Adaptive Inexact Sequential Quadratic Programming via Iterative Randomized Sketching

author names withheld

Under Review for OPT 2022

Abstract

We consider solving nonlinear optimization problems with equality constraints. We propose a randomized algorithm based on sequential quadratic programming (SQP) with a differentiable exact augmented Lagrangian as the merit function. In each SQP iteration, we solve the Newton system inexactly via iterative randomized sketching. The accuracy of the inexact solution and the penalty parameter of the augmented Lagrangian are adaptively controlled in the algorithm to ensure that the inexact random search direction is a descent direction of the augmented Lagrangian. This allows us to establish global convergence almost surely. Moreover, we show that a unit stepsize is admissible for the inexact search direction provided the iterate lies in a neighborhood of the solution. Based on this result, we show that the proposed algorithm exploits local linear convergence. We apply the algorithm on benchmark nonlinear problems in CUTEst test set and on constrained logistic regression with datasets from LIBSVM to demonstrate its superior performance.

1. Introduction

We consider the nonlinear equality-constrained optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \qquad \text{s.t. } c(\boldsymbol{x}) = \mathbf{0}, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and $c : \mathbb{R}^n \to \mathbb{R}^m$ are equality constraints. There exist numerous methods for solving Problem (1.1), including projected first- and second-order methods, penalty methods, augmented Lagrangian methods, and sequential quadratic programming (SQP). In this paper, we focus on solving (1.1) via SQP, which is one of the leading second-order methods for constrained optimization problems [9, 10, 13]. The algorithms in this class typically enjoy global convergence guarantees, and require a few iterations to find a local solution. However, the computational cost of SQP algorithms is dominated by solving one (or more) linear system in each iteration, which can be prohibitive for large-scale problems.

To reduce the per-iteration computational cost, [4] proposed an inexact SQP algorithm where, in each iteration, the linear system is approximately solved using an iterative solver and the stepsizes are chosen based on a penalized merit function. With suitable conditions on the quality of the inexact solution, the authors showed that the inexact search direction is still a descent direction of the merit function and the algorithm enjoys global convergence. Despite the solid theoretical underpinnings, the algorithm of [4] suffers from few drawbacks. First, for each SQP iteration, the algorithm relies on a few fixed tuning parameters (κ_1, κ_2, ϵ, β) for bounding the residuals of the iterative solver. These parameters may substantially affect the performance of the algorithm and have to be chosen

© .

carefully. In particular, a tighter residual bound will lead to more inner loop iterations to compute a more precise step. However, the cost of more inner loop iterations must be balanced against a possible decrease in the outer loop iterations for finding the local solution. Second, the algorithm uses a nonsmooth merit function $\phi_\pi(\boldsymbol{x}) = f(\boldsymbol{x}) + \pi\|c(\boldsymbol{x})\|$ when performing the line search, which is known to cause the Maratos effect—a unit stepsize may not be accepted near the solution. Such an effect leads to a slow local convergence [5]. Third, the local behavior of that algorithm has not been rigorously analyzed.

In this paper, we propose a randomized SQP algorithm to solve (1.1) in which the linear system in the inner loop is solved using the iterative randomized sketching (IRS) [8]. We use a differentiable *exact* augmented Lagrangian as the merit function of the form

$$\mathcal{L}_{\boldsymbol{\eta}}(\boldsymbol{x}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) + \frac{\eta_1}{2}\|c(\boldsymbol{x})\|^2 + \frac{\eta_2}{2}\|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})\|^2, \tag{1.2}$$

where $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T c(\boldsymbol{x})$ is the Lagrangian function of Problem (1.1) with $\boldsymbol{\lambda} \in \mathbb{R}^m$ being the Lagrangian multipliers, and $\boldsymbol{\eta} = (\eta_1, \eta_2)$ is the penalty parameter. The benefit of using an *exact* penalty function is that a stationary point $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ of (1.2) is also a stationary point of (1.1) and vice versa, provided that η_1 is sufficiently large and η_2 is sufficiently small [2, Proposition 4.15]. Further, the smoothness of the merit function in (1.2) effectively overcomes the Maratos effect [3]. We emphasize three novelties of the proposed algorithm. First, we use the iterative randomized sketching [8] to compute an inexact solution of the linear system. Projecting a large linear system of equations into a smaller one and obtaining an approximate solution leads to large computational savings [8, 16, 17]. Second, the algorithm adaptively selects a parameter that controls a bound on the residuals when accepting the step. As a result, the inexact solution of the SQP subproblem is a descent direction of the merit function, and is accurate enough to guarantee the global and local linear convergence of the algorithm. Empirically, our adaptive algorithm results in smaller KKT residuals (the sum of the feasibility error and the optimality error) and fewer gradient evaluations. Third, despite the randomness in the inexact search direction brought by the randomized solver, we establish the almost sure global convergence. Furthermore, we show that the algorithm locally selects a unit stepsize even with the adaptive step acceptance condition, which leads to a local linear convergence rate. Such a local result complements the existing literature on inexact SQP algorithms.

2. Method

We propose an adaptive inexact SQP algorithm that uses iterative randomized sketching to solve the Newton system in the inner loop. We suppose f and c are thrice continuously differentiable. We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the operator norm for matrices. At the k-th iteration, we let $f^k = f(\boldsymbol{x}^k)$, etc., to simplify the notation.

