Fully Stochastic Trust-Region Sequential Quadratic Programming for Equality-Constrained Optimization Problems

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 propose a fully stochastic trust-region sequential quadratic programming (TR-StoSQP) algorithm to solve nonlinear optimization problems. The problems involve a stochastic objective and deterministic equality constraints. Under the fully stochastic setup, we suppose that only a single sample is generated in each iteration to estimate the objective gradient. Compared to the existing line-search StoSQP schemes, our algorithm allows one to employ indefinite Hessian matrices for SQP subproblems. The algorithm adaptively selects the radius of the trust region based on an input sequence $\{\beta_k\}$, the estimated KKT residual, and the estimated Lipschitz constants of the objective gradients and constraint Jacobians. To address the infeasibility issue of trust-region methods that arises in constrained optimization, we propose an *adaptive relaxation technique* to compute the trial step. In particular, we decompose the trial step into a normal step and a tangential step. Based on the ratios of the feasibility and optimality residuals to the full KKT residual, we decompose the full trust-region radius into two segments that are used to control the size of the normal and tangential steps, respectively. The normal step has a closed form, while the tangential step is solved from a trust-region subproblem, of which the Cauchy point is sufficient for our study. We establish the global *almost sure* convergence guarantee of TR-StoSQP, and demonstrate its empirical performance on a subset of problems in CUTEst test set.

1. Introduction

We consider the following constrained stochastic optimization problem

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

where $f:\mathbb{R}^d\to\mathbb{R}$ is a stochastic objective with $F(\cdot;\xi)$ being one of its realizations, $c:\mathbb{R}^d\to\mathbb{R}^m$ are deterministic equality constraints, and $\xi\sim\mathcal{P}$ is a random variable following the distribution $\mathcal{P}$. Problem (1) appears in numerous applications, including deep neural networks [6], optimal control [4], and PDE-constrained optimization [15].

There are various methods for solving constrained deterministic optimization problems, among which sequential quadratic programming (SQP) methods enjoy superior performance in practice for both small and large problems. The recent literature has focused on the design of different stochastic SQP algorithms (StoSQP) to solve constrained stochastic optimization problems in (1) under different problem setups. In the fully stochastic setup, where a single sample is generated in each iteration to estimate the objective gradient, [2] proposed an StoSQP algorithm in which a random projection procedure is designed to select the stepsize. Inspired by the line search procedure, the projection procedure constructs an interval in each iteration based on a prespecified sequence $\{\beta_k\}$ and the estimated Lipschitz constants of the objective gradients and constraint Jacobians, projects a random quantity onto the interval to decide the stepsize, and ensures that the projected stepsize satisfies a model reduction condition on the ℓ_1 merit function. Based on the design in [2], [1] developed a procedure that can handle rank-deficient Jacobians, [9] solved Newton systems inexactly, and [3] applied SVRG techniques to accelerate the algorithm. In the random model setup, where a batch of samples is generated in each iteration, [12] proposed an StoSQP algorithm that utilizes an exact augmented Lagrangian merit function and incorporates a stochastic line search procedure to select the stepsize. [11] further generalized the algorithm by allowing inequality constraints using an active-set method.

The aforementioned existing algorithms provably converge globally either in expectation or almost surely and have satisfying numerical performance under suitable simulation settings. However, they have three limitations. First, they are all line-search-based, that is, a search direction is first computed by solving the SQP subproblem either exactly or inexactly, and then a stepsize is selected either by a random projection or by stochastic line search. However, for deterministic problems it is known that computing the search direction and stepsize jointly, such as in trust-region methods, can result in a better performing procedure [13, Chapter 4]. Second, to make the SQP subproblems solvable, the above literature requires the approximation of the Lagrangian Hessian to be positive definite in the null space of constraint Jacobians. Such a condition is common in the SQP literature, but is often achieved by Hessian modifications and excludes promising choices of the Hessian matrix, such as the unperturbed Hessian at the current iterate. Third, to show the global convergence, the above literature requires the merit parameter to be not only stabilized but at a sufficiently large (or small, depending on the context) value with an unknown threshold. To achieve this goal, [11, 12] imposed a condition on the feasibility error when selecting the merit parameter, while [1–3, 9] imposed a condition on the noise distribution. In principle, both resolutions are not necessary because standard deterministic SQP schemes only require a stabilized merit parameter.

Motivated by the above limitations, we design a fully stochastic trust-region SQP method (TR-StoSQP). As a trust-region method, TR-StoSQP computes the search direction and stepsize jointly and has the ability to explore negative curvatures of the Hessian matrix. The fully stochastic trust-region method for solving unconstrained problems was originally proposed in [8], where the authors used a linear model of the objective. [7] generalized that method and used a quadratic model of the objective. Our TR-StoSQP method is based on [7], however, the design is significantly different due to the existence of constraints. In particular, we need to overcome the challenge that the SQP subproblem with trust-region constraints may not have a feasibility point. To address this issue, we design a novel *adaptive relaxation technique*, where we decompose the trial step into a normal step and a tangential step, and decompose the radius of the trust-region into two segments. The two segments, proportional to the ratios of the feasibility and optimality residuals to the full KKT residual, are used to control the size of the normal and tangential steps, respectively. To our knowledge, this is the first trust-region SQP algorithm for constrained problems in a fully stochastic setup. With a stabilized merit parameter (not necessarily large or small enough), we establish the global convergence of TR-StoSQP by showing that the KKT residual converges to zero *almost surely*. Numerical experiments on a subset of problems in the CUTEst test set demonstrate the superior performance of the proposed method.

Notation We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the operator norm for matrices. The identity matrix is denoted as I and $\mathbf{0}$ denotes the zero matrix or vector. Let $G(\boldsymbol{x}) = \nabla^T c(\boldsymbol{x}) \in \mathbb{R}^{m\times d}$ be the Jacobian matrix of the constraints and $P(\boldsymbol{x}) = I - G^T(\boldsymbol{x})[G(\boldsymbol{x})G^T(\boldsymbol{x})]^{-1}G(\boldsymbol{x})$ be the projection matrix onto the null space of $G(\boldsymbol{x})$. We use $\bar{g}(\boldsymbol{x}) = \nabla F(\boldsymbol{x};\xi)$ to denote an estimate of $\nabla f(\boldsymbol{x})$, and use $\bar{(\cdot)}$ to denote stochastic estimates. We define the Lagrangian of (1) as $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T c(\boldsymbol{x})$, and $\nabla\mathcal{L}(\boldsymbol{x})$ is its gradient. At the k-th iteration, we let $\bar{g}_k = \bar{g}(\boldsymbol{x}_k)$, $G_k = G(\boldsymbol{x}_k)$, $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k = \bar{g}_k + G_k^T\boldsymbol{\lambda}_k$ (a similar notation is used for $c_k, P_k, \nabla_{\boldsymbol{x}}\mathcal{L}_k, \nabla\mathcal{L}_k, \bar{\nabla}\mathcal{L}_k$, etc.).

2. Adaptive Relaxation Technique

In this section, we introduce a novel *adaptive relaxation technique* to address the infeasibility issue of trust-region methods for constrained optimization. We refer to [5, 14, 16] for some existing relaxation methods. We decompose a trial step into two orthogonal segments 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. To satisfy the linearized constraints $c_k + G_k\Delta\boldsymbol{x}_k = \mathbf{0}$, when G_k has full row rank, we have $\boldsymbol{w}_k = -G_k^T[G_kG_k^T]^{-1}c_k =: \boldsymbol{v}_k$. However, the existence of trust-region constraint $\|\Delta\boldsymbol{x}_k\| \le \Delta_k$ may prevent us from taking the entire length of $\boldsymbol{v}_k$. Thus, we introduce a scalar $\bar{\gamma}_k \in (0,1]$ and let the normal step be $\boldsymbol{w}_k = \bar{\gamma}_k\boldsymbol{v}_k$. On the other hand, the tangential step can always be written as $\boldsymbol{t}_k = P_k\boldsymbol{u}_k$ for some vector $\boldsymbol{u}_k \in \mathbb{R}^d$. To correctly choose γ_k and adjust $\|\boldsymbol{u}_k\|$ so that $\|\Delta\boldsymbol{x}_k\| \le \Delta_k$, we propose to *adaptively* decompose the trust-region radius into two segments as

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

which are trust-region radii of the normal and tangential step, respectively. Thus, we define $\bar{\gamma}_k$ as

$$\bar{\gamma}_k = \min\{\breve{\Delta}_k/\|\boldsymbol{v}_k\|, 1\}, \tag{2}$$

and compute $\boldsymbol{u}_k$ by solving the trust-region problem

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

When $\|\boldsymbol{v}_k\| = 0$, $\Delta\boldsymbol{x}_k = P_k\boldsymbol{u}_k$ and thus (2) is skipped. We mention that (3) is a standard trust-region problem for unconstrained optimization, and its Cauchy point is sufficient for our analysis.

3. A Fully Stochastic Trust-Region SQP Algorithm

Given the iterate $\boldsymbol{x}_k$, user-specified parameters $\beta_k \in (0,\beta_{\max}]$ with $\beta_{\max} > 0$, $\zeta > 0$, and the merit parameter $\bar{\mu}_{k-1}$ selected at the $(k-1)$-th iteration, the TR-StoSQP algorithm proceeds in the following three steps.

