymbol{x}_k, \Delta \boldsymbol{x}_k\right], \mathbb{E}\left[|f_{s_k} - \bar{f}_{s_k}| \mid \boldsymbol{x}_k, \Delta \boldsymbol{x}_k\right]\right\} \leq \bar{\epsilon}_k^2,$$

where $\mathcal{B}_k = \left\{\max(|f_k - \bar{f}_k|, |f_{s_k} - \bar{f}_{s_k}|) \leq \kappa_f \Delta_k^2\right\}$. The estimated actual reduction $\text{Ared}_k = \bar{\mathcal{L}}_{\bar{\mu}_k}^{s_k} - \bar{\mathcal{L}}_{\bar{\mu}_k}^{k}$ is computed based on the ℓ_2 merit function:

$$\bar{\mathcal{L}}_{\bar{\mu}_k}^{s_k} := \bar{\mathcal{L}}_{\bar{\mu}_k}(\boldsymbol{x}_{s_k}) = \bar{f}_{s_k} + \bar{\mu}_k \|c_{s_k}\| \quad \text{ and } \quad \bar{\mathcal{L}}_{\bar{\mu}_k}^{k} := \bar{\mathcal{L}}_{\bar{\mu}_k}(\boldsymbol{x}_k) = \bar{f}_k + \bar{\mu}_k \|c_k\|.$$

Step 4. We update the triple $(\boldsymbol{x}_k, \Delta_k, \bar{\epsilon}_k)$. Let $\bar{\rho}_1 = \text{Ared}_k / \text{Pred}_k$ and $\bar{\rho}_2 = \|\bar{\nabla} \mathcal{L}_k\| / \Delta_k$. If

$$\bar{\rho}_1 \geq \eta \qquad \text{and} \qquad \bar{\rho}_2 \geq \eta \cdot \max\{\|B_k\|, \|G_k\|, 1\}, \tag{5}$$

then the trial step $\Delta \boldsymbol{x}_k$ is successful, and we update the iterate and radius as $\boldsymbol{x}_{k+1} = \boldsymbol{x}_{s_k}$ and $\Delta_{k+1} = \min\{\gamma \Delta_k, \Delta_{\max}\}$. Furthermore, if $-\text{Pred}_k \geq \bar{\epsilon}_k$ also holds, then $\Delta \boldsymbol{x}_k$ is a reliable step, and we set $\bar{\epsilon}_{k+1} = \gamma \bar{\epsilon}_k$ to relax the variance of the objective estimate for the next iteration; otherwise, we let $\bar{\epsilon}_{k+1} = \bar{\epsilon}_k / \gamma$. If either condition of (5) does not hold, then $\Delta \boldsymbol{x}_k$ is an unsuccessful step. We let $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$, $\Delta_{k+1} = \Delta_k / \gamma$, and $\bar{\epsilon}_{k+1} = \bar{\epsilon}_k / \gamma$.

The following assumptions are used to establish the global convergence guarantee.

Assumption 1 *The iterates $\boldsymbol{x}_k, \boldsymbol{x}_{s_k}$ lie in some open convex set Ω. The function $f(\boldsymbol{x})$ is continuously differentiable and bounded below by $f_{\inf}$. The gradient $\nabla f(\boldsymbol{x})$ is Lipschitz continuous on Ω with constant $L_{\nabla f}$. The constraints $c(\boldsymbol{x})$ are continuously differentiable. The Jacobian matrix $G(\boldsymbol{x})$ is Lipschitz continuous on Ω with constant L_G. We also assume that there exist constants $\kappa_B \geq 1, \kappa_c, \kappa_{\nabla f}, \kappa_{1,G}, \kappa_{2,G} > 0$ such that*

$$\|B_k\| \leq \kappa_B, \|c_k\| \leq \kappa_c, \|\nabla f_k\| \leq \kappa_{\nabla f}, \kappa_{1,G} \cdot I \preceq G_k G_k^T \preceq \kappa_{2,G} \cdot I \textit{ for all } k \in \mathbb{N}.$$

Assumption 2 (Stabilization of the merit parameter) *There exist a stochastic $\bar{K} < \infty$ and a deterministic constant $\hat{\mu}$, such that for $\forall k > \bar{K}$, $\bar{\mu}_k = \bar{\mu}_{\bar{K}} \leq \hat{\mu}$.*

Assumption 1 is standard in the existing literature [3]. Assumption 2 holds if $\bar{g}_k$ satisfies a boundedness condition; detailed analysis can be found in Appendix B. Now we establish the global convergence guarantee of TR-SQP-STORM. The proof is deferred to Appendix A.2. We use the following potential function

$$\Phi^k_{\bar{\mu}_{\bar{K}}} = \nu \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} + \frac{1-\nu}{2}\Delta_k^2 + \frac{1-\nu}{2}\bar{\epsilon}_k,$$

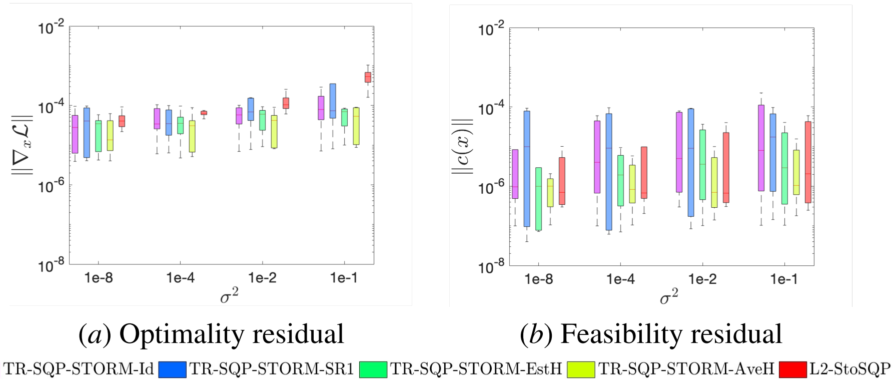
where $\mathcal{L}^k_{\bar{\mu}_{\bar{K}}} = f_k + \bar{\mu}_{\bar{K}}\|c_k\|$, and $\nu \in (0,1)$ is a deterministic parameter satisfying $\frac{\nu}{1-\nu} \geq \max\left\{\frac{\gamma^2}{\Upsilon}, \frac{2\gamma}{\eta}\right\}$ for constants Υ derived in the analysis.

Theorem 3 *Suppose that Assumptions 1, 2 hold, ν satisfies the above condition, and p_f, p_{grad} are large enough so that $p_g p_f \geq \frac{1}{2}$ and $p_f \geq 1 - \frac{(1-\nu)(\gamma-1)}{4\nu\gamma}$. Then $\lim_{k\to\infty}\|\nabla\mathcal{L}_k\| = 0$ a.s.*

Theorem 3 establishes the global convergence result for TR-SQP-STORM, which matches the conclusion in [8] for trust-region methods for unconstrained problems.

4. Numerical Experiments

We demonstrate the empirical performance of TR-SQP-STORM on a subset of problems in CUTEst set [10]. We compare our method with the ℓ_2-StoSQP [12, Algorithm 3]. Given a noise level σ^2 within $\{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}\}$, we draw estimates of f_k from the Gaussian distribution $\mathcal{N}(f_k, \sigma^2)$ and estimates of ∇f_k from $\mathcal{N}(\nabla f_k, \sigma^2(I + \mathbf{1}\mathbf{1}^T))$, where $\mathbf{1} \in \mathbb{R}^d$ denotes the all one vector. Both algorithms adaptively select sample sizes to satisfy respective conditions. For ℓ_2-StoSQP, we let $B_k = I$ (as used in [12]). For our method, we try four Hessian approximations: the identity matrix (Id), the symmetric rank-one (SR1) update, the estimated Hessian (EstH), and the averaged Hessian (AveH). When employing EstH and AveH, we estimate $\nabla^2 f_k$ with only one sample per iteration. The (i,j) (together with (j,i)) entry of the estimate of $\nabla^2 f_k$ is drawn from $\mathcal{N}((\nabla^2 f_k)_{i,j}, \sigma^2)$, where σ^2 is the same as for estimating the objective model. We draw boxplots for the optimality residual $\|\nabla_{\boldsymbol{x}}\mathcal{L}\|$ and the feasibility residual $\|c(\boldsymbol{x})\|$ of both methods in Figure 1.



(a) Optimality residual

(b) Feasibility residual

Figure 1: KKT residual boxplots of CUTEst problem set.

From Figure 1, we observe that TR-SQP-STORM generally outperforms ℓ_2-StoSQP for the optimality residual (especially when the noise level is high), while the two methods are comparable for the feasibility residual. Among the four choices of Hessian approximates, the averaged Hessian consistently performs the best, and the current estimated Hessian is in the second place. This observation is consistent with our intuition.