When a constraint qualification holds, the first-order necessary conditions for $\boldsymbol{x}^\star$ to be a solution to (1.1) are that there exist multipliers $\boldsymbol{\lambda}^\star$ such that

$$\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star) \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star) \end{pmatrix} = \begin{pmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^\star) + \boldsymbol{G}^T(\boldsymbol{x}^\star)\boldsymbol{\lambda}^\star \\ c(\boldsymbol{x}^\star) \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \tag{2.1}$$

where $\boldsymbol{G}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}}^T c(\boldsymbol{x}) = (\nabla_{\boldsymbol{x}} c_1(\boldsymbol{x}), \ldots, \nabla_{\boldsymbol{x}} c_m(\boldsymbol{x}))^T \in \mathbb{R}^{m\times n}$ is the constraint Jacobian matrix. In each iteration k, the SQP algorithm finds the primal-dual step $(\Delta\boldsymbol{x}^k, \Delta\boldsymbol{\lambda}^k)$ by solving the

following Newton system

$$\begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \Delta \boldsymbol{x}^k \\ \Delta \boldsymbol{\lambda}^k \end{pmatrix} = - \begin{pmatrix} \nabla_{\boldsymbol{x}} f^k + (\boldsymbol{G}^k)^T \boldsymbol{\lambda}^k \\ c^k \end{pmatrix}, \tag{2.2}$$

where $\boldsymbol{B}(\boldsymbol{x}, \boldsymbol{\lambda})$ is the Hessian of the Lagrangian $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{\lambda}) = \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ or its symmetric perturbation. Let $\boldsymbol{A}^k = \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$ and we rewrite the Newton system (2.2) by

$$\boldsymbol{A}^k \begin{pmatrix} \Delta \boldsymbol{x}^k \\ \Delta \boldsymbol{\lambda}^k \end{pmatrix} = -\nabla \mathcal{L}^k. \tag{2.3}$$

We apply the iterative randomized sketching to obtain an inexact solution $(\tilde{\Delta}\boldsymbol{x}^k, \tilde{\Delta}\boldsymbol{\lambda}^k)$ to (2.3). In particular, we let $\mathbf{S}^{k,j} \in \mathbb{R}^{(n+m)\times d}$ be a random matrix drawn from some probability distribution P at j-th inner iteration. At the outer iteration k and inner iteration j, we let $\mathbf{r}^{k,j} = \boldsymbol{A}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^{k,j} \\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j} \end{pmatrix} + \nabla \mathcal{L}^k$ denote the residual of the Newton system. The inner loop iteration updates the solution as

$$\begin{aligned} \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^{k,j+1} \\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j+1} \end{pmatrix} &= \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^{k,j} \\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j} \end{pmatrix} - (\boldsymbol{A}^k)^T \mathbf{S}^{k,j} \left((\mathbf{S}^{k,j})^T \boldsymbol{A}^k (\boldsymbol{A}^k)^T \mathbf{S}^{k,j} \right)^{-1} (\mathbf{S}^{k,j})^T \mathbf{r}^{k,j} \\ &= \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^{k,j} \\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j} \end{pmatrix} - \mathbf{W}^{k,j} \mathbf{r}^{k,j} \end{aligned} \tag{2.4}$$

where $\mathbf{W}^{k,j} = (\boldsymbol{A}^k)^T \mathbf{S}^{k,j} \left((\mathbf{S}^{k,j})^T \boldsymbol{A}^k (\boldsymbol{A}^k)^T \mathbf{S}^{k,j} \right)^{-1} (\mathbf{S}^{k,j})^T$. The following condition describes when a step will be accepted.

Step Acceptance Condition. Given $\eta_1, \eta_2 > 0$ and $0 < \delta < 1$, a step $(\tilde{\Delta}\boldsymbol{x}^k, \tilde{\Delta}\boldsymbol{\lambda}^k)$ that is computed via (2.4) is acceptable if

$$\left\| \mathbf{r}^k \right\| \leq \frac{\delta}{\|\boldsymbol{A}^k\| \, \Psi_k} \left\| \nabla \mathcal{L}^k \right\|, \tag{2.5}$$

where Ψ_k is an upper bound of $\left\| (\boldsymbol{A}^k)^{-1} \right\|$ defined in Lemma 7, **and**

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}^k_{\boldsymbol{\eta}} \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}^k_{\boldsymbol{\eta}} \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \leq -\frac{\eta_2}{2} \left\| \nabla \mathcal{L}^k \right\|^2. \tag{2.6}$$

In each SQP iteration, we first update the inexact search direction by (2.4) until the residual at the inner iteration j satisfies (2.5) with a given δ. Then, we check if the inexact search direction is a sufficient descent direction of (1.2); that is, whether (2.6) is satisfied for $\boldsymbol{\eta} = (\eta_1, \eta_2)$. If it does not satisfy (2.6), we increase η_1, and decrease η_2 and δ as

$$\eta_1 \leftarrow \eta_1 \nu^2, \quad \eta_2 \leftarrow \eta_2 / \nu, \quad \delta \leftarrow \delta / \nu^4 \tag{2.7}$$

where $\nu > 1$ is a given constant. We repeat the above two steps until we find an inexact search direction which satisfies (2.5) and (2.6) with appropriate (η_1, η_2, δ). Upon identifying the inexact search direction, we further require the threshold parameter δ is decreased to satisfy

$$\delta \leq \left(\frac{1}{2} - \beta \right) \frac{\eta_2}{2\Psi_k^2 (3\Upsilon_k + 4\eta_2 \Upsilon_k^2 + \eta_1 \Upsilon_k^2)} = \delta_k^{\text{trial}} \tag{2.8}$$

where $\Upsilon_k = \|\boldsymbol{B}^k\| \vee \|\boldsymbol{G}^k\| \vee \|\boldsymbol{H}^k\|$ and $\beta \in (0, 1/2)$. Finally, we update the inexact search direction by (2.4) once again until (2.5) holds with $\delta \leq \delta_k^{trial}$. This additional procedure is to ensure the algorithm to select a unit stepsize locally, so that it enjoys the local linear convergence near a stationary point of (1.1). In Appendix B, we show in Lemma 12 that the adaptive parameters (η_1, η_2, δ) will eventually be stabilized after some outer iterations. We design this scheme using double while loops in Algorithm 1. The stepsize α_k is selected to satisfy the Armijo condition