Step 1: Compute parameters. We obtain B_k to approximate the Lagrangian Hessian $\nabla^2_{\boldsymbol{x}}\mathcal{L}_k$ that is deterministic conditional on $\boldsymbol{x}_k$. Then we compute the parameters: $\eta_{1,k} = \zeta\min\{1/\|B_k\|, 6\beta_{\max}/\|G_k\|\}$, $\tau_k = L_{\nabla f,k} + L_{G,k}\bar{\mu}_{k-1} + \|B_k\|$, $\alpha_k = \beta_k/(4\eta_{1,k}\tau_k\beta_{\max} + 6\zeta\beta_{\max})$, and $\eta_{2,k} = \eta_{1,k} - \frac{1}{2}\zeta\eta_{1,k}\alpha_k$, where $L_{\nabla f,k}, L_{G,k} > 0$ are (estimated) Lipschitz constants of $\nabla f(\boldsymbol{x})$ and $G(\boldsymbol{x})$ at $\boldsymbol{x}_k$.

Step 2: Compute trust-region radius. We generate a realization ξ_g^k and compute $\bar{g}_k = \nabla F(\boldsymbol{x}_k;\xi_g^k)$. We further compute the Lagrangian multiplier $\bar{\boldsymbol{\lambda}}_k = -[G_kG_k^T]^{-1}G_k\bar{g}_k$, $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k$ and $\bar{\nabla}\mathcal{L}_k$. Given the parameters computed in Step 1, the trust-region radius Δ_k is generated as:

$$\Delta_k = \begin{cases} \eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\| & \text{if } \|\bar{\nabla}\mathcal{L}_k\| \in (0, 1/\eta_{1,k}), \\ \alpha_k & \text{if } \|\bar{\nabla}\mathcal{L}_k\| \in [1/\eta_{1,k}, 1/\eta_{2,k}], \\ \eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\| & \text{if } \|\bar{\nabla}\mathcal{L}_k\| \in (1/\eta_{2,k}, \infty). \end{cases}$$

Step 3: Compute the trial step and update the merit parameter. We compute the trial step $\Delta\boldsymbol{x}_k$ as introduced in Section 2 and update the iterate as $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta\boldsymbol{x}_k$. We also select the merit parameter $\bar{\mu}_k$ large enough such that

$$\text{Pred}_k \leq -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k - \frac{1}{2}\|c_k\|\breve{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \|B_k\|\breve{\Delta}_k\tilde{\Delta}_k, \tag{4}$$

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

In what follows, we present assumptions and state the main global convergence results.

Assumption 1 *Let $\Omega \subseteq \mathbb{R}^d$ be an open convex set containing the iterates $\{\boldsymbol{x}_k\}$. The function $f(\boldsymbol{x})$ is continuously differentiable and is bounded below by $f_{\inf}$ over Ω. The gradient $\nabla f(\boldsymbol{x})$ is Lipschitz continuous over Ω with constant $L_{\nabla f} > 0$, so that the (estimated) Lipschitz constant $L_{\nabla f,k}$ at $\boldsymbol{x}_k$ satisfies $L_{\nabla f,k} \leq L_{\nabla f}$, $\forall k \geq 0$. Similarly, the constraint $c(\boldsymbol{x})$ is continuously differentiable over Ω; its Jacobian $G(\boldsymbol{x})$ is Lipschitz continuous over Ω with constant $L_G > 0$; and $L_{G,k} \leq L_G$, $\forall k \geq 0$. We also assume there exist positive constants $\kappa_B, \kappa_c, \kappa_{\nabla f}, \kappa_{1,G}, \kappa_{2,G} > 0$ such that*

$$\|B_k\| \leq \kappa_B, \quad \|c_k\| \leq \kappa_c, \quad \|\nabla f_k\| \leq \kappa_{\nabla f}, \quad \kappa_{1,G} \cdot I \preceq G_k G_k^T \preceq \kappa_{2,G} \cdot I, \quad \forall k \geq 0.$$

Assumption 2 (Growth condition) *For $k \in \mathbb{N}$, we have $\mathbb{E}[\bar{g}_k \mid \boldsymbol{x}_k] = \nabla f_k$ and $\mathbb{E}[\|\nabla f_k - \bar{g}_k\|^2 \mid \boldsymbol{x}_k] \leq M_g + M_{g,1}(f_k - f_{\inf})$ for constants $M_g \geq 1, M_{g,1} \geq 0$.*

Assumption 3 (Stabilization of 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 [2]. Assumption 2 is weaker than the commonly used bounded variance condition. Assumption 3 assumes that the merit parameter $\bar{\mu}_k$ stabilizes for large k, which can be provably satisfied under a boundedness condition on $\bar{g}_k$. See detailed proofs in Appendix B. We also only provide a convergence guarantee for the case where β_k is a decaying sequence. Proof can be found in Appendix A.2.

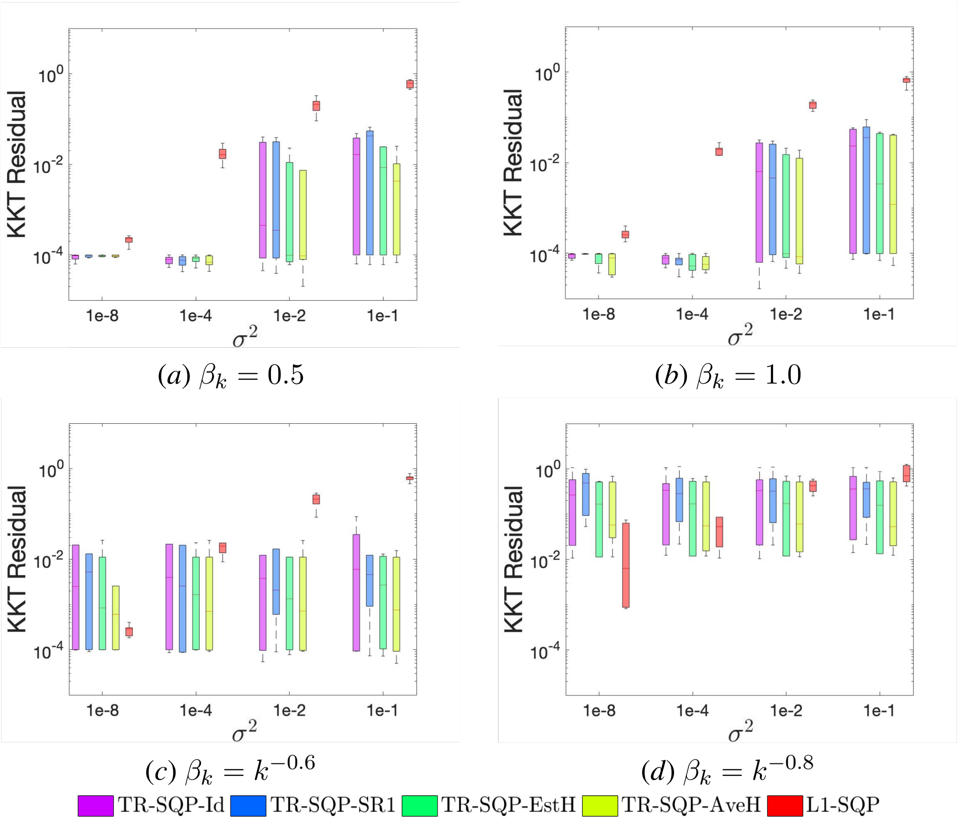
Theorem 4 (Global convergence with constant β_k) *Suppose Assumptions 1–3 hold and $\beta_k = \beta \in (0, \beta_{\max}]$ for $\forall k \geq 0$, then*

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{1}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k} \sum_{k=\bar{K}+1}^{\bar{K}+K} w_k\|\nabla\mathcal{L}_k\|^2\right] \leq \Upsilon\{M_{g,1}\mathbb{E}[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+1} - f_{\inf}] + M_g\}\beta,$$

where w_k, Υ are derived from the analysis.

Theorem 5 (Global convergence with decaying β_k) *Suppose Assumptions 1–3 hold and $\beta_k \in (0, \beta_{\max}]$ satisfies $\sum_{k=0}^{\infty}\beta_k = \infty$ and $\sum_{k=0}^{\infty}\beta_k^2 < \infty$, then $\lim_{k\to\infty}\|\nabla\mathcal{L}_k\| = 0$ almost surely.*

Theorems 4 and 5 establish the global convergence properties for TR-StoSQP. Compared with the conclusion for unconstrained problems in [7], the radius of neighborhood in Theorem 4 is proportional to β. The conclusion of Theorem 5 matches the result in [7], but under weaker assumption that the variance of gradient estimates satisfies the growth condition.



(*a*) $\beta_k = 0.5$

(*b*) $\beta_k = 1.0$

(*c*) $\beta_k = k^{-0.6}$

(*d*) $\beta_k = k^{-0.8}$

Figure 1: KKT residual boxplots of CUTEst datasets.