5. Conclusion

We developed TR-SQP-STORM to solve stochastic optimization problems with equality constraints. The algorithm is developed in a random model setup where the objective model is estimated sufficiently accurate with a high but fixed probability. We adopted the adaptive relaxation technique to resolve the infeasibility issue. The global almost sure convergence property of TR-SQP-STORM is established and its superior performance is demonstrated on a subset of CUTEst problems.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *International conference on machine learning*, pages 22–31, 2017.

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

[3] Albert S. Berahas, Frank E. Curtis, Daniel Robinson, and Baoyu Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, jan 2021.

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

[5] John T. Betts. *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*. Society for Industrial and Applied Mathematics, jan 2010.

[6] Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS Journal on Optimization*, 1(2):92–119, apr 2019.

[7] Changan Chen, Frederick Tung, Naveen Vedula, and Greg Mori. Constraint-aware deep neural network compression. *European Conference on Computer Vision*, pages 409–424, 2018.

[8] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, apr 2017.

[9] Frank E. Curtis, Daniel P. Robinson, and Baoyu Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv preprint arXiv:2107.03512*, 2021.

[10] Nicholas I. M. Gould, Dominique Orban, and Philippe L. Toint. CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60(3):545–557, aug 2014.

[11] Sen Na, Mihai Anitescu, and Mladen Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *arXiv preprint arXiv:2109.11502*, 2021.

[12] Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, jun 2022.

[13] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer New York, 2006.

Appendix A. Proofs

A.1. Fundamental lemmas

Lemma 4 *Suppose Assumptions 1 and 2 hold. When $\mathcal{A}_k$ happens, for $k \ge \bar{K}$ we have*

$$\left|\mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \overline{Pred}_k\right| \le \frac{1}{2}(2\kappa_g + L_{\nabla f} + \kappa_B + \hat{\mu} L_G)\Delta_k^2. \tag{6}$$

Proof Let $\mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} := \mathcal{L}_{\bar{\mu}_{\bar{K}}}(\boldsymbol{x}_{s_k}), \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} := \mathcal{L}_{\bar{\mu}_{\bar{K}}}(\boldsymbol{x}_k)$, then for all $k \ge \bar{K}$, it follows from definitions of $\mathcal{L}_{\bar{\mu}_{\bar{K}}}(\boldsymbol{x})$ and $\overline{\text{Pred}}_k$ that

$$\begin{aligned}
\left|\mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \overline{\text{Pred}}_k\right| &= \left|f_{s_k} + \bar{\mu}_{\bar{K}}\|c_{s_k}\| - f_k - \bar{g}_k^T \Delta\boldsymbol{x}_k - \frac{1}{2}\Delta\boldsymbol{x}_k^T B_k \Delta\boldsymbol{x}_k - \bar{\mu}_{\bar{K}}\|c_k + G_k^T \Delta\boldsymbol{x}_k\|\right| \\
&\le \left|(\nabla f_k - \bar{g}_k)^T \Delta\boldsymbol{x}_k + \frac{1}{2}(L_{\nabla f} + \|B_k\| + \bar{\mu}_{\bar{K}} L_G)\|\Delta\boldsymbol{x}_k\|^2\right| \\
&\le \|\nabla f_k - \bar{g}_k\|\|\Delta\boldsymbol{x}_k\| + \frac{1}{2}(L_{\nabla f} + \kappa_B + \hat{\mu} L_G)\|\Delta\boldsymbol{x}_k\|^2,
\end{aligned} \tag{7}$$

in the first inequality, we use Taylor expansion and Lipschitz continuity of $\nabla f(\boldsymbol{x})$ and $G(\boldsymbol{x})$. In the last inequality, we use Cauchy Schwartz inequality, $\|B_k\| \le \kappa_B$ and $\bar{\mu}_{\bar{K}} \le \hat{\mu}$. Since $\mathcal{A}_k$ holds, we have $\|\nabla f_k - \bar{g}_k\| \le \kappa_g \Delta_k$, which combined with (7) and $\|\Delta\boldsymbol{x}_k\| \le \Delta_k$ yields the result. ■

Lemma 5 *Suppose Assumptions 1 and 2 hold, and also suppose $\mathcal{A}_k \cap \mathcal{B}_k$ happens and for $k \ge \bar{K}$,*

$$\|\bar{\nabla}\mathcal{L}_k\| \ge \max\left\{\max\{1, \tilde{\eta}_0\} \cdot \max\{\kappa_B, \sqrt{\kappa_{2,G}}\}, \frac{\Upsilon_1}{\kappa_{fcd}(1-\eta_0)}\right\}\Delta_k, \tag{8}$$

where $\Upsilon_1 = 4\kappa_f + 2\kappa_g + L_{\nabla f} + \kappa_B + \hat{\mu} L_G$, then $\Delta\boldsymbol{x}_k$ is a successful trial step.

Proof To show that $\Delta\boldsymbol{x}_k$ is a successful step, we need to show that both conditions in (5) hold. Noticing that $\|B_k\| \le \kappa_B, \|G_k\| \le \sqrt{\kappa_{2,G}}$, it is implied by (8) that $\bar{\rho}_2 \ge \tilde{\eta}_0 \cdot \max\{\|B_k\|, \|G_k\|, 1\}$, so the second condition in (5) holds trivially. Next we need to show that $\bar{\rho}_1 > \eta_0$. From the definition of $\bar{\rho}_1$, we have

$$\bar{\rho}_1 = \frac{\overline{\text{Ared}}_k}{\overline{\text{Pred}}_k} = \frac{\bar{\mathcal{L}}^{s_k}_{\bar{\mu}_{\bar{K}}} - \bar{\mathcal{L}}^{k}_{\bar{\mu}_{\bar{K}}}}{\overline{\text{Pred}}_k} = \frac{\bar{\mathcal{L}}^{s_k}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} + \mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} - m^k_{\bar{\mu}_{\bar{K}}}(\Delta\boldsymbol{x}_k) + \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \bar{\mathcal{L}}^{k}_{\bar{\mu}_{\bar{K}}}}{\overline{\text{Pred}}_k} + 1,$$

since $\overline{\text{Pred}}_k = m^k_{\bar{\mu}_{\bar{K}}}(\Delta\boldsymbol{x}_k) - m^k_{\bar{\mu}_{\bar{K}}}(\boldsymbol{0})$ and $m^k_{\bar{\mu}_{\bar{K}}}(\boldsymbol{0}) - \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} = 0$. Rearranging the terms, we find

$$|\bar{\rho}_1 - 1| \le \frac{\left|\bar{\mathcal{L}}^{s_k}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}}\right| + \left|\mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} - m^k_{\bar{\mu}_{\bar{K}}}(\Delta\boldsymbol{x}_k)\right| + \left|\mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \bar{\mathcal{L}}^{k}_{\bar{\mu}_{\bar{K}}}\right|}{\left|\overline{\text{Pred}}_k\right|}. \tag{9}$$

Now we consider the four terms on the right-hand-side of (9). We first note that (8) shows $\Delta_k \le \frac{1}{\max\{\kappa_B, \sqrt{\kappa_{2,G}}\}}\|\bar{\nabla}\mathcal{L}_k\|$, which combined with (2) and $\|B_k\| \le \kappa_B, \|G_k\| \le \sqrt{\kappa_{2,G}}$ yields

$$\tilde{\Delta}_k = \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}\Delta_k \le \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}, \text{ and } \breve{\Delta}_k = \frac{\|c_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}\Delta_k \le \frac{\|c_k\|}{\|G_k\|}. \tag{10}$$

Therefore,

$$\begin{aligned}
\overline{\text{Pred}}_k &\overset{(4)}{\leq} -\frac{\kappa_{fcd}}{2}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\min\left\{\tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{2}\|c_k\|\min\left\{\breve{\Delta}_k, \frac{\|c_k\|}{\|G_k\|}\right\} \\
&\overset{(8),(10)}{=} -\frac{\kappa_{fcd}}{2}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k - \frac{\kappa_{fcd}}{2}\|c_k\|\breve{\Delta}_k \\
&\overset{(2)}{=} -\frac{\kappa_{fcd}}{2}\frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2}{\|\bar{\nabla}\mathcal{L}_k\|}\Delta_k - \frac{\kappa_{fcd}}{2}\frac{\|c_k\|^2}{\|\bar{\nabla}\mathcal{L}_k\|}\Delta_k \\
&= -\frac{\kappa_{fcd}}{2}\|\bar{\nabla}\mathcal{L}_k\|\Delta_k, \qquad (\text{since}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2 = \|\bar{\nabla}\mathcal{L}_k\|^2) \qquad (11)
\end{aligned}$$

which implies

$$\left|\overline{\text{Pred}}_k\right| \geq \frac{\kappa_{fcd}}{2}\|\bar{\nabla}\mathcal{L}_k\|\Delta_k. \qquad (12)$$