$$\mathcal{L}_{\boldsymbol{\eta}}^{k+1} \leq \mathcal{L}_{\boldsymbol{\eta}}^{k} + \alpha_k \beta \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta} \boldsymbol{x}^k \\ \tilde{\Delta} \boldsymbol{\lambda}^k \end{pmatrix}, \tag{2.9}$$

and the iterate is updated as

$$\begin{pmatrix} \boldsymbol{x}^{k+1} \\ \boldsymbol{\lambda}^{k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{k} \\ \boldsymbol{\lambda}^{k} \end{pmatrix} + \alpha_k \begin{pmatrix} \tilde{\Delta} \boldsymbol{x}^k \\ \tilde{\Delta} \boldsymbol{\lambda}^k \end{pmatrix}.$$

Our Algorithm 1 is presented in Appendix A.

3. Convergence Analysis

We now study well-posedness of Algorithm 1, and establish global and local linear convergence guarantees. We emphasize that the randomness plays a key role in the analysis since the inexact search direction is calculated by the iterative randomized solver. Compared to an algorithm that uses a deterministic iterative solver, our inexact search direction is stochastic. As a result, all the components of the algorithm that are affected by the search direction are also random; for example, (2.5), (2.6), and (2.9). Our analysis relies on the following assumption.

Assumption 1 *Iterates $\{\boldsymbol{x}^k\}$ and $\{\boldsymbol{x}^k + \alpha_k \tilde{\Delta} \boldsymbol{x}^k\}$ belong to an open convex set $\mathcal{X}$ where $\alpha_k \in [0, 1]$. The functions $f, c_i \in C^3(\mathcal{X})$ are bounded over $\mathcal{X}$ for $i \in \{1, ..., m\}$. Moreover, the first, second, and third derivatives of f and c_i are bounded over $\mathcal{X}$.*

Assumption 2 *The Jacobian matrices $\{\boldsymbol{G}^k\}_{k \geq 0}$ have full row rank. There exist constants $\xi_{\boldsymbol{B}}, \Upsilon_{\boldsymbol{B}} > 0$, such that, for any iteration $k \geq 0$, $\boldsymbol{z}^T \boldsymbol{B}^k \boldsymbol{z} \geq \xi_{\boldsymbol{B}} \|\boldsymbol{z}\|^2$ for any $\boldsymbol{z} \in \{\boldsymbol{z} : \boldsymbol{G}^k \boldsymbol{z} = 0, \boldsymbol{z} \neq 0\}$ and $\|\boldsymbol{B}^k\| \leq \Upsilon_{\boldsymbol{B}}$.*

Assumption 3 *The random matrices $\{\mathbf{S}^{k,j}\}_{k \geq 0, j \geq 0}$ are independent and identically distributed from a probability distribution P such that, for any inner iteration $j \geq 0$, $p((\mathbf{S}^{k,j})^T \boldsymbol{z} \neq 0) > 0$ for any $\boldsymbol{z} \in \mathbb{R}^{n+m} \backslash \{0\}$.*

Assumption 4 *For any iteration $k \geq 0$, $\|(\boldsymbol{H}^k - \boldsymbol{B}^k)\| = o(1)$.*

Assumption 1 does not make any assumptions about the set Λ that contains the dual iterates $\{\boldsymbol{\lambda}^k\}$ and $\{\boldsymbol{\lambda}^k + \alpha_k \tilde{\Delta} \boldsymbol{\lambda}^k\}$. The boundedness of Λ can be proven based on the algorithm itself; see Lemma 14 in Appendix B. Furthermore, when using the augmented Lagrangian as the merit function (see (1.2)), it is common to assume the existence of third derivatives of f and c_i, since $\nabla_{\boldsymbol{x}}^2 \mathcal{L}_{\boldsymbol{\eta}}$ requires $\nabla_{\boldsymbol{x}}^3 f$ and $\nabla_{\boldsymbol{x}}^3 c_i$ to exist. The existence of third derivatives is only necessary for analysis, and they are never computed in practice. Assumption 2 implies that $\boldsymbol{A}^k$ in (2.3) is invertible. Therefore, for any iteration k, the linear system (2.3) has a unique solution. This is a standard assumption

in the SQP literature [3]. Assumption 3 is used specifically to establish the well-posedness of Algorithm 1. Under Assumption 3, we first define in Lemma 8 the subsequence of the inner iteration $\{j_l\}_{l\geq 0}$ where the reduction of the error step occurs, and then show that the set $\{j_l\} \subset \mathbb{N} \cup \{0\}$ is bounded above. Thus, the error linearly decays in those iterations almost surely (see Lemma 9). With these results, we show in Lemma 10–12 that the step acceptance conditions (2.5), (2.6) are satisfied almost surely. The formal statements of Lemma 8–12 are presented in Appendix B. Assumption 4 is standard in the SQP literature and is needed to show local superlinear or quadratic convergence [3].

Theorem 1 (Global convergence) *Suppose Assumption 1, 2, and 3 hold for the iterates $\{(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\}_{k\geq 0}$ generated by Algorithm 1. Then $\left\|\nabla \mathcal{L}^k\right\| \to 0$ as $k \to \infty$ almost surely.*

Theorem 2 (Local linear convergence) *Let $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ be a stationary point of (1.1). Suppose Assumption 1, 2, 3, and 4 hold for the iterates $\{(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\}_{k\geq 0}$ generated by Algorithm 1, and $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \to (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$. Then for all sufficiently large k, almost surely, $\alpha_k = 1$, and*

$$\left\| \begin{pmatrix} \boldsymbol{x}^{k+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \leq \delta \left\| \begin{pmatrix} \boldsymbol{x}^{k} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^\star \end{pmatrix} \right\|$$

where $0 < \delta < 1$.