4. Numerical Experiments

We demonstrate the empirical performance of TR-StoSQP on a subset of equality constrained problems in CUTEst test set [10]. We compare TR-StoSQP with the ℓ_1-StoSQP algorithm [2, Algorithm 3]. We try two constants $\beta_k \in \{0.5, 1\}$ and two decaying sequences $\beta_k \in \{k^{-0.6}, k^{-0.8}\}$. For each β_k, we draw $\bar{g}_k$ from $\mathcal{N}(\nabla f_k, \sigma^2(I + \mathbf{1}\mathbf{1}^T))$, where $\mathbf{1}$ denotes the d-dimensional all one vector. For the noise level σ^2, we try four values in $\{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}\}$. For ℓ_1-StoSQP, we let $B_k = I$ (as used in [2]). For TR-StoSQP, we try four Hessian approximates: the identity matrix (Id), the symmetric rank-one (SR1) update, the estimated Hessian (EstH), and the averaged Hessian (AveH). When EstH and AveH are employed, we draw the (i, j) (also (j, i)) entry of the estimate of $\nabla^2 f_k$ from $\mathcal{N}((\nabla^2 f_k)_{i,j}, \sigma^2)$ with identical σ^2 for estimating the gradient. We draw the boxplots of the KKT residuals of ℓ_1-StoSQP and TR-StoSQP in Figure 1.

From Figure 1, we observe that TR-StoSQP consistently outperforms ℓ_1-StoSQP for constant β_k and decaying β_k with large noise levels. However, for decaying β_k with small noise levels, ℓ_1-StoSQP generally performs better. Moreover, for decaying β_k, TR-StoSQP is more robust to different noise levels. Among the four choices of the Hessian approximates, TR-StoSQP generally performs the best with the averaged Hessian and second best with the estimated Hessian, especially when the noise level is high.

5. Conclusion

We proposed a fully stochastic trust-region SQP (TR-StoSQP) algorithm for solving equality-constrained problems. We developed a novel adaptive relaxation technique to relax the linearized constraints and decompose the trust-region radius adaptively to control the size of normal and tangential steps. We established the global almost sure convergence guarantee and demonstrated the superior performance of TR-StoSQP on a subset of problems in CUTEst collection set.

References

[1] 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.

[2] 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. doi: 10.1137/20m1354556.

[3] 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.

[4] John R. Birge. State-of-the-art-survey—stochastic programming: Computation and applications. *INFORMS Journal on Computing*, 9(2):111–133, may 1997. doi: 10.1287/ijoc.9.2.111.

[5] Richard H. Byrd, Robert B. Schnabel, and Gerald A. Shultz. A trust region algorithm for nonlinearly constrained optimization. *SIAM Journal on Numerical Analysis*, 24(5):1152–1170, oct 1987. doi: 10.1137/0724076.

[6] Changan Chen, Frederick Tung, Naveen Vedula, and Greg Mori. Constraint-aware deep neural network compression. In *Computer Vision – ECCV 2018*, pages 409–424. Springer International Publishing, 2018. doi: 10.1007/978-3-030-01237-3_25.

[7] Frank E. Curtis and Rui Shi. A fully stochastic second-order trust region method. *Optimization Methods and Software*, pages 1–34, nov 2020. doi: 10.1080/10556788.2020.1852403.

[8] Frank E. Curtis, Katya Scheinberg, and Rui Shi. A stochastic trust region algorithm based on careful step normalization. *INFORMS Journal on Optimization*, 1(3):200–220, jul 2019. doi: 10.1287/ijoo.2018.0010.

[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. doi: 10.1007/s10589-014-9687-3.

[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. doi: 10.1007/s10107-022-01846-z.

[13] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer New York, 2006. doi: 10.1007/978-0-387-40065-5.

[14] Emmanuel Omotayo Omojokun. *Trust region algorithms for optimization with nonlinear equality and inequality constraints*. PhD thesis, University of Colorado, Boulder, CO, 1989.

[15] Tyrone Rees, H. Sue Dollar, and Andrew J. Wathen. Optimal solvers for PDE-constrained optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, jan 2010. doi: 10.1137/080727154.

[16] Avi Vardi. A trust region algorithm for equality constrained minimization: Convergence properties and implementation. *SIAM Journal on Numerical Analysis*, 22(3):575–591, jun 1985. doi: 10.1137/0722035.

Appendix A. Proofs

A.1. Fundamental lemmas

Lemma 6 *Suppose $\boldsymbol{u}_k$ is solved (approximately) from (3) and achieved at least Cauchy reduction, then for all $k \in \mathbb{N}$, we have*

$$m(\boldsymbol{u}_k) - m(\mathbf{0}) = \bar{g}_k^T P_k \boldsymbol{u}_k + \frac{1}{2}\boldsymbol{u}_k^T P_k B_k P_k \boldsymbol{u}_k \le -\|P_k\bar{g}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2.$$

Proof Since $\boldsymbol{u}_k$ achieves at least Cauchy reduction in $m(\boldsymbol{u})$, we only need to analyze the reduction achieved by the Cauchy point $\boldsymbol{u}_k^C$, denoted as $m(\boldsymbol{u}_k^C) - m(\mathbf{0})$. From Lemma 4.3 of [13], one finds that the Cauchy point $\boldsymbol{u}_k^C$ lies in the interior of the trust region if $\|P_k\bar{g}_k\|^3 \le \tilde{\Delta}_k\bar{g}_k^T P_k B_k P_k \bar{g}_k$, and lies on the boundary otherwise. If the Cauchy point lies in the interior, we have

$$\boldsymbol{u}_k^C = -\frac{\|P_k\bar{g}_k\|^2}{\bar{g}_k^T P_k B_k P_k \bar{g}_k} P_k\bar{g}_k,$$

noting that $P_k^2 = P_k$, we find

$$m(\boldsymbol{u}_k^C) - m(\mathbf{0}) = \bar{g}_k^T P_k \boldsymbol{u}_k^C + \frac{1}{2}\boldsymbol{u}_k^{CT} P_k B_k P_k \boldsymbol{u}_k^C = -\frac{1}{2}\frac{\|P_k\bar{g}_k\|^4}{\bar{g}_k^T P_k B_k P_k \bar{g}_k} \le -\frac{1}{2}\frac{\|P_k\bar{g}_k\|^2}{\|B_k\|}.$$

And if the Cauchy point lies on the boundary, it is given by

$$\boldsymbol{u}_k^C = -\frac{\tilde{\Delta}_k}{\|P_k\bar{g}_k\|} P_k\bar{g}_k.$$

We similarly have

$$\begin{aligned} m(\boldsymbol{u}_k^C) - m(\mathbf{0}) =& \bar{g}_k^T P_k \boldsymbol{u}_k^C + \frac{1}{2}\boldsymbol{u}_k^{CT} P_k B_k P_k \boldsymbol{u}_k^C \\ =& -\|P_k\bar{g}_k\|\tilde{\Delta}_k + \frac{1}{2}\frac{\bar{g}_k^T P_k B_k P_k \bar{g}_k}{\|P_k\bar{g}_k\|^2}\tilde{\Delta}_k^2 \\ \le& -\|P_k\bar{g}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2. \end{aligned}$$

Combining the above two cases, one finds

$$m(\boldsymbol{u}_k^C) - m(\mathbf{0}) = \bar{g}_k^T P_k \boldsymbol{u}_k^C + \frac{1}{2}\boldsymbol{u}_k^{CT} P_k B_k P_k \boldsymbol{u}_k^C \le -\min\left\{\|P_k\bar{g}_k\|\tilde{\Delta}_k - \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2, \frac{1}{2}\frac{\|P_k\bar{g}_k\|^2}{\|B_k\|}\right\}.$$

Noting that $\|P_k\bar{g}_k\|\tilde{\Delta}_k - \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 \le \frac{1}{2}\frac{\|P_k\bar{g}_k\|^2}{\|B_k\|}$ always holds and $m(\boldsymbol{u}_k) - m(\mathbf{0}) \le m(\boldsymbol{u}_k^C) - m(\mathbf{0})$, we complete the proof. ■

Lemma 7 *Suppose Assumptions 1 and 3 hold, then for all $k \ge \bar{K} + 1$, we have*

$$\begin{aligned} \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \le& -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k - \frac{1}{2}\|c_k\|\breve{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \|B_k\|\breve{\Delta}_k\tilde{\Delta}_k + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \\ &+ \|P_k(\nabla f_k - \bar{g}_k)\|\tilde{\Delta}_k + \frac{1}{2}\tau_k\Delta_k^2. \end{aligned} \qquad (5)$$