Next, we consider $\left|\bar{\mathcal{L}}^k_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}}\right|$ and $\left|\bar{\mathcal{L}}^{s_k}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}}\right|$. From the definitions of $\mathcal{L}_{\bar{\mu}}(\boldsymbol{x})$ and $\bar{\mathcal{L}}_{\bar{\mu}}(\boldsymbol{x})$, it is easy to check that $\mathcal{B}_k$ holds is equivalently to

$$\max\left\{\left|\bar{\mathcal{L}}^k_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}}\right|, \left|\bar{\mathcal{L}}^{s_k}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}}\right|\right\} \leq \kappa_f\Delta_k^2. \qquad (13)$$

For the last term $\left|\mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} - m^k_{\bar{\mu}_{\bar{K}}}(\Delta\boldsymbol{x}_k)\right|$, we note that since $\overline{\text{Pred}}_k = m^k_{\bar{\mu}_{\bar{K}}}(\Delta\boldsymbol{x}_k) - m^k_{\bar{\mu}_{\bar{K}}}(\mathbf{0})$ and $m^k_{\bar{\mu}_{\bar{K}}}(\mathbf{0}) - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} = 0$, when $\mathcal{A}_k$ holds

$$\begin{aligned}
\left|\mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} - m^k_{\bar{\mu}_{\bar{K}}}(\Delta\boldsymbol{x}_k)\right| &= \left|\mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} + m^k_{\bar{\mu}_{\bar{K}}}(\mathbf{0}) - m^k_{\bar{\mu}_{\bar{K}}}(\Delta\boldsymbol{x}_k)\right| \\
&= \left|\mathcal{L}^{s_k}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} - \overline{\text{Pred}}_k\right| \\
&\overset{(6)}{\leq} \frac{1}{2}(2\kappa_g + L_{\nabla f} + \kappa_B + \hat{\mu}L_G)\Delta_k^2. \qquad (14)
\end{aligned}$$

Combining (9), (12), (13) and (14), we have

$$|\bar{\rho}_1 - 1| \leq \frac{(4\kappa_f + 2\kappa_g + L_{\nabla f} + \kappa_B + \hat{\mu}L_G)\Delta_k}{\kappa_{fcd}\|\bar{\nabla}\mathcal{L}_k\|} \overset{(8)}{\leq} 1 - \eta_0,$$

equivalently, $\bar{\rho}_1 \geq \eta_0$. Thus (5) holds and $\Delta\boldsymbol{x}_k$ is a successful step. ■

Lemma 6 *Suppose Assumptions 1 and 2 hold, and also suppose $\mathcal{B}_k$ happens and $\Delta\boldsymbol{x}_k$ is a successful step, then for $k \geq \bar{K}$, we have*

$$\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} \leq -\Upsilon_2\Delta_k^2,$$

where $\Upsilon_2 = \frac{\kappa_{fcd}}{4}\eta_0\tilde{\eta}_0\min\{1, \tilde{\eta}_0\}$.

Proof Since the $\Delta\boldsymbol{x}_k$ is successful, we have $\boldsymbol{x}_{k+1} = \boldsymbol{x}_{s_k}$ and both conditions in (5) hold. Note that $\bar{\rho}_2 \geq \tilde{\eta}_0 \cdot \max\{\|B_k\|, \|G_k\|, 1\}$ is equivalent to $\min\left\{\|\bar{\nabla}\mathcal{L}_k\|, \frac{\|\bar{\nabla}\mathcal{L}_k\|}{\max\{\|B_k\|,\|G_k\|\}}\right\} \geq \tilde{\eta}_0\Delta_k$. It is easy to check that $\|\bar{\nabla}\mathcal{L}_k\| \geq \tilde{\eta}_0\Delta_k$ leads to $\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| \geq \tilde{\eta}_0\tilde{\Delta}_k$ and $\|c_k\| \geq \tilde{\eta}_0\breve{\Delta}_k$, $\frac{\|\bar{\nabla}\mathcal{L}_k\|}{\max\{\|B_k\|,\|G_k\|\}} \geq \tilde{\eta}_0\Delta_k$ leads to $\frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|} \geq \tilde{\eta}_0\tilde{\Delta}_k$ and $\frac{\|c_k\|}{\|G_k\|} \geq \tilde{\eta}_0\breve{\Delta}_k$. Therefore, when the trial step is successful, we have

$$\begin{aligned}\overline{\text{Ared}}_k \overset{(5)}{\leq} \eta_0\overline{\text{Pred}}_k \overset{(4)}{\leq} & -\frac{\kappa_{fcd}}{2}\eta_0\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\min\left\{\tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{2}\eta_0\|c_k\|\min\left\{\breve{\Delta}_k, \frac{\|c_k\|}{\|G_k\|}\right\} \\ \leq & -\frac{\kappa_{fcd}}{2}\eta_0\tilde{\eta}_0\min\{1,\tilde{\eta}_0\}\tilde{\Delta}_k^2 - \frac{\kappa_{fcd}}{2}\eta_0\tilde{\eta}_0\min\{1,\tilde{\eta}_0\}\breve{\Delta}_k^2 \\ = & -\frac{\kappa_{fcd}}{2}\eta_0\tilde{\eta}_0\min\{1,\tilde{\eta}_0\}\Delta_k^2, \end{aligned} \tag{15}$$

in the last equality, we use the fact that $\breve{\Delta}_k^2 + \tilde{\Delta}_k^2 = \Delta_k^2$. When $\mathcal{B}_k$ holds and $\Delta\boldsymbol{x}_k$ is successful, we have

$$\max\left\{\left|\bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^k - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^k\right|, \left|\bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}\right|\right\} \leq \kappa_f\Delta_k^2 = \frac{\kappa_{fcd}}{16}\eta_0\tilde{\eta}_0\min\{1,\tilde{\eta}_0\}\Delta_k^2. \tag{16}$$

Therefore, since $\overline{\text{Ared}}_k = \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k+1} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^k$ we have

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^k = & \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k+1} + \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k+1} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^k + \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^k - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^k \\ \leq & \left|\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k+1}\right| + \overline{\text{Ared}}_k + \left|\bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^k - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^k\right|. \end{aligned} \tag{17}$$

The conclusion follows by combining (15), (16) and (17). ■

In the condition of ν, Υ_2 is defined in Lemma 6,

$$\Upsilon_3 = \frac{\kappa_{fcd}}{8}\max\left\{\frac{\max\{1,\tilde{\eta}_0\}\cdot\max\{\kappa_B,\sqrt{\kappa_{2,G}}\}}{\max\{1,\tilde{\eta}_0\}\cdot\max\{\kappa_B,\sqrt{\kappa_{2,G}}\}+\kappa_g}, \frac{4\Upsilon_1}{4\Upsilon_1+(1-\eta_0)\kappa_{fcd}\kappa_g}\right\},$$

where Υ_1 is defined in Lemma 5. We choose ζ such that

$$\zeta \geq \kappa_g + \max\left\{\max\{1,\tilde{\eta}_0\}\cdot\max\{\kappa_B,\sqrt{\kappa_{2,G}}\}, \frac{4\Upsilon_1}{\kappa_{fcd}(1-\eta_0)}\right\}. \tag{18}$$

Lemma 7 *Suppose conditions of Theorem 3 are satisfied, then for $k \geq \bar{K}$, we have*

$$\mathbb{E}_k[\Phi_{\bar{\mu}_{\bar{K}}}^{k+1}] - \Phi_{\bar{\mu}_{\bar{K}}}^k \leq \frac{1-\nu}{4}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2. \tag{19}$$

Proof First note that if the step is not successful, then $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k, \Delta_{k+1} = \Delta_k/\gamma, \bar{\epsilon}_{k+1} = \bar{\epsilon}_k/\gamma$, and

$$\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^k = \frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar{\epsilon}_k. \tag{20}$$

For $k \geq \bar{K}$, we consider two cases separately: $\|\nabla\mathcal{L}_k\| \geq \zeta\Delta_k$ and $\|\nabla\mathcal{L}_k\| < \zeta\Delta_k$, where ζ is chosen to satisfy (18).