Proof of Theorem 1 and 2 are given in Appendix C.

4. Experiments

We implement three inexact SQP algorithms to solve benchmark nonlinear problems in CUTEst test set [7] and solve constrained logistic regression with datasets from LIBSVM [6]. The considered three algorithms are Algorithm 1 (the proposed algorithm), Algorithm 2: [4] with the ℓ_1 penalized merit function, and Algorithm 3: adaptive version of Algorithm 2. We use two randomized iterative solvers and one deterministic iterative solver for (2.3): Gaussian vector sketch [8, Section 3.2], Randomized Kaczmarz [8, Section 3.3], and GMRES [15]. We evaluate each algorithm with the following three criteria: (1) the KKT residual $\left(\left\|\nabla \mathcal{L}^k\right\|\right)$, (2) the number of gradient and Jacobian evaluations, and (3) the number of objective and constraints evaluations. We first present the comparison between Algorithms 1 and 3 on CUTEst set in Figure 1.

From Figure 1, we observe that Algorithm 1 outperforms Algorithm 3 in terms of the KKT residual and number of objective and constraints evaluations. This result is expected as Algorithm 1 uses tighter bounds on the residuals of the iterative solver to guarantee fast local convergence. This results in steeper decrease in the merit function at each iteration and fewer number of outer iterations. However, as we mentioned earlier, smaller number of outer iterations yields possible increase in the number of inner iterations required to satisfy (2.5). We can reduce this cost by applying IRS, which substantially saves the computational complexities by projecting (2.3) into a smaller space for each inner iteration.

On the one hand, we see Algorithm 3 shows slightly lower number of gradient and Jacobian evaluations than Algorithm 1. This is because for each iteration, Algorithm 1 finds a stepsize α_k to satisfy the Armijo condition (2.9), and $\mathcal{L}_{\boldsymbol{\eta}}^{k+1}$ in (2.9) requires the gradient and Jacobian to be evaluated at each new trial point. For Algorithm 3, however, the gradient and Jacobian are not involved in the evaluations of the ℓ_1 penalized merit function at new trial points.

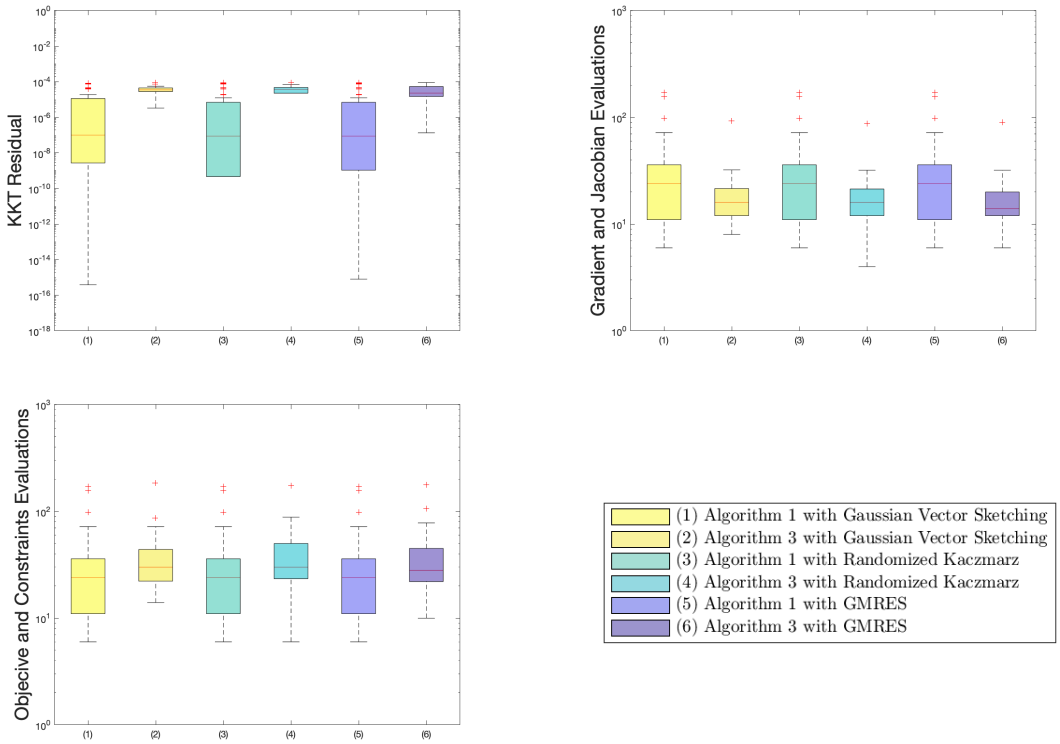

Figure 1: KKT residual, number of gradient and Jacobian evaluations, and number of objective and constraints evaluations boxplots for Algorithm 1 and Algorithm 3 on CUTEst problems.

Further comparisons between Algorithms 2 and 3, and experiments on LIBSVM datasets are in Appendix D due to the limitation of space.

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] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[3] Paul T Boggs and Jon W Tolle. Sequential quadratic programming. *Acta numerica*, 4:1–51, 1995.

[4] Richard H Byrd, Frank E Curtis, and Jorge Nocedal. An inexact sqp method for equality constrained optimization. *SIAM Journal on Optimization*, 19(1):351–369, 2008.

[5] RM Chamberlain, MJD Powell, C Lemarechal, and HC Pedersen. The watchdog technique for forcing convergence in algorithms for constrained optimization. In *Algorithms for Constrained Minimization of Smooth Nonlinear Functions*, pages 1–17. Springer, 1982.