Proof From the definitions of $\mathcal{L}_{\bar{\mu}_{\bar{K}}}(\boldsymbol{x})$ and Pred_k, we have

$$\begin{aligned}\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \text{Pred}_k =& f_{k+1} - f_k - \bar{g}_k^T \Delta\boldsymbol{x}_k - \frac{1}{2}\Delta\boldsymbol{x}_k^T B_k \Delta\boldsymbol{x}_k + \bar{\mu}_{\bar{K}}(\|c_{k+1}\| - \|c_k + G_k\Delta\boldsymbol{x}_k\|) \\ \leq& (\nabla f_k - \bar{g}_k)^T\Delta\boldsymbol{x}_k + \frac{1}{2}(L_{\nabla f,k} + \|B_k\| + L_{G,k}\bar{\mu}_{\bar{K}})\|\Delta\boldsymbol{x}_k\|^2,\end{aligned}$$

in the inequality, we use Taylor expansion and Lipschitz continuity of $\nabla f(\boldsymbol{x})$ and $G(\boldsymbol{x})$. Since the merit parameter stabilizes, we have $L_{\nabla f,k} + \|B_k\| + L_{G,k}\bar{\mu}_{\bar{K}} = \tau_k$. By $\Delta\boldsymbol{x}_k = \gamma_k\boldsymbol{v}_k + P_k\boldsymbol{u}_k$, we further have

$$\begin{aligned}\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}} - \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \text{Pred}_k \leq& \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + (\nabla f_k - \bar{g}_k)^T P_k\boldsymbol{u}_k + \frac{1}{2}\tau_k\|\Delta\boldsymbol{x}_k\|^2 \\ =& \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + (\nabla f_k - \bar{g}_k)^T P_k P_k\boldsymbol{u}_k + \frac{1}{2}\tau_k\|\Delta\boldsymbol{x}_k\|^2 \\ \leq& \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \|P_k(\nabla f_k - \bar{g}_k)\|\|P_k\boldsymbol{u}_k\| + \frac{1}{2}\tau_k\|\Delta\boldsymbol{x}_k\|^2, \quad (6)\end{aligned}$$

in the equality, we use the fact that $P_k^2 = P_k$ and in the last inequality, we use Cauchy-Schwartz inequality. The result follows from the combination of (4) and (6) and facts that $\|P_k\boldsymbol{u}_k\| \leq \tilde{\Delta}_k$, $\|\Delta\boldsymbol{x}_k\| \leq \Delta_k$. ■

Lemma 8 *Suppose Assumptions 1 and 2 hold, then for all $k \in \mathbb{N}$, we have*

$$\mathbb{E}_k[\gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k] \leq \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|].$$

Proof When $\boldsymbol{v}_k = \boldsymbol{0}$, the inequality holds trivially. Now we consider $\boldsymbol{v}_k \neq \boldsymbol{0}$. It is implied by the computation of $\boldsymbol{v}_k$ that $c_k = -G_k\boldsymbol{v}_k$, which leads to $\|c_k\| = \|G_k\boldsymbol{v}_k\| \leq \|G_k\|\|\boldsymbol{v}_k\|$, equivalently, $\|\boldsymbol{v}_k\| \geq \|c_k\|/\|G_k\|$. From the definitions of $\eta_{1,k}$ and α_k, it is easy to check that $\eta_{1,k}\alpha_k \leq \frac{1}{\|G_k\|}$. Since $\breve{\Delta}_k \leq \eta_{1,k}\alpha_k\|c_k\|$ for all $k \in \mathbb{N}$, we have

$$\breve{\Delta}_k \leq \frac{\|c_k\|}{\|G_k\|} \leq \|\boldsymbol{v}_k\|.$$

Therefore, from (2) we have

$$\gamma_k\|\boldsymbol{v}_k\| = \min\left\{\breve{\Delta}_k, \|\boldsymbol{v}_k\|\right\} = \breve{\Delta}_k, \quad (7)$$

which is equivalent to $\gamma_k = \breve{\Delta}_k/\|\boldsymbol{v}_k\|$. Since $\eta_{2,k}\alpha_k\|c_k\| \leq \breve{\Delta}_k \leq \eta_{1,k}\alpha_k\|c_k\|$ holds for all $k \in \mathbb{N}$, we have

$$\gamma_{k,\min} := \eta_{2,k}\alpha_k\frac{\|c_k\|}{\|\boldsymbol{v}_k\|} \leq \gamma_k \leq \eta_{1,k}\alpha_k\frac{\|c_k\|}{\|\boldsymbol{v}_k\|} =: \gamma_{k,\max}, \quad (8)$$

where both $\gamma_{k,\min}$ and $\gamma_{k,\max}$ are deterministic conditioned on $\boldsymbol{x}_k$. Let E_k be the event that $(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \geq 0$, E_k^c be its complement and $\mathbb{P}_k[\cdot]$ denote the probability conditioned on $\boldsymbol{x}_k$. By the law of total expectation, we have

$$\begin{aligned}&\mathbb{E}_k[\gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k] \\ =&\mathbb{E}_k[\gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k|E_k]\mathbb{P}_k[E_k] + \mathbb{E}_k[\gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k|E_k^c]\mathbb{P}_k[E_k^c] \\ \overset{(8)}{\leq}&\gamma_{k,\max}\mathbb{E}_k[(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k|E_k]\mathbb{P}_k[E_k] + \gamma_{k,\min}\mathbb{E}_k[(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k|E_k^c]\mathbb{P}_k[E_k^c].\end{aligned}$$

Since $\bar{g}_k$ is an unbiased estimator of ∇f_k, it follows from the equality above and the law of total expectation that

$$\begin{aligned}&\mathbb{E}_k[\gamma_k(\nabla f_k - \bar{g}_k)^T \boldsymbol{v}_k]\\ \leq&\gamma_{k,\min}\mathbb{E}_k[(\nabla f_k - \bar{g}_k)^T \boldsymbol{v}_k | E_k]\mathbb{P}_k[E_k] + \gamma_{k,\min}\mathbb{E}_k[(\nabla f_k - \bar{g}_k)^T \boldsymbol{v}_k | E_k^c]\mathbb{P}_k[E_k^c]\\ &+ (\gamma_{k,\max} - \gamma_{k,\min})\mathbb{E}_k[(\nabla f_k - \bar{g}_k)^T \boldsymbol{v}_k | E_k]\mathbb{P}_k[E_k]\\ =&(\gamma_{k,\max} - \gamma_{k,\min})\mathbb{E}_k[(\nabla f_k - \bar{g}_k)^T \boldsymbol{v}_k | E_k]\mathbb{P}_k[E_k]. \end{aligned} \tag{9}$$

By Cauchy-Schwarz inequality and the law of total expectation, we have

$$\begin{aligned}\mathbb{E}_k[(\nabla f_k - \bar{g}_k)^T \boldsymbol{v}_k | E_k]\mathbb{P}_k[E_k] \leq&\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|\|\boldsymbol{v}_k\| | E_k]\mathbb{P}_k[E_k]\\ =&\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|\|\boldsymbol{v}_k\|] - \mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|\|\boldsymbol{v}_k\| | E_k^c]\mathbb{P}_k[E_k^c]\\ \leq&\|\boldsymbol{v}_k\|\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|]. \end{aligned} \tag{10}$$

Combining (9) and (10), we have

$$\begin{aligned}\mathbb{E}_k[\gamma_k(\nabla f_k - \bar{g}_k)^T \boldsymbol{v}_k] \leq&(\gamma_{k,\max} - \gamma_{k,\min})\|\boldsymbol{v}_k\|\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|]\\ \overset{(8)}{\leq}& (\eta_{1,k} - \eta_{2,k})\, \alpha_k \|c_k\| \mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|]\\ =&\frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|]. \end{aligned}$$

In the last inequality we use the relation $\eta_{1,k} - \eta_{2,k} = \frac{1}{2}\zeta\eta_{1,k}\alpha_k$ and $\|c_k\| \leq \kappa_c$. We complete the proof. ■

Lemma 9 *Suppose Assumptions 1, 2, and 3 hold, then for all* $k \geq \bar{K} + 1$, *we have*

$$\begin{aligned}\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}] \leq&\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - \frac{1}{4}\eta_{2,k}\alpha_k\|\nabla\mathcal{L}_k\|^2 + \frac{1}{2}\left(2\zeta + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|^2]\\ &+ \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|]. \end{aligned}$$

Proof We divide the proof into three cases according to (3).
Case 1. $\|\bar{\nabla}\mathcal{L}_k\| \in (0, 1/\eta_{1,k})$. Since $\tilde{\Delta}_k = \eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|$, for all $k \geq \bar{K} + 1$, we have

$$\begin{aligned}-\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 =& - \eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2\|B_k\|\\ \leq& - \eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\alpha_k^2\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2\\ =& - \left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2. \end{aligned} \tag{11}$$

By the definition of α_k, it is easy to check that $\alpha_k < \frac{2}{\zeta}$, hence we always have $1 - \frac{1}{2}\alpha_k\zeta > 0$. It follows from $\Delta_k = \eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|, \breve{\Delta}_k = \eta_{1,k}\alpha_k\|c_k\|, \tilde{\Delta}_k = \eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|$ that

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \overset{(5)}{\leq} & -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k - \frac{1}{2}\|c_k\|\breve{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \|B_k\|\breve{\Delta}_k\tilde{\Delta}_k + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \\ & + \|P_k(\nabla f_k - \bar{g}_k)\|\tilde{\Delta}_k + \frac{1}{2}\tau_k\Delta_k^2 \\ \overset{(11)}{\leq} & -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{1,k}\alpha_k\|c_k\|^2 + \eta_{1,k}^2\alpha_k^2\|B_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\|c_k\| \\ & + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \frac{1}{2}\tau_k\eta_{1,k}^2\alpha_k^2\|\bar{\nabla}\mathcal{L}_k\|^2.\end{aligned}$$