Case 1: $\|\nabla\mathcal{L}_k\| \geq \zeta\Delta_k$

a. $\mathcal{A}_k \cap \mathcal{B}_k$ **happens.** When $\mathcal{A}_k$ holds, we have

$$\left|\|\nabla\mathcal{L}_k\| - \|\bar{\nabla}\mathcal{L}_k\|\right| \leq \|\nabla\mathcal{L}_k - \bar{\nabla}\mathcal{L}_k\| = \|P_k(\nabla f_k - \bar{g}_k)\| \leq \|\nabla f_k - \bar{g}_k\| \leq \kappa_g\Delta_k,$$

therefore, $\|\bar{\nabla}\mathcal{L}_k\| \geq \|\nabla\mathcal{L}_k\| - \kappa_g\Delta_k \geq (\zeta - \kappa_g)\Delta_k$. Thus $\|\nabla\mathcal{L}_k\| \geq \zeta\Delta_k$ and (18) together imply

$$\|\bar{\nabla}\mathcal{L}_k\| \geq \max\left\{\max\{1, \tilde{\eta}_0\} \cdot \max\{\kappa_B, \sqrt{\kappa_{2,G}}\}, \frac{4\Upsilon_1}{\kappa_{fcd}(1-\eta_0)}\right\}\Delta_k. \tag{21}$$

Therefore, (8) holds and Lemma 5 shows that the trial step is successful.
Reliable step: When $\Delta\boldsymbol{x}_k$ is a reliable step, we find

$$\begin{aligned}
\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{s_k} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} &\overset{(6)}{\leq} \overline{\text{Pred}}_k + \frac{1}{2}(2\kappa_g + L_{\nabla f} + \kappa_B + \hat{\mu}L_G)\Delta_k^2 \\
&\leq \frac{1}{2}\overline{\text{Pred}}_k - \frac{1}{2}\bar{\epsilon}_k + \frac{1}{2}(2\kappa_g + L_{\nabla f} + \kappa_B + \hat{\mu}L_G)\Delta_k^2 \\
&\overset{(4)}{\leq} -\frac{\kappa_{fcd}}{4}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| \min\left\{\tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{4}\|c_k\| \min\left\{\breve{\Delta}_k, \frac{\|c_k\|}{\|G_k\|}\right\} \\
&\quad - \frac{1}{2}\bar{\epsilon}_k + \frac{1}{2}(2\kappa_g + L_{\nabla f} + \kappa_B + \hat{\mu}L_G)\Delta_k^2.
\end{aligned}$$

Using derivation similar to (11), we have

$$\begin{aligned}
\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{s_k} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} &\leq -\frac{\kappa_{fcd}}{4}\|\bar{\nabla}\mathcal{L}_k\|\Delta_k - \frac{1}{2}\bar{\epsilon}_k + \frac{1}{2}(2\kappa_g + L_{\nabla f} + \kappa_B + \hat{\mu}L_G)\Delta_k^2 \\
&\overset{(21)}{\leq} -\frac{\kappa_{fcd}}{8}\|\bar{\nabla}\mathcal{L}_k\|\Delta_k - \frac{1}{2}\bar{\epsilon}_k.
\end{aligned} \tag{22}$$

Combining (21), $\|\nabla\mathcal{L}_k\| \geq \zeta\Delta_k$ and $\|\bar{\nabla}\mathcal{L}_k\| \geq \|\nabla\mathcal{L}_k\| - \kappa_g\Delta_k$, we also have

$$\begin{aligned}
\|\bar{\nabla}\mathcal{L}_k\| \geq &\|\nabla\mathcal{L}_k\| \\
&- \kappa_g \min\left\{\frac{1}{\max\{1, \tilde{\eta}_0\} \cdot \max\{\kappa_B, \sqrt{\kappa_{2,G}}\} + \kappa_g}, \frac{\kappa_{fcd}(1-\eta_0)}{4\Upsilon_1 + (1-\eta_0)\kappa_{fcd}\kappa_g}\right\}\|\nabla\mathcal{L}_k\| \\
\geq &\max\left\{\frac{\max\{1, \tilde{\eta}_0\} \cdot \max\{\kappa_B, \sqrt{\kappa_{2,G}}\}}{\max\{1, \tilde{\eta}_0\} \cdot \max\{\kappa_B, \sqrt{\kappa_{2,G}}\} + \kappa_g}, \frac{4\Upsilon_1}{4\Upsilon_1 + (1-\eta_0)\kappa_{fcd}\kappa_g}\right\}\|\nabla\mathcal{L}_k\|.
\end{aligned} \tag{23}$$

Combining (22) with (23), we have

$$\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{s_k} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq -\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k - \frac{1}{2}\bar{\epsilon}_k,$$

where $\Upsilon_3 = \frac{\kappa_{fcd}}{8} \cdot \max\left\{\frac{\max\{1,\tilde{\eta}_0\}\cdot\max\{\kappa_B,\sqrt{\kappa_{2,G}}\}}{\max\{1,\tilde{\eta}_0\}\cdot\max\{\kappa_B,\sqrt{\kappa_{2,G}}\}+\kappa_g}, \frac{4\Upsilon_1}{4\Upsilon_1+(1-\eta_0)\kappa_{fcd}\kappa_g}\right\}$. Therefore, when $\Delta\boldsymbol{x}_k$ is a reliable step, we have

$$\begin{aligned}
\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} &\leq -\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k - \frac{1}{2}\nu\bar{\epsilon}_k + \frac{1-\nu}{2}(\gamma^2 - 1)\Delta_k^2 + \frac{1-\nu}{2}(\gamma - 1)\bar{\epsilon}_k \\
&\leq -\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k - \frac{1}{4}\nu\bar{\epsilon}_k + \frac{1-\nu}{2}(\gamma^2 - 1)\Delta_k^2,
\end{aligned} \tag{24}$$

since $\frac{\nu}{1-\nu} \geq \frac{2\gamma}{\eta_0}$ and $\eta_0 \leq 1$ imply

$$-\frac{1}{2}\nu\bar{\epsilon}_k + \frac{1-\nu}{2}(\gamma - 1)\bar{\epsilon}_k \leq -\frac{1}{4}\nu\bar{\epsilon}_k.$$

Moreover, since $\|\nabla\mathcal{L}_k\| \geq \zeta\Delta_k$, we have

$$-\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k + \frac{1-\nu}{2}(\gamma^2-1)\Delta_k^2 \leq \left[-\nu\Upsilon_3\zeta + \frac{1-\nu}{2}(\gamma^2-1)\right]\Delta_k^2 \leq 0.$$

Unreliable step: When $\Delta\boldsymbol{x}_k$ is unreliable, one finds

$$\begin{aligned}\mathcal{L}^{s_k}_{\bar\mu_{\bar K}} - \mathcal{L}^{k}_{\bar\mu_{\bar K}} &\overset{(6)}{\leq} \overline{\text{Pred}}_k + \frac{1}{2}(2\kappa_g + L_{\nabla f} + \kappa_B + \hat\mu L_G)\Delta_k^2 \\ &\overset{(4)}{\leq} -\frac{\kappa_{fcd}}{2}\|\bar\nabla_{\boldsymbol{x}}\mathcal{L}_k\|\min\left\{\tilde\Delta_k, \frac{\|\bar\nabla_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{2}\|c_k\|\min\left\{\breve\Delta_k, \frac{\|c_k\|}{\|G_k\|}\right\} \\ &\quad + \frac{1}{2}(2\kappa_g + L_{\nabla f} + \kappa_B + \hat\mu L_G)\Delta_k^2.\end{aligned}$$

Using similar derivation to (22), we have

$$\mathcal{L}^{s_k}_{\bar\mu_{\bar K}} - \mathcal{L}^{k}_{\bar\mu_{\bar K}} \overset{(21)}{\leq} -\frac{\kappa_{fcd}}{4}\|\bar\nabla\mathcal{L}_k\|\Delta_k. \tag{25}$$

Combining (23) and (25), we have

$$\mathcal{L}^{s_k}_{\bar\mu_{\bar K}} - \mathcal{L}^{k}_{\bar\mu_{\bar K}} \leq -2\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k,$$

where Υ_3 is defined above, and thus

$$\Phi^{k+1}_{\bar\mu_{\bar K}} - \Phi^{k}_{\bar\mu_{\bar K}} \leq -2\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k + \frac{1-\nu}{2}(\gamma^2-1)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar\epsilon_k. \tag{26}$$

Since $\gamma - 1 \geq 1 - \frac{1}{\gamma}$ that

$$-\frac{1}{4}\nu\bar\epsilon_k \leq \frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar\epsilon_k.$$