[6] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[7] Nicholas IM 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, 2015.

[8] Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[9] S-P Han and Olvi L Mangasarian. Exact penalty functions in nonlinear programming. *Mathematical programming*, 17(1):251–269, 1979.

[10] Shih-Ping Han. A globally convergent method for nonlinear programming. *Journal of optimization theory and applications*, 22(3):297–309, 1977.

[11] Sen Na, Mihai Anitescu, and Mladen Kolar. A fast temporal decomposition procedure for long-horizon nonlinear dynamic programming. *arXiv preprint arXiv:2107.11560*, 2021.

[12] Vivak Patel, Mohammad Jahangoshahi, and Daniel A Maldonado. An implicit representation and iterative solution of randomly sketched linear systems. *SIAM Journal on Matrix Analysis and Applications*, 42(2):800–831, 2021.

[13] Michael JD Powell. A fast algorithm for nonlinearly constrained optimization calculations. In *Numerical analysis*, pages 144–157. Springer, 1978.

[14] Nathaniel Pritchard and Vivak Patel. Residual tracking and stopping for iterative random sketching. *arXiv preprint arXiv:2201.05741*, 2022.

[15] Youcef Saad and Martin H Schultz. Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, 7 (3):856–869, 1986.

[16] Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[17] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Appendix A.

Algorithm 1: Adaptive Inexact SQP via Iterative Randomized Sketching

Input: Initial iterate $(\boldsymbol{x}^0, \boldsymbol{\lambda}^0)$;
Scalars $\eta_{1,0}$, $\eta_{2,0}$, $\delta_0 \in (0,1)$; $\xi_B \in (0,1]$, $\beta \in (0, \frac{1}{2})$; $\rho \in (0,1)$; $\nu > 1$;
for $k = 0, 1, 2, \ldots$ **do**
  Compute f^k, $\nabla_{\boldsymbol{x}} f^k$, c^k, $\boldsymbol{G}^k$, and $\boldsymbol{H}^k$
  Generate $\boldsymbol{B}^k$
  Set $\boldsymbol{A}^k = \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix}$
  Compute Ψ_k and δ_k^{trial}
  Set $\begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} \leftarrow \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}$ and $\mathbf{r} \leftarrow \nabla \mathcal{L}^k$
  while $\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} > -\frac{\eta_{2,k}}{2} \left\|\nabla \mathcal{L}^k\right\|^2$ *OR* $\|\mathbf{r}\| > \frac{\delta_k}{\|\boldsymbol{A}^k\| \Psi_k} \left\|\nabla \mathcal{L}^k\right\|$ **do**
    while $\|\mathbf{r}\| > \frac{\delta_k}{\|\boldsymbol{A}^k\| \Psi_k} \left\|\nabla \mathcal{L}^k\right\|$ **do**
      Generate $\mathbf{S} \sim P$
      Set $\begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} \leftarrow \begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} - \mathbf{W}^k \mathbf{r}$
      Set $\mathbf{r} \leftarrow \boldsymbol{A}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} + \nabla \mathcal{L}^k$
    end
    if $\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} > -\frac{\eta_{2,k}}{2} \left\|\nabla \mathcal{L}^k\right\|^2$ **then**
      Set $\eta_{1,k} \leftarrow \eta_{1,k} \nu^2$, $\eta_{2,k} \leftarrow \eta_{2,k}/\nu$, and $\delta_k \leftarrow \delta_k/\nu^4$
    else if $\delta_k > \delta_k^{trial}$ **then**
      Set $\delta_k \leftarrow \delta_k^{\text{trial}}$
    else
      Set $\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \leftarrow \begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix}$
    end
  end
  Set $\alpha_k \leftarrow 1$
  while $\mathcal{L}_{\boldsymbol{\eta}}^{k+1} > \mathcal{L}_{\boldsymbol{\eta}}^k + \alpha_k \beta \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}$ **do**
    $\alpha_k \leftarrow \rho \alpha_k$
  end
  Set $\begin{pmatrix} \boldsymbol{x}^{k+1} \\ \boldsymbol{\lambda}^{k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^k \\ \boldsymbol{\lambda}^k \end{pmatrix} + \alpha_k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}$
  Set $\eta_{1,k+1} \leftarrow \eta_{1,k}$, $\eta_{2,k+1} \leftarrow \eta_{2,k}$, and $\delta_{k+1} \leftarrow \delta_k$
end

Algorithm 2: [4] with ℓ_1 penalized merit function

Input: Initial iterate $(\boldsymbol{x}^0, \boldsymbol{\lambda}^0)$;
Scalars $\kappa_1, \epsilon, \tau, \sigma, \eta \in (0,1)$; $\xi_B \in (0,1]$; $\pi_0, \beta, \kappa, \kappa_2 > 0$;
Function $\tilde{D}_\phi(\tilde{\Delta}\boldsymbol{x}^k; \pi_k) = (\nabla_{\boldsymbol{x}} f^k)^T \tilde{\Delta}\boldsymbol{x}^k - \pi_k \left(\left\|c^k\right\|_1 - \left\|\mathbf{r}^k\right\|_1\right)$, $\phi^k_{\pi_k} = f^k + \pi_k \left\|c^k\right\|_1$