Noting that $ab \leq a^2/2 + b^2/2$ and $\|\bar{\nabla}\mathcal{L}_k\|^2 = \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2$, we have

$$\begin{aligned} & \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \\ \leq & -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{1,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\|B_k\|\left(\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2\right) \\ & + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \frac{1}{2}\eta_{1,k}\alpha_k\left(\|P_k(\nabla f_k - \bar{g}_k)\|^2 + \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2\right) \\ & + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\tau_k\left(\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2\right) \\ = & -\frac{1}{2}\left(1 - \alpha_k\zeta\right)\eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{1,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\|B_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 \\ & + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\|B_k\|\|c_k\|^2 + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \frac{1}{2}\eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2 \\ & + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\tau_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\tau_k\|c_k\|^2. \end{aligned} \tag{12}$$

Since $\eta_{1,k}\alpha_k \leq \frac{1}{4\tau_k} \leq \frac{1}{4\|B_k\|}$, we have $\frac{1}{2}\eta_{1,k}^2\alpha_k^2\|B_k\|\|c_k\|^2 + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\tau_k\|c_k\|^2 \leq \frac{1}{4}\eta_{1,k}\alpha_k\|c_k\|^2$. Rearranging terms of (12), we have

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq & -\frac{1}{2}\left(1 - \alpha_k\zeta - \eta_{1,k}\alpha_k\tau_k - \eta_{1,k}\alpha_k\|B_k\|\right)\eta_{1,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{1,k}\alpha_k\|c_k\|^2 \\ & + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \frac{1}{2}\eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2. \end{aligned} \tag{13}$$

Taking expectation conditional on $\boldsymbol{x}_k$ on both sides of (13), using the conclusion of Lemma 8 and the relation $\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] = \mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2$ twice, we have

$$\begin{aligned} & \mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}] - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \\ \leq & -\frac{1}{2}\left(1 - \alpha_k\zeta - \eta_{1,k}\alpha_k\tau_k - \eta_{1,k}\alpha_k\|B_k\|\right)\eta_{1,k}\alpha_k\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \frac{1}{4}\eta_{1,k}\alpha_k\|c_k\|^2 \\ & + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] + \frac{1}{2}\eta_{1,k}\alpha_k\left(\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2\right) \\ = & -\frac{1}{2}\eta_{1,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\left(\zeta + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] \\ & - \frac{1}{4}\eta_{1,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \end{aligned}$$

$$
\begin{aligned}
=& -\frac{1}{2}\eta_{1,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{1,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
& + \frac{1}{2}\left(\zeta + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\left(\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] + \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2\right) \\
=& -\frac{1}{2}\left(1 - \alpha_k\zeta - \eta_{1,k}\alpha_k(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
& - \frac{1}{4}\eta_{1,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\left(\zeta + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2].
\end{aligned}
$$

Since $\beta_k \leq 1$ holds for all $k \in \mathbb{N}$, we have $\alpha_k = \frac{\beta_k}{4\eta_{1,k}\tau_k + 6\zeta} \leq \frac{1}{2\eta_{1,k}\tau_k + 4\zeta}$. Meanwhile, $\eta_{1,k}\|B_k\| \leq \zeta$ leads to

$$
\alpha_k \leq \frac{1}{2(\zeta + \eta_{1,k}(\tau_k + \|B_k\|))}. \tag{14}
$$

Rearranging the terms, we have $\alpha_k\zeta + \eta_{1,k}\alpha_k(\tau_k + \|B_k\|) \leq 1/2$, hence

$$
\begin{aligned}
\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}] - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq& -\frac{1}{4}\eta_{1,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{1,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
& + \frac{1}{2}\left(\zeta + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\
=& -\frac{1}{4}\eta_{1,k}\alpha_k\|\nabla\mathcal{L}_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
& + \frac{1}{2}\left(\zeta + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2],
\end{aligned}
$$

in the last equality, we use $\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2 = \|\nabla\mathcal{L}_k\|^2$.

Case 2. $\|\bar{\nabla}\mathcal{L}_k\| \in [1/\eta_{1,k}, 1/\eta_{2,k}]$. Since $\tilde{\Delta}_k = \alpha_k\frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}$, we have

$$
\begin{aligned}
-\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 =& -\alpha_k\frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2}{\|\bar{\nabla}\mathcal{L}_k\|} + \alpha_k^2\frac{\|B_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2}{2\|\bar{\nabla}\mathcal{L}_k\|^2} \\
\leq& -\alpha_k\frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2}{\|\bar{\nabla}\mathcal{L}_k\|} + \alpha_k^2\frac{\zeta\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2}{2\eta_{1,k}\|\bar{\nabla}\mathcal{L}_k\|^2} \\
\leq& -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\alpha_k\frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2}{\|\bar{\nabla}\mathcal{L}_k\|} \\
\leq& -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2, \qquad (15)
\end{aligned}
$$

in the last inequality, we also use the fact that $1 - \frac{1}{2}\alpha_k\zeta > 0$. Meanwhile, it follows from $\breve{\Delta}_k = \alpha_k\frac{\|c_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}$ and $\eta_{2,k}\|\bar{\nabla}\mathcal{L}_k\| \leq 1$ that

$$
-\frac{1}{2}\|c_k\|\breve{\Delta}_k = -\frac{\alpha_k}{2}\frac{\|c_k\|^2}{\|\bar{\nabla}\mathcal{L}_k\|} \leq -\frac{1}{2}\eta_{2,k}\alpha_k\|c_k\|^2. \tag{16}
$$

Since $\Delta_k = \alpha_k, \breve{\Delta}_k = \alpha_k \frac{\|c_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}$ and $\tilde{\Delta}_k = \alpha_k \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}$ we have

$$\begin{aligned}
\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} &\overset{(5)}{\leq} -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k - \frac{1}{2}\|c_k\|\breve{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \|B_k\|\breve{\Delta}_k\tilde{\Delta}_k + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \\
&\quad + \|P_k(\nabla f_k - \bar{g}_k)\|\tilde{\Delta}_k + \frac{1}{2}\tau_k\Delta_k^2 \\
&\overset{(15),(16)}{\leq} -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{2,k}\alpha_k\|c_k\|^2 + \alpha_k^2\frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\|B_k\|\|c_k\|}{\|\bar{\nabla}\mathcal{L}_k\|^2} \\
&\quad + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \alpha_k\frac{\|P_k(\nabla f_k - \bar{g}_k)\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|\bar{\nabla}\mathcal{L}_k\|} + \frac{1}{2}\tau_k\alpha_k^2 \\
&\leq -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{2,k}\alpha_k\|c_k\|^2 + \eta_{1,k}^2\alpha_k^2\|B_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\|c_k\| \\
&\quad + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \frac{1}{2}\tau_k\eta_{1,k}^2\alpha_k^2\|\bar{\nabla}\mathcal{L}_k\|^2,
\end{aligned}$$

in the last inequality, we use $\eta_{1,k}\|\bar{\nabla}\mathcal{L}_k\| \geq 1$. Using similar derivation as (12), we have

$$\begin{aligned}
&\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \\
&\leq -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{2,k}\alpha_k\|c_k\|^2 \\
&\quad + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\|B_k\|\left(\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2\right) + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \\
&\quad + \frac{1}{2}\eta_{1,k}\alpha_k\left(\|P_k(\nabla f_k - \bar{g}_k)\|^2 + \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2\right) + \frac{1}{2}\tau_k\eta_{1,k}^2\alpha_k^2\left(\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2\right).
\end{aligned} \tag{17}$$

Since $\beta_k \leq 1$ holds for all $k \in \mathbb{N}$, we have $\alpha_k = \frac{\beta_k}{4\eta_{1,k}\tau_k + 6\zeta} = \frac{2\beta_k}{8\eta_{1,k}\tau_k + 12\zeta} \leq \frac{2}{8\eta_{1,k}\tau_k + \zeta}$. Rearranging the terms, it follows that

$$\eta_{1,k}\alpha_k\tau_k \leq \frac{1}{4} - \frac{1}{8}\zeta\alpha_k \quad \Rightarrow \quad \eta_{1,k}^2\alpha_k\tau_k \leq \frac{1}{4}\eta_{1,k} - \frac{1}{8}\eta_{1,k}\zeta\alpha_k = \frac{1}{4}\eta_{2,k},$$

which leads to $\eta_{1,k}^2\alpha_k^2\tau_k \leq \frac{1}{4}\eta_{2,k}\alpha_k$. Since $\tau_k \geq \|B_k\|$, we also have $\eta_{1,k}^2\alpha_k^2\|B_k\| \leq \frac{1}{4}\eta_{2,k}\alpha_k$, hence

$$\frac{1}{2}\eta_{1,k}^2\alpha_k^2\|B_k\|\|c_k\|^2 + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\tau_k\|c_k\|^2 \leq \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2. \tag{18}$$

Rearranging terms in (17), we have

$$\begin{aligned}
\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} &\overset{(18)}{\leq} -\left(\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \frac{1}{2}\alpha_k\eta_{1,k}^2(\tau_k + \|B_k\|)\right)\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 \\
&\quad - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \frac{1}{2}\eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2.
\end{aligned}$$