Thus when $\mathcal{A}_k \cap \mathcal{B}_k$ holds, we always have

$$\Phi^{k+1}_{\bar\mu_{\bar K}} - \Phi^{k}_{\bar\mu_{\bar K}} \leq -\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k + \frac{1-\nu}{2}(\gamma^2-1)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar\epsilon_k. \tag{27}$$

b. $\mathcal{A}_k \cap \mathcal{B}_k^c$ happens. Using similar analysis as in **Case 1. a.**, we have that if the trial step is reliable, then (24) holds, if the trial step is unreliable, then (26) holds. Thus (27) is guaranteed. If the step is unsuccessful, then (20) holds. It is implied by $\|\nabla\mathcal{L}_k\| \geq \zeta\Delta_k$ that

$$-\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k + \frac{1-\nu}{2}(\gamma^2-1)\Delta_k^2 \leq \frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2. \tag{28}$$

Therefore when $\mathcal{A}_k \cap \mathcal{B}_k^c$ holds, we always have

$$\Phi^{k+1}_{\bar\mu_{\bar K}} - \Phi^{k}_{\bar\mu_{\bar K}} \leq \frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar\epsilon_k. \tag{29}$$

c. $\mathcal{A}_k^c \cap \mathcal{B}_k$ happens. We consider the following three cases.
Reliable step: When $\Delta \boldsymbol{x}_k$ is reliable, we find

$$\begin{aligned}\overline{\text{Ared}}_k \overset{(5)}{\leq} \eta_0 \overline{\text{Pred}}_k \leq & \frac{1}{2}\eta_0 \overline{\text{Pred}}_k - \frac{1}{2}\eta_0 \bar{\epsilon}_k \\ \overset{(15)}{\leq} & -\frac{\kappa_{fcd}}{4}\eta_0 \tilde{\eta}_0 \min\{1, \tilde{\eta}_0\} \Delta_k^2 - \frac{1}{2}\eta_0 \bar{\epsilon}_k \\ = & -\Upsilon_2 \Delta_k^2 - \frac{1}{2}\eta_0 \bar{\epsilon}_k. \end{aligned} \tag{30}$$

Combining (16), (17) with (30), we have

$$\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{s_k} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq -\frac{1}{2}\Upsilon_2 \Delta_k^2 - \frac{1}{2}\eta_0 \bar{\epsilon}_k.$$

Therefore,

$$\begin{aligned}\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq & -\frac{1}{2}\nu \Upsilon_2 \Delta_k^2 - \frac{1}{2}\nu \eta_0 \bar{\epsilon}_k + \frac{1-\nu}{2}(\gamma^2 - 1)\Delta_k^2 + \frac{1-\nu}{2}(\gamma - 1)\bar{\epsilon}_k \\ \leq & -\frac{1}{2}\nu \Upsilon_2 \Delta_k^2 - \frac{1}{4}\nu \eta_0 \bar{\epsilon}_k + \frac{1-\nu}{2}(\gamma^2 - 1)\Delta_k^2, \end{aligned} \tag{31}$$

since $\frac{\nu}{1-\nu} \geq \frac{2\gamma}{\eta_0}$ implies

$$-\frac{1}{2}\nu \eta_0 \bar{\epsilon}_k + \frac{1-\nu}{2}(\gamma - 1)\bar{\epsilon}_k \leq -\frac{1}{4}\nu \eta_0 \bar{\epsilon}_k. \tag{32}$$

Unreliable step: When the step $\Delta \boldsymbol{x}_k$ is unreliable, it follows from Lemma 6 that

$$\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{s_k} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq -\Upsilon_2 \Delta_k^2,$$

therefore

$$\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq -\nu \Upsilon_2 \Delta_k^2 + \frac{1-\nu}{2}(\gamma^2 - 1)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma} - 1\right)\bar{\epsilon}_k. \tag{33}$$

Unsuccessful step: When $\Delta \boldsymbol{x}_k$ is unsuccessful, (20) holds.

Combining (31), (33) and (20), noting that

$$-\frac{1}{2}\nu \Upsilon_2 \Delta_k^2 + \frac{1-\nu}{2}(\gamma^2 - 1)\Delta_k^2 \leq \frac{1-\nu}{2}\left(\frac{1}{\gamma^2} - 1\right)\Delta_k^2, \tag{34}$$

and since $\gamma - 1 \geq 1 - 1/\gamma$,

$$-\frac{1}{4}\nu \eta_0 \bar{\epsilon}_k \leq \frac{1-\nu}{2}\left(\frac{1}{\gamma} - 1\right)\bar{\epsilon}_k, \tag{35}$$

when $\mathcal{A}_k^c \cap \mathcal{B}_k$ holds, we always have

$$\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq \frac{1-\nu}{2}\left(\frac{1}{\gamma^2} - 1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma} - 1\right)\bar{\epsilon}_k. \tag{36}$$

d. $\mathcal{A}_k^c \cap \mathcal{B}_k^c$ happens. We consider the following three cases. When $\Delta \boldsymbol{x}_k$ is successful, we have

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} =& \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{s_k} + \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{s_k} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k} + \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \\ \leq& \left|\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{s_k}\right| + \overline{\text{Ared}}_k + \left|\bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k}\right|, \\ =& \left|f_{s_k} - \bar{f}_{s_k}\right| + \overline{\text{Ared}}_k + \left|\bar{f}_k - f_k\right|,\end{aligned}$$

since $\overline{\text{Ared}}_k = \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k+1} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k}$, $\left|\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{s_k}\right| = \left|f_{s_k} - \bar{f}_{s_k}\right|$ and $\left|\bar{\mathcal{L}}_{\bar{\mu}_{\bar{K}}}^{k} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k}\right| = \left|\bar{f}_k - f_k\right|$.
Reliable step: When $\Delta \boldsymbol{x}_k$ is reliable, one finds

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq& \left|f_{s_k} - \bar{f}_{s_k}\right| + \overline{\text{Ared}}_k + \left|\bar{f}_k - f_k\right| \\ \overset{(30)}{\leq}& \left|\bar{f}_k - f_k\right| + \left|\bar{f}_{s_k} - f_{s_k}\right| - \Upsilon_2 \Delta_k^2 - \frac{1}{2}\eta_0 \bar{\epsilon}_k.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq& \nu\left|\bar{f}_k - f_k\right| + \nu\left|\bar{f}_{s_k} - f_{s_k}\right| - \nu\Upsilon_2\Delta_k^2 - \frac{1}{2}\nu\eta_0\bar{\epsilon}_k + \frac{1-\nu}{2}(\gamma^2 - 1)\Delta_k^2 \\ &+ \frac{1-\nu}{2}(\gamma - 1)\bar{\epsilon}_k \\ \overset{(32),(34),(35)}{\leq}& \nu\left|\bar{f}_k - f_k\right| + \nu\left|\bar{f}_{s_k} - f_{s_k}\right| + \frac{1-\nu}{2}\left(\frac{1}{\gamma^2} - 1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma} - 1\right)\bar{\epsilon}_k.\end{aligned}$$

Unreliable step: When $\Delta \boldsymbol{x}_k$ is unreliable, one finds

$$\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \overset{(15)}{\leq} \left|\bar{f}_k - f_k\right| + \left|\bar{f}_{s_k} - f_{s_k}\right| - 2\Upsilon_2\Delta_k^2,$$

and therefore

$$\begin{aligned}\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq& \nu\left|\bar{f}_k - f_k\right| + \nu\left|\bar{f}_{s_k} - f_{s_k}\right| - 2\nu\Upsilon_2\Delta_k^2 + \frac{1-\nu}{2}(\gamma^2 - 1)\Delta_k^2 \\ &+ \frac{1-\nu}{2}\left(\frac{1}{\gamma} - 1\right)\bar{\epsilon}_k \\ \overset{(34)}{\leq}& \nu\left|\bar{f}_k - f_k\right| + \nu\left|\bar{f}_{s_k} - f_{s_k}\right| + \frac{1-\nu}{2}\left(\frac{1}{\gamma^2} - 1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma} - 1\right)\bar{\epsilon}_k.\end{aligned}$$

Unsuccessful step: If the step is not successful, we have (20).