for $k = 0, 1, 2, \ldots$ **do**
- Compute f^k, $\nabla_{\boldsymbol{x}} f^k$, c^k, $\boldsymbol{G}^k$, and $\boldsymbol{H}^k$
- Generate $\boldsymbol{B}^k$
- Set $\boldsymbol{A}^k = \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \mathbf{0} \end{pmatrix}$
- Set $\begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} \leftarrow \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{r} \leftarrow \nabla \mathcal{L}^k$
- **while** *Termination Test 1 OR Termination Test 2 is satisfied* **do**
  - Set $\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}$ as the jth solution of an iterative solver
- **end**
- **if** *Termination Test 2 is satisfied and model reduction condition does not hold* **then**
  - Set $\pi_k \leftarrow \pi_k^{\text{trial}} + 10^{-4}$
- **end**
- Set $\alpha_k \leftarrow 1$
- **while** $\phi^{k+1}_{\pi_k} > \phi^k_{\pi_k} + \alpha_k \eta \tilde{D}_\phi(\tilde{\Delta}\boldsymbol{x}^k; \pi_k)$ **do**
  - $\alpha_k \leftarrow \alpha_k / 2$
- **end**
- Set $\begin{pmatrix} \boldsymbol{x}^{k+1} \\ \boldsymbol{\lambda}^{k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^k \\ \boldsymbol{\lambda}^k \end{pmatrix} + \alpha_k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}$
- Set $\pi_{k+1} \leftarrow \pi_k$

end

Algorithm 3: Adaptive version of Algorithm 2

Input: Initial iterate $(\boldsymbol{x}^0, \boldsymbol{\lambda}^0)$;
Scalars $\kappa_0, \eta \in (0,1)$, $\xi_B \in (0,1]$, $\pi_0 > 0$; $\nu > 1$;
Function $\tilde{D}_\phi(\tilde{\Delta}\boldsymbol{x}^k; \pi_k) = (\nabla_{\boldsymbol{x}} f^k)^T \tilde{\Delta}\boldsymbol{x}^k - \pi_k \left(\left\|c^k\right\|_1 - \left\|\mathbf{r}^k\right\|_1\right)$, $\phi^k_{\pi_k} = f^k + \pi_k \left\|c^k\right\|_1$

for $k = 0, 1, 2, \ldots$ **do**
- Compute f^k, $\nabla_{\boldsymbol{x}} f^k$, c^k, $\boldsymbol{G}^k$, and $\boldsymbol{H}^k$
- Generate $\boldsymbol{B}^k$
- Set $\boldsymbol{A}^k = \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \mathbf{0} \end{pmatrix}$
- Set $\begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} \leftarrow \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{r} \leftarrow \nabla \mathcal{L}^k$
- **while** *Termination Test 1 is satisfied* **do**
  - **while** $\|\mathbf{r}\|_1 > \kappa \left\|\nabla \mathcal{L}^k\right\|_1$ **do**
    - Set $\begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix}$ as the jth iterative solver solution
    - Set $\mathbf{r} \leftarrow \boldsymbol{A}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix} + \nabla \mathcal{L}^k$
  - **end**
  - **if** *Model Reduction Condition is not satisfied* **then**
    - Set $\pi_k \leftarrow \pi_k \nu$, and $\kappa_k \leftarrow \kappa_k / \nu^2$
  - **else**
    - Set $\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \leftarrow \begin{pmatrix} \tilde{\Delta}\boldsymbol{x} \\ \tilde{\Delta}\boldsymbol{\lambda} \end{pmatrix}$
  - **end**
- **end**
- Set $\alpha_k \leftarrow 1$
- **while** $\phi^{k+1}_{\pi_k} > \phi^k_{\pi_k} + \alpha_k \eta \tilde{D}_\phi(\tilde{\Delta}\boldsymbol{x}^k; \pi_k)$ **do**
  - $\alpha_k \leftarrow \alpha_k / 2$
- **end**
- Set $\begin{pmatrix} \boldsymbol{x}^{k+1} \\ \boldsymbol{\lambda}^{k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^k \\ \boldsymbol{\lambda}^k \end{pmatrix} + \alpha_k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}$
- Set $\pi_{k+1} \leftarrow \pi_k$ and $\kappa_{k+1} \leftarrow \kappa_k$

end

Appendix B.

Lemma 3 (Upper bound on Hessian of Lagrangian) *For any iteration k, there exists a uniform constant $\Upsilon_{\boldsymbol{H}} > 0$, independent of k, such that for any k, $\left\|\boldsymbol{H}^k\right\| \leq \Upsilon_{\boldsymbol{H}}$.*

Proof Assume that for any iteration k, Λ is bounded at the moment. Using Assumption 1, we get

$$\left\|\boldsymbol{H}^k\right\| = \left\|\nabla^2_{\boldsymbol{xx}}\mathcal{L}^k\right\| = \left\|\nabla^2_{\boldsymbol{xx}}f^k + \sum_{i=1}^{m}\lambda_i^k\nabla^2_{\boldsymbol{xx}}c_i^k\right\| \leq \left\|\nabla^2_{\boldsymbol{xx}}f^k\right\| + \sum_{i=1}^{m}\left|\lambda_i^k\right|\left\|\nabla^2_{\boldsymbol{xx}}c_i^k\right\| \leq \Upsilon_{\boldsymbol{H}}.$$

■

Lemma 4 (Lower bound on Jacobian of constraints) *There exists a constant $\xi_{\boldsymbol{G}} > 0$ such that for any iteration k, $\boldsymbol{G}^k(\boldsymbol{G}^k)^T \succeq \xi_{\boldsymbol{G}}\boldsymbol{I}$.*

Proof Assumption 2 implies that $\boldsymbol{G}^k(\boldsymbol{G}^k)^T$ is positive definite. Let $\xi_{m,k} \geq \cdots \geq \xi_{1,k} > 0$ be the eigenvalues of $\boldsymbol{G}^k(\boldsymbol{G}^k)^T$. Then we can show $\xi_{1,k}\boldsymbol{I} \preceq \boldsymbol{G}^k(\boldsymbol{G}^k)^T \preceq \xi_{m,k}\boldsymbol{I}$. If we let $\xi_{\boldsymbol{G}} = \inf_{k\geq 0}\{\xi_{1,k}\}$, then for any iteration k, $\boldsymbol{G}^k(\boldsymbol{G}^k)^T \succeq \xi_{\boldsymbol{G}}\boldsymbol{I}$. ■