Taking expectation conditional on $\mathcal{F}_{k-1}$ on both sides, using the conclusion of Lemma 8 and the relation $\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] = \mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2$ twice, we have

$$
\begin{aligned}
&\mathbb{E}_k[\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}}] - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} \\
\le& -\left(\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \frac{1}{2}\alpha_k\eta^2_{1,k}(\tau_k + \|B_k\|)\right)\alpha_k\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 \\
&+ \frac{1}{2}\eta_{1,k}\alpha_k(\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2) + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
=& -\frac{1}{2}\eta_{1,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \left(\eta_{1,k} - \eta_{2,k} + \frac{1}{2}\left(\eta_{2,k}\zeta + \eta^2_{1,k}(\tau_k + \|B_k\|)\right)\alpha_k\right)\alpha_k\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] \\
&- \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
=& \left(\eta_{1,k} - \eta_{2,k} + \frac{1}{2}\left(\eta_{2,k}\zeta + \eta^2_{1,k}(\tau_k + \|B_k\|)\right)\alpha_k\right)\alpha_k(\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] + \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2) \\
&- \frac{1}{2}\eta_{1,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
=& -\left(\frac{1}{2}\eta_{1,k} - (\eta_{1,k} - \eta_{2,k}) - \frac{1}{2}\left(\eta_{2,k}\zeta + \eta^2_{1,k}(\tau_k + \|B_k\|)\right)\alpha_k\right)\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 \\
&+ \left(\eta_{1,k} - \eta_{2,k} + \frac{1}{2}\left(\eta_{2,k}\zeta + \eta^2_{1,k}(\tau_k + \|B_k\|)\right)\alpha_k\right)\alpha_k\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\
&- \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|].
\end{aligned}
$$

Substituting $\eta_{2,k}$ with $\eta_{1,k} - \frac{1}{2}\zeta\eta_{1,k}\alpha_k$, we then have

$$
\begin{aligned}
\mathbb{E}_k[\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}}] - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} \le& -\left(\frac{1}{2} - \alpha_k\zeta + \frac{1}{4}\zeta^2\alpha_k^2 - \frac{1}{2}\eta_{1,k}(\tau_k + \|B_k\|)\alpha_k\right)\eta_{1,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 \\
&+ \frac{1}{2}\left(2\zeta - \frac{1}{2}\zeta^2\alpha_k + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\
&- \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|].
\end{aligned}
$$

Since $\beta_k \le 1$ and $\|B_k\| \le \tau_k$, we have $\alpha_k = \frac{\beta_k}{4\eta_{1,k}\tau_k + 6\zeta} = \frac{2\beta_k}{8\eta_{1,k}\tau_k + 12\zeta} \le \frac{2}{4\eta_{1,k}(\tau_k + \|B_k\|) + 7\zeta}$. Rearranging the terms, we have

$$
\begin{aligned}
&\frac{1}{4} - \frac{1}{2}\eta_{1,k}\alpha_k(\tau_k + \|B_k\|) \ge \frac{7}{8}\alpha_k\zeta \\
\Rightarrow\quad &\frac{1}{4}\eta_{1,k} - \eta_{1,k}\alpha_k\zeta - \frac{1}{2}\eta^2_{1,k}\alpha_k(\tau_k + \|B_k\|) \ge -\frac{1}{8}\eta_{1,k}\alpha_k\zeta \\
\Rightarrow\quad &\frac{1}{4}\eta_{1,k} - \eta_{1,k}\alpha_k\zeta - \frac{1}{2}\eta^2_{1,k}\alpha_k(\tau_k + \|B_k\|) \ge -\frac{1}{4}(\eta_{1,k} - \eta_{2,k}).
\end{aligned}
$$

Rearranging the terms, we have $\left(\frac{1}{2} - \alpha_k\zeta + \frac{1}{4}\zeta^2\alpha_k^2 - \frac{1}{2}\eta_{1,k}(\tau_k + \|B_k\|)\alpha_k\right)\eta_{1,k} \geq \frac{1}{4}\eta_{2,k}$. Hence,

$$\begin{aligned}
&\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}] - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \\
\leq& -\frac{1}{4}\eta_{2,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
&+ \frac{1}{2}\left(2\zeta - \frac{1}{2}\zeta^2\alpha_k + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\
\leq& -\frac{1}{4}\eta_{2,k}\alpha_k\|\nabla\mathcal{L}_k\|^2 + \frac{1}{2}\left(2\zeta + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\
&+ \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|],
\end{aligned}$$

in the last equality, we use $\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2 = \|\nabla\mathcal{L}_k\|^2$.
Case 3. $\|\bar{\nabla}\mathcal{L}_k\| \in (1/\eta_{2,k}, \infty)$. It follows from $\tilde{\Delta}_k = \eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|$ that for all $k \geq \bar{K}+1$,

$$\begin{aligned}
-\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 =& -\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\eta_{2,k}^2\alpha_k^2\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2\|B_k\| \\
\leq& -\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\zeta\eta_{2,k}\alpha_k^2\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 \\
=& -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2. \qquad (19)
\end{aligned}$$

Combining $\Delta_k = \eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|, \breve{\Delta}_k = \eta_{2,k}\alpha_k\|c_k\|, \tilde{\Delta}_k = \eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|$ with (19), we have

$$\begin{aligned}
&\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \\
\overset{(5)}{\leq}& -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k - \frac{1}{2}\|c_k\|\breve{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \|B_k\|\breve{\Delta}_k\tilde{\Delta}_k + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \\
&+ \|P_k(\nabla f_k - \bar{g}_k)\|\tilde{\Delta}_k + \frac{1}{2}\tau_k\Delta_k^2 \\
\leq& -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{2,k}\alpha_k\|c_k\|^2 + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \\
&+ \eta_{2,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \eta_{2,k}^2\alpha_k^2\|B_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\|c_k\| + \frac{1}{2}\eta_{2,k}^2\alpha_k^2\tau_k\|\bar{\nabla}\mathcal{L}_k\|^2.
\end{aligned}$$

Using similar derivation as (12), we have

$$\begin{aligned}
\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq& -\left(1 - \frac{1}{2}\alpha_k\zeta\right)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{2,k}\alpha_k\|c_k\|^2 + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \\
&+ \frac{1}{2}\eta_{2,k}\alpha_k\left(\|P_k(\nabla f_k - \bar{g}_k)\|^2 + \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2\right) + \frac{1}{2}\eta_{2,k}^2\alpha_k^2\|B_k\|\left(\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2\right) \\
&+ \frac{1}{2}\eta_{2,k}^2\alpha_k^2\tau_k\left(\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2\right) \\
=& -\frac{1}{2}(1 - \alpha_k\zeta)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{2}\eta_{2,k}\alpha_k\|c_k\|^2 + \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k \\
&+ \frac{1}{2}\eta_{2,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2 + \frac{1}{2}\eta_{2,k}^2\alpha_k^2\|B_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 \\
&+ \frac{1}{2}\eta_{2,k}^2\alpha_k^2\|B_k\|\|c_k\|^2 + \frac{1}{2}\eta_{2,k}^2\alpha_k^2\tau_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\tau_k\eta_{2,k}^2\alpha_k^{\|}c_k\|^2 \qquad (20)
\end{aligned}$$

Since $\eta_{2,k}\alpha_k \leq \eta_{1,k}\alpha_k \leq \frac{1}{4\tau_k} \leq \frac{1}{4\|B_k\|}$, one finds $\frac{1}{2}\eta_{2,k}^2\alpha_k^2\|B_k\|\|c_k\|^2 + \frac{1}{2}\eta_{2,k}^2\alpha_k^2\tau_k\|c_k\|^2 \leq \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2$. It then follows from (20) that

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq& -\frac{1}{2}\left(1 - \alpha_k\zeta - \eta_{2,k}\alpha_k\tau_k - \eta_{2,k}\alpha_k\|B_k\|\right)\eta_{2,k}\alpha_k\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 \\ &+ \gamma_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \frac{1}{2}\eta_{2,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2. \end{aligned} \tag{21}$$

Taking expectation conditional on $\mathcal{F}_{k-1}$ on both sides of (21), using the conclusion of Lemma 8 and the relation $\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] = \mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2$ twice, we have

$$\begin{aligned}&\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}] - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \\ \leq& -\frac{1}{2}\left(1 - \alpha_k\zeta - \eta_{2,k}\alpha_k\tau_k - \eta_{2,k}\alpha_k\|B_k\|\right)\eta_{2,k}\alpha_k\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 \\ &+ \frac{1}{2}\eta_{2,k}\alpha_k\left(\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2\right) + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\ =& -\frac{1}{2}\eta_{2,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \frac{1}{2}\left(\zeta + \eta_{2,k}(\tau_k + \|B_k\|)\right)\eta_{2,k}\alpha_k^2\mathbb{E}_k[\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|^2] - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 \\ &+ \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\ =& -\frac{1}{2}\eta_{2,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\ &+ \frac{1}{2}\left(\zeta + \eta_{2,k}(\tau_k + \|B_k\|)\right)\eta_{2,k}\alpha_k^2\left(\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] + \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2\right) \\ =& -\frac{1}{2}\left(1 - \alpha_k\zeta - \eta_{2,k}\alpha_k(\tau_k + \|B_k\|)\right)\eta_{2,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 \\ &+ \frac{1}{2}\left(\zeta + \eta_{2,k}(\tau_k + \|B_k\|)\right)\eta_{2,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|].\end{aligned}$$