It is easy to check that when $\mathcal{A}_k^c \cap \mathcal{B}_k^c$ holds, we always have

$$\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq \nu\left|\bar{f}_k - f_k\right| + \nu\left|\bar{f}_{s_k} - f_{s_k}\right| + \frac{1-\nu}{2}\left(\frac{1}{\gamma^2} - 1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma} - 1\right)\bar{\epsilon}_k. \quad (37)$$

Now we take expectation on reduction of Φ when $\|\nabla\mathcal{L}_k\| \geq \zeta\Delta_k$ given $\mathcal{F}_{k-1}$. Note that the event $\mathcal{A}_k \cap \mathcal{B}_k$ holds with probability at least $p_{grad}p_f$, event $\mathcal{A}_k^c \cap \mathcal{B}_k^c$ holds with probability at most $(1 - p_{grad})(1 - p_f)$, otherwise $\mathcal{A}_k^c \cap \mathcal{B}_k$ and $\mathcal{A}_k \cap \mathcal{B}_k^c$ hold. We use $\mathbb{E}_k$ to denote $\mathbb{E}[\cdot|\mathcal{F}_{k-1}]$, then when $\|\nabla\mathcal{L}_k\| \geq \zeta\Delta_k$, we have

$$\begin{aligned}\mathbb{E}_k[\Phi_{\bar{\mu}_{\bar{K}}}^{k+1}] - \Phi_{\bar{\mu}_{\bar{K}}}^{k} =& \mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k\cap\mathcal{B}_k}(\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k})] + \mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k^c\cap\mathcal{B}_k}(\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k})] \\ &+ \mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k\cap\mathcal{B}_k^c}(\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k})] + \mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k^c\cap\mathcal{B}_k^c}(\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k})].\end{aligned}$$

From the analysis of **Case 1. a. b. c.**, combining (27), (28), (29), and (36), we have

$$\begin{aligned}
&\mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k\cap\mathcal{B}_k}(\Phi^{k+1}_{\bar{\mu}_{\bar{K}}}-\Phi^{k}_{\bar{\mu}_{\bar{K}}})]+\mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k^c\cap\mathcal{B}_k}(\Phi^{k+1}_{\bar{\mu}_{\bar{K}}}-\Phi^{k}_{\bar{\mu}_{\bar{K}}})]+\mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k\cap\mathcal{B}_k^c}(\Phi^{k+1}_{\bar{\mu}_{\bar{K}}}-\Phi^{k}_{\bar{\mu}_{\bar{K}}})]\\
\le& p_{grad}p_f\left[-\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k+\frac{1-\nu}{2}(\gamma^2-1)\Delta_k^2+\frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar{\epsilon}_k\right]\\
&+[(1-p_{grad})p_f+(1-p_f)p_{grad}]\left[\frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2+\frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar{\epsilon}_k\right]. \qquad (38)
\end{aligned}$$

For **Case 1. d.**, by (37) and $\max\left\{\mathbb{E}_{\xi_f^k}\left[|\bar{f}_k-f_k|^2\right],\mathbb{E}_{\xi_f^k}\left[|\bar{f}_{s_k}-f_{s_k}|^2\right]\right\}\le\bar{\epsilon}_k^2$, we have

$$\begin{aligned}
&\mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k^c\cap\mathcal{B}_k^c}(\Phi^{k+1}_{\bar{\mu}_{\bar{K}}}-\Phi^{k}_{\bar{\mu}_{\bar{K}}})]\\
\le&\frac{1}{2}(1-p_{grad})(1-p_f)(1-\nu)\left[\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2+\left(\frac{1}{\gamma}-1\right)\bar{\epsilon}_k\right]+\nu\mathbb{E}_k\left[\mathbf{1}_{\mathcal{A}_k^c\cap\mathcal{B}_k^c}(|\bar{f}_k-f_k|)\right]\\
&+\nu\mathbb{E}_k\left[\mathbf{1}_{\mathcal{A}_k^c\cap\mathcal{B}_k^c}(|\bar{f}_{s_k}-f_{s_k}|)\right]\\
\le&\frac{1}{2}(1-p_{grad})(1-p_f)(1-\nu)\left[\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2+\left(\frac{1}{\gamma}-1\right)\bar{\epsilon}_k\right]\\
&+2\nu\sqrt{(1-p_{grad})(1-p_f)}\bar{\epsilon}_k, \qquad (39)
\end{aligned}$$

in the last inequality we use Hölder's inequality. Combining (38), (39) and rearranging the terms, we have

$$\begin{aligned}
&\mathbb{E}_k[\Phi^{k+1}_{\bar{\mu}_{\bar{K}}}]-\Phi^{k}_{\bar{\mu}_{\bar{K}}}\\
\le&\frac{1-\nu}{2}\left[p_{grad}p_f-\frac{1}{\gamma^2}[(1-p_{grad})(1-p_f)+(1-p_{grad})p_f+(1-p_f)p_{grad}]\right](\gamma^2-1)\Delta_k^2\\
&-p_{grad}p_f\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k+\left[\frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)+2\nu\sqrt{(1-p_{grad})(1-p_f)}\right]\bar{\epsilon}_k. \qquad (40)
\end{aligned}$$

Noting that

$$p_{grad}p_f-\frac{1}{\gamma^2}[(1-p_{grad})(1-p_f)+(1-p_{grad})p_f+(1-p_f)p_{grad}]\le p_{grad}p_f,$$

it follows from the combination of $\|\nabla\mathcal{L}_k\|\ge\zeta\Delta_k$ and (40) that

$$\begin{aligned}
\mathbb{E}_k[\Phi^{k+1}_{\bar{\mu}_{\bar{K}}}]-\Phi^{k}_{\bar{\mu}_{\bar{K}}}\le&-p_{grad}p_f\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k+\frac{1}{2}p_{grad}p_f(1-\nu)(\gamma^2-1)\Delta_k^2\\
&+\left[\frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)+2\nu\sqrt{(1-p_{grad})(1-p_f)}\right]\bar{\epsilon}_k\\
\le&-p_{grad}p_f\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k+\frac{1}{2}p_{grad}p_f(1-\nu)(\gamma^2-1)\Delta_k^2\\
\le&-\frac{1}{2}p_{grad}p_f\nu\Upsilon_3\|\nabla\mathcal{L}_k\|\Delta_k \qquad (41)
\end{aligned}$$

Case 2: $\|\nabla\mathcal{L}_k\|<\zeta\Delta_k$

a. $\mathcal{B}_k$ happens. Using identical proof as in **Case 1. c.**, it follows that no matter the trial step is successful or not, we always have

$$\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq \frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar{\epsilon}_k. \tag{42}$$

b. $\mathcal{B}_k^c$ happens. Using identical proof as in **Case 1. d.**, it follows that no matter the trial step is successful or not, we always have

$$\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq \nu|\bar{f}_k - f_k| + \nu|\bar{f}_{s_k} - f_{s_k}| + \frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar{\epsilon}_k. \tag{43}$$

Now we take expectation on $\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k}$ when $\|\nabla \mathcal{L}_k\| \leq \zeta\Delta_k$ given $\mathcal{F}_{k-1}$. Since $\mathcal{B}_k$ holds with probability at least p_f, otherwise $\mathcal{B}_k^c$ holds, it follows from the combination of (42) and (43) that

$$\begin{aligned}
&\mathbb{E}_k[\Phi_{\bar{\mu}_{\bar{K}}}^{k+1}] - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \\
=&\mathbb{E}_k[\mathbf{1}_{B_k}(\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k})] + \mathbb{E}_k[\mathbf{1}_{\mathcal{B}_k^c}(\Phi_{\bar{\mu}_{\bar{K}}}^{k+1} - \Phi_{\bar{\mu}_{\bar{K}}}^{k})] \\
\leq&\frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2 + \frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right)\bar{\epsilon}_k + \nu\mathbb{E}_k[\mathbf{1}_{\mathcal{B}_k^c}\left(\left|\bar{f}_k - f_k\right|\right)] + \nu\mathbb{E}_k[\mathbf{1}_{\mathcal{B}_k^c}\left(\left|\bar{f}_{s_k} - f_{s_k}\right|\right)] \\
\leq&\frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2 + \left[\frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right) + 2\nu\sqrt{1-p_f}\right]\bar{\epsilon}_k,
\end{aligned}$$

where in the last inequality, we use Hölder inequality. Notice that

$$\frac{1-\nu}{2}\left(\frac{1}{\gamma}-1\right) + 2\nu\sqrt{1-p_f} \leq 0.$$

Therefore, when $\|\nabla \mathcal{L}_k\| \leq \zeta\Delta_k$, we have

$$\mathbb{E}_k[\Phi_{\bar{\mu}_{\bar{K}}}^{k+1}] - \Phi_{\bar{\mu}_{\bar{K}}}^{k} \leq \frac{1-\nu}{2}\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2. \tag{44}$$

Combining the conclusion (41) in **Case 1** and the conclusion (44) in **Case 2**, also noting that $1 - 1/\gamma^2 \leq \gamma^2 - 1$ and $p_{grad}p_f \geq 1/2$, we show that (19) holds. ■