Lemma 5 (Upper bound on Hessian of Augmented Lagrangian) *For any iteration k, there exists a uniform constant $\Upsilon_{\boldsymbol{\eta}} > 0$, independent of k, such that for any k, $\left\|\nabla^2 L^k_{\boldsymbol{\eta}}\right\| \leq \Upsilon_{\boldsymbol{\eta}}$.*

Proof Using the fact that for any iteration k,

$$\nabla\mathcal{L}^k_{\boldsymbol{\eta}} = \begin{pmatrix} (\boldsymbol{I} + \eta_2\boldsymbol{H}^k)\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_1(\boldsymbol{G}^k)^T c^k \\ c^k + \eta_2\boldsymbol{G}^k\nabla_{\boldsymbol{x}}\mathcal{L}^k \end{pmatrix}$$

and

$$\nabla(\boldsymbol{M}\cdot\boldsymbol{c}) = \boldsymbol{c}\cdot\nabla(\boldsymbol{M}^T) + \boldsymbol{M}\cdot\nabla\boldsymbol{c}$$

where $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{c} \in \mathbb{R}^n$, we get

$$\begin{aligned}
\nabla^2_{\boldsymbol{xx}}\mathcal{L}^k_{\boldsymbol{\eta}} &= \nabla_{\boldsymbol{x}}\left(\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_2\boldsymbol{H}^k\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_1(\boldsymbol{G}^k)^T c^k\right) \\
&= \boldsymbol{H}^k + \eta_2\left(\nabla_{\boldsymbol{x}}\mathcal{L}^k\cdot\nabla_{\boldsymbol{x}}\boldsymbol{H}^k + (\boldsymbol{H}^k)^2\right) + \eta_1\left(c^k\cdot\nabla_{\boldsymbol{x}}((\boldsymbol{G}^k)^T) + (\boldsymbol{G}^k)^T\boldsymbol{G}^k\right), \\
\nabla^2_{\boldsymbol{\lambda\lambda}}\mathcal{L}^k_{\boldsymbol{\eta}} &= \nabla_{\boldsymbol{\lambda}}\left(c^k + \eta_2\boldsymbol{G}^k\nabla_{\boldsymbol{x}}\mathcal{L}^k\right) = \eta_2\boldsymbol{G}^k\nabla^2_{\boldsymbol{x\lambda}}\mathcal{L}^k = \eta_2\boldsymbol{G}^k(\boldsymbol{G}^k)^T, \\
\nabla^2_{\boldsymbol{x\lambda}}\mathcal{L}^k_{\boldsymbol{\eta}} &= \nabla_{\boldsymbol{\lambda}}\left(\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_2\boldsymbol{H}^k\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_1(\boldsymbol{G}^k)^T c^k\right) = (\boldsymbol{G}^k)^T + \eta_2\left(\nabla_{\boldsymbol{x}}\mathcal{L}^k\cdot\nabla_{\boldsymbol{\lambda}}\boldsymbol{H}^k + \boldsymbol{H}^k(\boldsymbol{G}^k)^T\right),
\end{aligned}$$

where $\nabla_{\boldsymbol{x}}\boldsymbol{H}^k = \nabla^3_{\boldsymbol{xxx}}f^k + \sum_{i=1}^m\lambda_i^k\nabla^3_{\boldsymbol{xxx}}c_i^k$ and $\nabla_{\boldsymbol{\lambda}}\boldsymbol{H}^k = \nabla^2_{\boldsymbol{xx}}c^k$. Using Assumption 1, we get

$$\left\|\nabla^2 L^k_{\boldsymbol{\eta}}\right\| \leq \left\|\nabla^2_{\boldsymbol{xx}}\mathcal{L}^k_{\boldsymbol{\eta}}\right\| + 2\left\|\nabla^2_{\boldsymbol{x\lambda}}\mathcal{L}^k_{\boldsymbol{\eta}}\right\| + \left\|\nabla^2_{\boldsymbol{\lambda\lambda}}\mathcal{L}^k_{\boldsymbol{\eta}}\right\| \leq \Upsilon_{\boldsymbol{\eta}}.$$

■

Lemma 6 (Upper bound on KKT matrix) *For any iteration k, there exists a uniform constant $\Upsilon_{\boldsymbol{A}} > 0$, independent of k, such that for any k, $\left\|\boldsymbol{A}^k\right\| \leq \Upsilon_{\boldsymbol{A}}$.*

Proof Using Assumption 1, we let $\Upsilon_{\boldsymbol{G}} > 0$ be the uniform upper bound of $\left\|\boldsymbol{G}^k\right\|$. Then we get

$$\left\|\boldsymbol{A}^k\right\| \leq \left\|\boldsymbol{B}^k\right\| + 2\left\|\boldsymbol{G}^k\right\| \leq \Upsilon_{\boldsymbol{B}} + 2\Upsilon_{\boldsymbol{G}} \leq \Upsilon_{\boldsymbol{A}}.$$

■

Lemma 7 (Upper bound on KKT matrix inverse) *For any iteration k, let $\Psi_k = \dfrac{7\left\|\boldsymbol{B}^k\right\|^2}{\xi_{\boldsymbol{G}}\xi_{\boldsymbol{B}}} \vee \dfrac{7}{\xi_{\boldsymbol{G}}\xi_{\boldsymbol{B}}}$. Then for any k, $\left\|(\boldsymbol{A}^k)^{-1}\right\| \leq \Psi_k$.*