Using similar derivation as (14), since $\eta_{2,k} \leq \eta_{1,k}$, we have $\zeta\alpha_k + \eta_{2,k}\alpha_k(\tau_k + \|B_k\|) \leq \frac{1}{2}$, hence

$$\begin{aligned}&\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}] - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \\ \leq& -\frac{1}{4}\eta_{2,k}\alpha_k\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 - \frac{1}{4}\eta_{2,k}\alpha_k\|c_k\|^2 + \frac{1}{2}\left(\zeta + \eta_{2,k}(\tau_k + \|B_k\|)\right)\eta_{2,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\ &+ \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\ =& -\frac{1}{4}\eta_{2,k}\alpha_k\|\nabla\mathcal{L}_k\|^2 + \frac{1}{2}\left(\zeta + \eta_{2,k}(\tau_k + \|B_k\|)\right)\eta_{2,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\ &+ \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|],\end{aligned}$$

in the last equality, we use $\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|^2 + \|c_k\|^2 = \|\nabla\mathcal{L}_k\|^2$ again. The conclusion follows from combining three cases and noting that $\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \leq \mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|^2]$. ■

Lemma 10 *Suppose Assumptions 1 and 3 hold, then for all $k \in \mathbb{N}$, we have:*
(a) There exist positive parameters $\eta_{\min}, \eta_{\max} \in \mathbb{R}$ such that $\eta_{\min} \leq \eta_{2,k} \leq \eta_{1,k} \leq \eta_{\max}$.
(b) There exists positive parameter $\tau_{\max} \in \mathbb{R}$ such that $\tau_k \leq \tau_{\max}$.
(c) There exist positive parameters $\alpha_l, \alpha_u \in \mathbb{R}$ such that $\alpha_k \in [\alpha_l\beta_k, \alpha_u\beta_k]$.

Proof (a) To show that $\eta_{\min}$ exists, we find that since $\eta_{1,k}$ and τ_k are both non-negative, $\alpha_k = \frac{\beta_k}{4\eta_{1,k}\tau_k+6\zeta} \le \frac{\beta_k}{6\zeta}$ holds for all $k \in \mathbb{N}$, which leads to $\eta_{2,k} = \eta_{1,k}\left(1 - \frac{1}{2}\zeta\alpha_k\right) \ge \eta_{1,k}\left(1 - \frac{1}{12}\beta_k\right)$. Noticing that $\beta_k \le 1$ and $\eta_{1,k} \ge \zeta \min\left\{1/\kappa_B, 6/\sqrt{\kappa_{2,G}}\right\}$, we have $\eta_{2,k} \ge \frac{11}{12}\zeta \min\left\{\frac{1}{\kappa_B}, \frac{6}{\sqrt{\kappa_{2,G}}}\right\} := \eta_{\min} > 0$. Since $\eta_{2,k} = \eta_{1,k} - \frac{1}{2}\zeta\eta_{1,k}\alpha_k$, it is straightforward that $\eta_{2,k} \le \eta_{1,k}$. The existence of $\eta_{\max}$ follows from $\eta_{1,k} = \zeta \min\left\{\frac{1}{\|B_k\|}, \frac{6}{\|G_k\|}\right\}$ and $\|G_k\| \ge \sqrt{\kappa_{1,G}}$, which leads to $\eta_{1,k} \le 6\zeta/\sqrt{\kappa_{1,G}} := \eta_{\max}$. (b) From Assumption 1, we have that for all $k \in \mathbb{N}$, $L_{\nabla f,k} \le L_{\nabla f}$, $L_{G,k} \le L_G$ and $\|B_k\| \le \kappa_B$. Since $\bar{\mu}_k \le \hat{\mu}$ for all $k \in \mathbb{N}$, we have $\tau_k = L_{\nabla f,k} + L_{G,k}\bar{\mu}_{k-1} + \|B_k\| \le L_{\nabla f} + L_G\hat{\mu} + \kappa_B := \tau_{\max}$. (c) Since $\alpha_k = \beta_k/(4\eta_{1,k}\tau_k + 6\zeta)$, we have $\alpha_l\beta_k := \frac{\beta_k}{4\eta_{\max}\tau_{\max}+6\zeta} \le \alpha_k \le \frac{\beta_k}{6\zeta} =: \alpha_u\beta_k$. ■

A.2. Proof of Theorem 5

Proof Using the conclusion of Lemma 9 and Assumption 2, we have

$$\begin{aligned}\mathbb{E}_k[\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}}] \le& \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \frac{1}{4}\eta_{2,k}\alpha_k\|\nabla\mathcal{L}_k\|^2 + \frac{1}{2}\zeta\eta_{1,k}\kappa_c\alpha_k^2\sqrt{M_g + M_{g,1}(f_k - f_{\inf})} \\ &+ \frac{1}{2}\left(2\zeta + \eta_{1,k}(\tau_k + \|B_k\|)\right)\eta_{1,k}\alpha_k^2[M_g + M_{g,1}(f_k - f_{\inf})].\end{aligned}$$

Since $M_g \ge 1$, we have $\sqrt{M_g + M_{g,1}(f_k - f_{\inf})} \le M_g + M_{g,1}(f_k - f_{\inf})$. Conclusions of Lemma 10 further lead to

$$\mathbb{E}_k\left[\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}}\right] \le \mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \frac{1}{4}\eta_{\min}\alpha_l\beta\|\nabla\mathcal{L}_k\|^2 + \Upsilon_1\beta^2[M_g + M_{g,1}(f_k - f_{\inf})],$$

where $\Upsilon_1 := \frac{1}{2}\left(2\zeta + \eta_{\max}(\tau_{\max} + \kappa_B) + \zeta\kappa_c\right)\eta_{\max}\alpha_u^2$. Due to the relation that $f_k - f_{\inf} \le f_k - f_{\inf} + \|c_k\| = \mathcal{L}^k_{\bar{\mu}_{\bar{K}}}$, we have

$$\mathbb{E}_k\left[\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}}\right] \le \left(1 + \Upsilon_1 M_{g,1}\beta_k^2\right)\mathcal{L}^{k}_{\bar{\mu}_{\bar{K}}} - \frac{1}{4}\eta_{\min}\alpha_l\beta_k\|\nabla\mathcal{L}_k\|^2 + \Upsilon_1 M_g\beta_k^2, \tag{22}$$

where $\Upsilon_1 := \frac{1}{2}\left(2\zeta + \eta_{\max}(\tau_{\max} + \kappa_B) + \zeta\kappa_c\right)\eta_{\max}\alpha_u^2$. Taking total expectation on both sides of (22), since $\sum_{k=\bar{K}+1}^{\infty}\beta_k^2 < \infty$, it follows from Robbins-Siegmund theorem that

$$\sup_{k\ge\bar{K}+1}\mathbb{E}[\mathcal{L}^k_{\bar{\mu}_{\bar{K}}}] < \infty, \tag{23a}$$

$$\mathbb{E}\left[\sum_{k=\bar{K}+1}^{\infty}\beta_k\|\nabla\mathcal{L}_k\|^2\right] < \infty, \tag{23b}$$

thus the first result in (**??**) holds. Since $\sum_{k=0}^{\infty}\beta_k = \infty$ and $\bar{K}$ is finite, one finds $\sum_{k=\bar{K}+1}^{\infty}\beta_k = \infty$. Thus from (23b) we find that if dividing $\mathbb{E}\left[\sum_{k=\bar{K}+1}^{k=\bar{K}+K}\beta_k\|\nabla\mathcal{L}_k\|^2\right]$ by $\sum_{k=\bar{K}+1}^{k=\bar{K}+K}\beta_k$ and letting $K \to \infty$, we have the second result in (**??**). Moreover, by Fubini's theorem, (23b) implies

$$\sum_{k=\bar{K}+1}^{\infty}\beta_k\|\nabla\mathcal{L}_k\|^2 < \infty \tag{24}$$

with probability 1, which combined with $\sum_{k=\bar{K}+1}^{\infty}\beta_k = \infty$ yields $\liminf_{k\to\infty}\|\nabla\mathcal{L}_k\|^2 = 0$ almost surely. Next, we prove $\lim_{k\to\infty}\|\nabla\mathcal{L}_k\| = 0$ almost surely. Suppose that $\lim_{k\to\infty}\|\nabla\mathcal{L}_k\| = 0$ almost surely does not hold, then 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, we already show that with probability 1, there exists an infinite index set $\mathcal{K}_2$ such that $\|\nabla\mathcal{L}_k\| \le \epsilon$ for all $k \in \mathcal{K}_2$. They imply that with some nonzero probability, there are index sets $\{m_i\} \subset \mathbb{N}$ and $\{n_i\} \subset \mathbb{N}$ with $m_i < n_i$ for all $i \in \mathbb{N}$ such that

$$\|\nabla\mathcal{L}_{m_i}\| \ge 2\epsilon, \|\nabla\mathcal{L}_{n_i}\| < \epsilon, \text{and } \|\nabla\mathcal{L}_k\| \ge \epsilon \text{ for all } k \in \{m_i+1, \cdots, n_i-1\}. \tag{25}$$