Corollary 8 *Under the conditions of Lemma 7, we have*

$$\lim_{k\to\infty} \Delta_k = 0 \ \textit{ with probability 1.}$$

Proof Taking total expectation on both sides of (19), summing up for all $k \geq \bar{K}$, and noting that Φ is bounded below, we find that

$$\mathbb{E}\left[\sum_{k=\bar{K}}^{\infty} \Delta_k^2\right] < \infty.$$

By Fubini's theorem, with probability 1,

$$\sum_{k=\bar{K}}^{\infty} \Delta_k^2 < \infty,$$

which implies that $\Delta_k \to 0$ with probability 1. This completes the proof. ■

Lemma 9 *Suppose the conditions of Lemma 7 are satisfied, then we have*

$$\liminf_{k\to\infty} \|\nabla \mathcal{L}_k\| = 0 \quad \textit{almost surely.}$$

Proof Suppose that the almost sure event $\Delta_k \to 0$ happens and we prove this theorem by contradiction. Let us assume that there exists $\epsilon > 0$ such that, with positive probability, for all $k \geq \bar{K}$, we have

$$\|\nabla \mathcal{L}_k\| \geq \epsilon.$$

Since $\Delta_k \to 0$, there exists k_0 such that for all $k \geq k_0$,

$$\Delta_k \leq b := \min\left\{\frac{\Delta_{\max}}{\gamma}, \frac{\epsilon}{\zeta}\right\}, \tag{45}$$

where ζ is chosen to satisfies (18). We define $r_k = \log_\gamma\left(\frac{\Delta_k}{b}\right)$, which satisfies $r_k \leq 0$ for all $k \geq k_0$. The main idea of the proof is to show that such r_k occurs only with probability zero, hence obtaining a contradiction with the initial assumption of $\|\nabla \mathcal{L}_k\| \geq \epsilon, \forall k \geq \bar{K}$.

We first show that $\{r_k\}_k$ is a submartingale. Consider some $k \geq k_0$ for which $\mathcal{A}_k$ and $\mathcal{B}_k$ both hold, which happens with probability at least $p_{grad}p_f$. Due to (45) and $\|\nabla \mathcal{L}_k\| \geq \epsilon$, we have exactly the same situation as in **Case 1. a**. Therefore, we conclude that the trial step obtained at the k-th iteration is successful. Since $\Delta_k \leq \frac{\Delta_{\max}}{\gamma}$, we have $\Delta_{k+1} = \gamma\Delta_k$. Consequently, $r_{k+1} = r_k + 1$. For all other outcomes of $\mathcal{A}_k$ and $\mathcal{B}_k$, which occur with total probability of at most $1 - p_{grad}p_f$, we have $\Delta_{k+1} \geq \gamma^{-1}\Delta_k$, consequently, $r_{k+1} \geq r_k - 1$. Moreover, since $p_{grad}p_f \geq 1/2$, we find

$$\mathbb{E}_k[r_{k+1}] \geq p_{grad}p_f(r_k + 1) + (1 - p_{grad}p_f)(r_k - 1) \geq r_k,$$

which implies that $\{r_k\}_k$ is a submartingale. Now we define $w_k = \sum_{i=0}^{k}(2 \cdot \mathbf{1}_{\mathcal{A}_k} \cdot \mathbf{1}_{\mathcal{B}_k} - 1)$. Note that $\{w_k\}_k$ is a submartingale since

$$\mathbb{E}_k[w_k] = \mathbb{E}_k[w_{k-1}] + \mathbb{E}_k[2 \cdot \mathbf{1}_{\mathcal{A}_k} \cdot \mathbf{1}_{\mathcal{B}_k} - 1] = w_{k-1} + 2\mathbb{E}_k[\mathbf{1}_{\mathcal{A}_k} \cdot \mathbf{1}_{\mathcal{B}_k}] - 1 \geq w_{k-1},$$

the last equality holds because $p_{grad}p_f \geq 1/2$. Also note that $\{w_k\}_k$ is on the same probability space as $\{r_k\}_k$. Since w_k has only ±1 increments, Theorem 4.4 of [8] shows that $\limsup_{k\to\infty} w_k = \infty$ holds with probability 1. By the construction of $\{r_k\}_k$ and $\{w_k\}_k$, we know that $r_k - r_{k_0} \geq w_k - w_{k_0}$. Therefore, r_k has to be positive infinitely often with probability one. This implies that for the sequence $\{r_k\}_k$, $r_k \leq 0$ for all $k \geq k_0$ occurs with probability zero. Therefore our assumption that $\|\nabla \mathcal{L}_k\| \geq \epsilon$ holds for all $k > \bar{K}$ with positive probability is false and we have

$$\liminf_{k\to\infty} \|\nabla \mathcal{L}_k\| = 0$$

holds almost surely. This completes the proof. ■

Lemma 10 *Suppose the conditions of Lemma 7 are satisfied. Fix $\epsilon > 0$ and define the sequence $\{K_\epsilon\}$ consisting of the natural numbers k for which $\|\nabla \mathcal{L}_k\| \geq \epsilon$. Then we have*

$$\sum_{k \in \{K_\epsilon\}} \Delta_k < \infty \quad \textit{with probability 1.}$$

Proof Suppose that $\Delta_k \to 0$ happens, then there exists k_0 such that $\Delta_k \leq \epsilon/\zeta, \forall k \geq k_0$, where ζ is chosen to satisfy (18). WLOG, we assume $k_0 \geq \bar{K}$. Let K denote the sequence of indices k such that $k \in K_\epsilon$ and $k \geq k_0$. Then for all $k \in K$, $\|\nabla \mathcal{L}_k\| \geq \zeta \Delta_k$ holds. It follows from (41) that

$$\mathbb{E}_k[\Phi^{k+1}_{\bar{\mu}_{\bar{K}}}] - \Phi^k_{\bar{\mu}_{\bar{K}}} \leq -\frac{1}{2} p_{grad} p_f \nu \Upsilon_3 \epsilon \Delta_k, \quad \forall k \in K.$$

Taking total expectation, noting that Lemma 7 implies that $\mathbb{E}[\Phi^k_{\bar{\mu}_{\bar{K}}}]$ is non-increasing and bounded below, we sum up the above inequalities for all $k \in K$ and get

$$\sum_{k \in K} \mathbb{E}[\Delta_k] < \infty.$$

By Fubini's theorem, this implies

$$\sum_{k \in K} \Delta_k < \infty \text{ with probability 1.}$$

Since $K_\epsilon \subset K \cup \{k \leq k_0\}$, k_0 is finite and $\Delta_k \leq \Delta_{\max}$, the statement follows. ■

A.2. Proof of Theorem 3

Proof Suppose that $\Delta_k \to 0$ happens and we will prove this theorem by contradiction. Let us assume that with some probability, there exists $\epsilon > 0$ and an infinite index set $\mathcal{K}_1 \subseteq \mathbb{N}$ such that $\|\nabla \mathcal{L}_k\| > 2\epsilon$ for all $k \in \mathcal{K}_1$. On the other hand, Lemma 9 shows that with probability 1, there exists an infinite index set $\mathcal{K}_2$ such that $\|\nabla \mathcal{L}_k\| \leq \epsilon$ for all $k \in \mathcal{K}_2$. They imply that with some nonzero probability, there are index sets $\{m_i\}_{i=0}^\infty \subset \mathbb{N}$ and $\{n_i\}_{i=0}^\infty \subset \mathbb{N}$ with $m_i < n_i$ for all $i \in \mathbb{N}$ such that

$$\|\nabla \mathcal{L}_{m_i}\| \geq 2\epsilon, \|\nabla \mathcal{L}_{n_i}\| < \epsilon, \text{ and } \|\nabla \mathcal{L}_k\| \geq \epsilon \text{ for all } k \in \{m_i + 1, \cdots, n_i - 1\}.$$

Since $\bar{K}$ is finite, WLOG, we assume that $\bar{K} + 1 \leq m_i < n_i$ for all $i \in \mathbb{N}$. By triangular inequality, we have

$$\epsilon < |\|\nabla \mathcal{L}_{n_i}\| - \|\nabla \mathcal{L}_{m_i}\|| \leq \sum_{j=m_i}^{n_i - 1} |\|\nabla \mathcal{L}_{j+1}\| - \|\nabla \mathcal{L}_j\|| \tag{46}$$

From Assumption 1, we find that $\nabla \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ is Lipschitz continuous in both $\boldsymbol{x}$ and $\boldsymbol{\lambda}$. Since we define $\boldsymbol{\lambda}_k = -[G_k G_k^T]^{-1} G_k \nabla f_k$, Assumption 1 implies that $\boldsymbol{\lambda}$ is also Lipschitz in $\boldsymbol{x}$, so there is a constant $L_{\nabla \mathcal{L}} > 0$ such that $\|\nabla \mathcal{L}_{j+1} - \nabla \mathcal{L}_j\| \leq L_{\nabla \mathcal{L}} \|\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\|$ holds for $\forall j \in \mathbb{N}$. It follows from (46) that

$$\begin{aligned}\epsilon \leq &\sum_{j=m_i}^{n_i-1} |\|\nabla \mathcal{L}_{j+1}\| - \|\nabla \mathcal{L}_j\|| \leq \sum_{j=m_i}^{n_i-1} \|\nabla \mathcal{L}_{j+1} - \nabla \mathcal{L}_j\| \\ \leq & L_{\nabla \mathcal{L}} \sum_{j=m_i}^{n_i-1} \|\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\| \leq L_{\nabla \mathcal{L}} \left(\Delta_{m_i} + \sum_{j=m_i+1}^{n_i-1} \Delta_j \right).\end{aligned}$$

Since Δ_k converges to zero, for any i large enough, $\Delta_{m_i} < \frac{\epsilon}{2L_{\nabla \mathcal{L}}}$ holds. Then $L_{\nabla \mathcal{L}} \sum_{j=m_i+1}^{n_i-1} \Delta_j > \frac{\epsilon}{2} > 0$. Since $\sum_i \sum_{j=m_i+1}^{n_i-1} \Delta_j \leq \sum_{j \in \{K_\epsilon\}} \Delta_j$, we then have $\sum_{j \in \{K_\epsilon\}} \Delta_j = \infty$. The above proof shows that if $\lim_{k \to \infty} \|\nabla \mathcal{L}_k\| = 0$ doesn't hold almost surely, then with positive probability, $\sum_{j \in \{K_\epsilon\}} \Delta_j = \infty$. This yields a contradiction to Lemma 10. Therefore, $\lim_{k \to \infty} \|\nabla \mathcal{L}_k\| = 0$ holds almost surely. ■

Appendix B. Behavior of the merit parameter

Assumption 11 *For all $k \in \mathbb{N}$, there exists some positive deterministic parameter $M_1 \in \mathbb{R}$, such that $\|\nabla f_k - \bar{g}_k\| \leq M_1$.*

Lemma 12 *Under the Assumptions 1 and 11, there exist a stochastic $\bar{K} < \infty$ and a deterministic constant $\hat{\mu}$, such that for $\forall k > \bar{K}$, $\bar{\mu}_k = \bar{\mu}_{\bar{K}} \leq \hat{\mu}$.*

Proof It suffices to show that (4) is always satisfied if $\bar{\mu}_k$ is larger than a threshold independent of k. First note that by the adaptive relaxation technique, $\|c_k + G_k \Delta \boldsymbol{x}_k\| - \|c_k\| = -\gamma_k \|c_k\|$, then

$$\begin{aligned}\overline{\text{Pred}}_k =& \bar{g}_k^T \Delta \boldsymbol{x}_k + \frac{1}{2} \Delta \boldsymbol{x}_k^T B_k \Delta \boldsymbol{x}_k + \bar{\mu}_k (\|c_k + G_k \Delta \boldsymbol{x}_k\| - \|c_k\|) \\ =& \bar{g}_k^T P_k \boldsymbol{u}_k + \frac{1}{2} \boldsymbol{u}_k^T P_k B_k P_k \boldsymbol{u}_k + \gamma_k (\bar{g}_k - \nabla f_k)^T \boldsymbol{v}_k + \gamma_k \nabla f_k^T \boldsymbol{v}_k + \gamma_k \boldsymbol{v}_k^T B_k P_k \boldsymbol{u}_k \\ &+ \frac{1}{2} \gamma_k^2 \boldsymbol{v}_k^T B_k \boldsymbol{v}_k - \bar{\mu}_k \gamma_k \|c_k\| \\ \leq& -\frac{\kappa_{fcd}}{2} \|\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_k\| \min \left\{ \tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_k\|}{\|B_k\|} \right\} + \gamma_k \|\bar{g}_k - \nabla f_k\| \|\boldsymbol{v}_k\| + \gamma_k \|\nabla f_k\| \|\boldsymbol{v}_k\| \\ &+ \gamma_k \|B_k\| \|\boldsymbol{v}_k\| \|P_k \boldsymbol{u}_k\| + \frac{1}{2} \gamma_k^2 \|\boldsymbol{v}_k\| \|B_k\| \|\boldsymbol{v}_k\| - \bar{\mu}_k \gamma_k \|c_k\| \\ =& -\frac{\kappa_{fcd}}{2} \|\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_k\| \min \left\{ \tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_k\|}{\|B_k\|} \right\} - \frac{\kappa_{fcd}}{2} \gamma_k \|\boldsymbol{v}_k\| \|c_k\| + \frac{\kappa_{fcd}}{2} \gamma_k \|\boldsymbol{v}_k\| \|c_k\| \\ &+ \gamma_k \|\bar{g}_k - \nabla f_k\| \|\boldsymbol{v}_k\| + \gamma_k \|\nabla f_k\| \|\boldsymbol{v}_k\| + \gamma_k \|B_k\| \|\boldsymbol{v}_k\| \|P_k \boldsymbol{u}_k\| \\ &+ \frac{1}{2} \gamma_k^2 \|\boldsymbol{v}_k\| \|B_k\| \|\boldsymbol{v}_k\| - \bar{\mu}_k \gamma_k \|c_k\|.\end{aligned}$$

Using Assumptions 1 and 11, also noting that $\|\boldsymbol{v}_k\| \le \|G_k^T[G_kG_k^T]^{-1}\|\|c_k\| \le \frac{1}{\sqrt{\kappa_{1,G}}}\|c_k\|, \gamma_k \le 1$, and $\|P_k\boldsymbol{u}_k\| \le \Delta_k \le \Delta_{\max}$, we have

$$\begin{aligned}\overline{\text{Pred}}_k \le & -\frac{\kappa_{fcd}}{2}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| \min\left\{\tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\| + \gamma_k\frac{\kappa_c\kappa_{fcd}}{2\sqrt{\kappa_{1,G}}}\|c_k\| \\ & + \gamma_k\frac{M_1}{\sqrt{\kappa_{1,G}}}\|c_k\| + \gamma_k\frac{\kappa_{\nabla f}}{\sqrt{\kappa_{1,G}}}\|c_k\| + \gamma_k\frac{\kappa_B\Delta_{\max}}{\sqrt{\kappa_{1,G}}}\|c_k\| + \gamma_k\frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\|c_k\| - \bar{\mu}_k\gamma_k\|c_k\| \\ = & -\frac{\kappa_{fcd}}{2}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| \min\left\{\tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\| \\ & + \left(\frac{2M_1 + 2\kappa_{\nabla f} + 2\kappa_B\Delta_{\max} + \kappa_c\kappa_{fcd}}{2\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}} - \bar{\mu}_k\right)\gamma_k\|c_k\|.\end{aligned}$$

Therefore, if

$$\bar{\mu}_k \ge \frac{2M_1 + 2\kappa_{\nabla f} + 2\kappa_B\Delta_{\max} + \kappa_c\kappa_{fcd}}{2\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}} := \hat{\mu}/\rho,$$

we have

$$\overline{\text{Pred}}_k \le -\frac{\kappa_{fcd}}{2}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| \min\left\{\tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\|. \tag{47}$$

It is implied by the computation of $\boldsymbol{v}_k$ that $c_k = -G_k\boldsymbol{v}_k$, which leads to $\|c_k\| = \|G_k\boldsymbol{v}_k\| \le \|G_k\|\|\boldsymbol{v}_k\|$, equivalently, $\|\boldsymbol{v}_k\| \ge \|c_k\|/\|G_k\|$. Hence,

$$-\frac{\kappa_{fcd}}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\| = -\frac{\kappa_{fcd}}{2}\|c_k\|\min\left\{\breve{\Delta}_k, \|\boldsymbol{v}_k\|\right\} \le -\frac{\kappa_{fcd}}{2}\|c_k\|\min\left\{\breve{\Delta}_k, \frac{\|c_k\|}{\|G_k\|}\right\}. \tag{48}$$

It follows from (47) and (48) that

$$\overline{\text{Pred}}_k \le -\frac{\kappa_{fcd}}{2}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| \min\left\{\tilde{\Delta}_k, \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|B_k\|}\right\} - \frac{\kappa_{fcd}}{2}\|c_k\|\min\left\{\breve{\Delta}_k, \frac{\|c_k\|}{\|G_k\|}\right\},$$

meaning that (4) is satisfied. At the same time, since we update $\bar{\mu}_k$ by a factor of ρ in each while loop, there must be a $\bar{K} < \infty$ such that $\bar{\mu}_k = \bar{\mu}_{\bar{K}}$ for all $k \ge \bar{K}$ and $\bar{\mu}_{\bar{K}} \le \hat{\mu}$. This completes the proof. ■