Thus, one finds

$$\infty \overset{(24)}{>} \sum_{k=\bar{K}+1}^{\infty} \beta_k\|\nabla\mathcal{L}_k\|^2 \ge \sum_{i=0}^{\infty}\sum_{k=m_i}^{n_i-1}\beta_k\|\nabla\mathcal{L}_k\|^2 \overset{(25)}{\ge} \epsilon^2\sum_{i=0}^{\infty}\sum_{k=m_i}^{n_i-1}\beta_k \text{ with probability 1,}$$

which means that

$$\lim_{i\to\infty}\sum_{k=m_i}^{n_i-1}\beta_k = 0 \text{ with probability 1.} \tag{26}$$

Note that $\|\bar{\nabla}\mathcal{L}_k\| \le \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \|c_k\| = \|P_k\bar{g}_k\| + \|c_k\| \le \|\nabla f_k\| + \|\nabla f_k - \bar{g}_k\| + \|c_k\|$, so Assumptions 1 and 2 imply $\mathbb{E}_k[\|\bar{\nabla}\mathcal{L}_k\|] \le \kappa_{\nabla f}+\kappa_c+\sqrt{M_g + M_{g,1}(f_k - f_{\inf})} \le \kappa_{\nabla f}+\kappa_c+M_g+M_{g,1}(f_k-f_{\inf})$. Therefore, from $\mathbb{E}_k\left[\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|\right] = \mathbb{E}_k\left[\|\Delta\boldsymbol{x}_k\|\right] \le \mathbb{E}_k[\Delta_k] \le \eta_{1,k}\alpha_k\mathbb{E}_k[\|\bar{\nabla}\mathcal{L}_k\|]$ we have

$$\mathbb{E}_k\left[\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|\right] \le \eta_{\max}\alpha_u\beta_k\left[\kappa_{\nabla f}+\kappa_c+M_g+M_{g,1}(f_k-f_{\inf})\right]. \tag{27}$$

Taking total expectation on (27), we have

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|\right] \le& \eta_{\max}\alpha_u\beta_k\left[\kappa_{\nabla f}+\kappa_c+M_g+M_{g,1}\mathbb{E}\left[f_k-f_{\inf}\right]\right]\\ \le& \eta_{\max}\alpha_u\beta_k\left[\kappa_{\nabla f}+\kappa_c+M_g+M_{g,1}\sup_{k\ge\bar{K}+1}\mathbb{E}\left[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^k\right]\right].\end{aligned}$$

In the last inequality, we use the relation that $\mathbb{E}[f_k - f_{\inf}] \le \mathbb{E}\left[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^k\right] \le \sup_{k\ge\bar{K}+1}\mathbb{E}\left[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^k\right]$. Taking summation over m_i to $n_i - 1$, we get

$$\mathbb{E}[\|\boldsymbol{x}_{n_i}-\boldsymbol{x}_{m_i}\|] \le \eta_{\max}\alpha_u\left[\kappa_{\nabla f}+\kappa_c+M_g+M_{g,1}\sup_{k\ge\bar{K}+1}\mathbb{E}\left[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^k\right]\right]\sum_{k=m_i}^{n_i-1}\beta_k. \tag{28}$$

Combining (23a), (26) and (28), we have $\lim_{i\to\infty}\|\boldsymbol{x}_{n_i}-\boldsymbol{x}_{m_i}\| = 0$ with probability 1. From Assumption 1, we find that $\nabla\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})$ is Lipschitz continuous in both $\boldsymbol{x}$ and $\boldsymbol{\lambda}$. Since $\boldsymbol{\lambda}_k = -[G_kG_k^T]^{-1}G_k\nabla f_k$, Assumption 1 implies that $\boldsymbol{\lambda}$ is also Lipschitz in $\boldsymbol{x}$, so there exists a constant $L_{\nabla\mathcal{L}} > 0$ such that $\|\nabla\mathcal{L}_{n_i}-\nabla\mathcal{L}_{m_i}\| \le L_{\nabla\mathcal{L}}\|\boldsymbol{x}_{n_i}-\boldsymbol{x}_{m_i}\|$. We thus have $0 \le \lim_{i\to\infty}\|\nabla\mathcal{L}_{n_i}-\nabla\mathcal{L}_{m_i}\| \le \lim_{i\to\infty}L_{\nabla\mathcal{L}}\|\boldsymbol{x}_{n_i}-\boldsymbol{x}_{m_i}\| = 0$ with probability 1. This yields contradiction to (25). ■

Appendix B. Merit parameter behavior

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

Lemma 12 *Suppose Assumptions 1 and 11 hold, then there exist a stochastic $\bar{K} < \infty$ and a deterministic constant $\hat{\mu}$, such that $\bar{\mu}_k = \bar{\mu}_{\bar{K}} \le \hat{\mu}$ for $\forall k > \bar{K}$, meaning that the merit parameter stabilizes after finite iterations.*

Proof To show the merit parameter stabilizes, it suffices to show that if $\bar{\mu}_k$ is larger than a deterministic threshold independent to k, (4) is always satisfied. From the definition of Pred_k, we have

$$\begin{aligned}
\text{Pred}_k =& \bar{g}_k^T \Delta\boldsymbol{x}_k + \frac{1}{2}\Delta\boldsymbol{x}_k^T B_k \Delta\boldsymbol{x}_k + \bar{\mu}_k(\|c_k + G_k\Delta\boldsymbol{x}_k\| - \|c_k\|) \\
=& \bar{g}_k^T P_k\boldsymbol{u}_k + \frac{1}{2}\boldsymbol{u}_k^T P_k B_k P_k \boldsymbol{u}_k + \gamma_k(\bar{g}_k - \nabla f_k)^T\boldsymbol{v}_k + \gamma_k \nabla f_k^T \boldsymbol{v}_k + \gamma_k \boldsymbol{v}_k^T B_k P_k \boldsymbol{u}_k \\
&+ \frac{1}{2}\gamma_k^2 \boldsymbol{v}_k^T B_k \boldsymbol{v}_k - \bar{\mu}_k\gamma_k\|c_k\| \\
\le& \bar{g}_k^T P_k\boldsymbol{u}_k + \frac{1}{2}\boldsymbol{u}_k^T P_k B_k P_k \boldsymbol{u}_k - \frac{1}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\| + \frac{1}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\| + \gamma_k\|\bar{g}_k - \nabla f_k\|\|\boldsymbol{v}_k\| \\
&+ \gamma_k\|\nabla f_k\|\|\boldsymbol{v}_k\| + \|B_k\|\|\gamma_k\boldsymbol{v}_k\|\|P_k\boldsymbol{u}_k\| + \frac{1}{2}\|B_k\|\|\gamma_k\boldsymbol{v}_k\|^2 - \bar{\mu}_k\gamma_k\|c_k\|,
\end{aligned}$$

in the last inequality, we use Cauchy-Schwartz inequality. Using Assumptions 1 and 11, noting that $\|\boldsymbol{v}_k\| \le \|G_k^T[G_kG_k^T]^{-1}\|\|c_k\| \le \|c_k\|/\sqrt{\kappa_{1,G}}$, $\gamma_k \le 1$, $\|\gamma_k\boldsymbol{v}_k\| \le \breve{\Delta}_k$ and $\|P_k\boldsymbol{u}_k\| \le \tilde{\Delta}_k$, we have

$$\begin{aligned}
\text{Pred}_k \le& \bar{g}_k^T P_k\boldsymbol{u}_k + \frac{1}{2}\boldsymbol{u}_k^T P_k B_k P_k \boldsymbol{u}_k - \frac{1}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\| + \gamma_k\frac{\kappa_c}{2\sqrt{\kappa_{1,G}}}\|c_k\| + \gamma_k\frac{M_1}{\sqrt{\kappa_{1,G}}}\|c_k\| \\
&+ \gamma_k\frac{\kappa_{\nabla f}}{\sqrt{\kappa_{1,G}}}\|c_k\| + \|B_k\|\breve{\Delta}_k\tilde{\Delta}_k + \gamma_k\frac{\kappa_c\kappa_B}{2\kappa_{1,G}}\|c_k\| - \bar{\mu}_k\gamma_k\|c_k\| \\
=& \bar{g}_k^T P_k\boldsymbol{u}_k + \frac{1}{2}\boldsymbol{u}_k^T P_k B_k P_k \boldsymbol{u}_k - \frac{1}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\| + \|B_k\|\breve{\Delta}_k\tilde{\Delta}_k \\
&+ \gamma_k\left(\frac{\kappa_c + 2M_1 + 2\kappa_{\nabla f}}{2\sqrt{\kappa_{1,G}}} + \frac{\kappa_c\kappa_B}{2\kappa_{1,G}} - \bar{\mu}_k\right)\|c_k\|.
\end{aligned}$$

Therefore, if

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

one finds

$$\text{Pred}_k \le \bar{g}_k^T P_k\boldsymbol{u}_k + \frac{1}{2}\boldsymbol{u}_k^T P_k B_k P_k \boldsymbol{u}_k - \frac{1}{2}\gamma_k\|\boldsymbol{v}_k\|\|c_k\| + \|B_k\|\breve{\Delta}_k\tilde{\Delta}_k. \tag{29}$$

The conclusion follows by combining the result of Lemma 6, (7), (29), and $P_k\bar{g}_k = \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k$. ■