Proof For any iteration k, let $\boldsymbol{Z}^k$ be a matrix which has orthonormal columns spanning the null space of $\boldsymbol{G}^k$. Assumption 2 leads to $(\boldsymbol{Z}^k)^T\boldsymbol{B}^k\boldsymbol{Z}^k \succeq \xi_{\boldsymbol{B}}\boldsymbol{I}$ and $\boldsymbol{G}^k(\boldsymbol{G}^k(\boldsymbol{G}^k)^T)^{-1}\boldsymbol{G}^k + \boldsymbol{Z}^k(\boldsymbol{Z}^k)^T = \boldsymbol{I}$. Then we can show

$$(\boldsymbol{A}^k)^{-1} = \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix}^{-1} = \begin{pmatrix} \mathcal{A}_1 & \mathcal{A}_2^T \\ \mathcal{A}_2 & \mathcal{A}_3 \end{pmatrix}$$

where

$$\mathcal{A}_1 = \boldsymbol{Z}^k((\boldsymbol{Z}^k)^T\boldsymbol{B}^k\boldsymbol{Z}^k)^{-1}(\boldsymbol{Z}^k)^T, \quad \mathcal{A}_2 = (\boldsymbol{G}^k(\boldsymbol{G}^k)^T)^{-1}\boldsymbol{G}^k(\boldsymbol{I} - \boldsymbol{B}^k\boldsymbol{Z}^k((\boldsymbol{Z}^k)^T\boldsymbol{B}^k\boldsymbol{Z}^k)^{-1}(\boldsymbol{Z}^k)^T)$$
$$\mathcal{A}_3 = (\boldsymbol{G}^k(\boldsymbol{G}^k)^T)^{-1}\boldsymbol{G}^k(\boldsymbol{B}^k\boldsymbol{Z}^k((\boldsymbol{Z}^k)^T\boldsymbol{B}^k\boldsymbol{Z}^k)^{-1}(\boldsymbol{Z}^k)^T\boldsymbol{B}^k - \boldsymbol{B}^k)\boldsymbol{G}^k(\boldsymbol{G}^k(\boldsymbol{G}^k)^T)^{-1}.$$

Using the fact that

$$\|\mathcal{A}_1\| \leq \frac{1}{\xi_{\boldsymbol{B}}}, \quad \|\mathcal{A}_2\| \leq \|(\boldsymbol{G}^k(\boldsymbol{G}^k)^T)^{-1}\boldsymbol{G}^k\|\left(1 + \frac{\left\|\boldsymbol{B}^k\right\|}{\xi_{\boldsymbol{B}}}\right) \leq \frac{1}{\sqrt{\xi_{\boldsymbol{G}}}}\left(1 + \frac{\left\|\boldsymbol{B}^k\right\|}{\xi_{\boldsymbol{B}}}\right),$$
$$\|\mathcal{A}_3\| \leq \|(\boldsymbol{G}^k(\boldsymbol{G}^k)^T)^{-1}\boldsymbol{G}^k\|^2\left(\left\|\boldsymbol{B}^k\right\| + \frac{\left\|\boldsymbol{B}^k\right\|^2}{\xi_B}\right)\| \leq \frac{1}{\xi_{\boldsymbol{G}}}\left(\left\|\boldsymbol{B}^k\right\| + \frac{\left\|\boldsymbol{B}^k\right\|^2}{\xi_{\boldsymbol{B}}}\right),$$

and without loss of generality, assuming $\left\|\boldsymbol{B}^k\right\| \geq 1 \geq \xi_{\boldsymbol{B}} \vee \xi_{\boldsymbol{G}}$, we get

$$\begin{aligned}
\left\|(\boldsymbol{A}^k)^{-1}\right\| &\leq \|\mathcal{A}_1\| + 2\|\mathcal{A}_2\| + \|\mathcal{A}_3\| \\
&\leq \frac{1}{\xi_{\boldsymbol{B}}} + \frac{2}{\sqrt{\xi_{\boldsymbol{G}}}}\left(1 + \frac{\left\|\boldsymbol{B}^k\right\|}{\xi_{\boldsymbol{B}}}\right) + \frac{1}{\xi_{\boldsymbol{G}}}\left(\left\|\boldsymbol{B}^k\right\| + \frac{\left\|\boldsymbol{B}^k\right\|^2}{\xi_{\boldsymbol{B}}}\right) \\
&\leq \frac{5\left\|\boldsymbol{B}^k\right\|}{\sqrt{\xi_{\boldsymbol{G}}}\xi_{\boldsymbol{B}}} + \frac{2\left\|\boldsymbol{B}^k\right\|^2}{\xi_{\boldsymbol{G}}\xi_{\boldsymbol{B}}} \leq \frac{7\left\|\boldsymbol{B}^k\right\|^2}{\xi_{\boldsymbol{G}}\xi_{\boldsymbol{B}}}.
\end{aligned}$$

If we let $\Psi_k = \dfrac{7\left\|\boldsymbol{B}^k\right\|^2}{\xi_{\boldsymbol{G}}\xi_{\boldsymbol{B}}} \vee \dfrac{7}{\xi_{\boldsymbol{G}}\xi_{\boldsymbol{B}}}$, then for any iteration k, $\left\|(\boldsymbol{A}^k)^{-1}\right\| \leq \Psi_k$. ■

For the later usage, we further let $\Psi = \sup_{k\geq 0}\{\Psi_k\}$ and $\Upsilon = \Upsilon_{\boldsymbol{B}} \vee \Upsilon_{\boldsymbol{G}} \vee \Upsilon_{\boldsymbol{H}}$.

Lemma 8 *Let $\mathbf{Q}^{k,j}$ be a random matrix with orthonormal columns that form a basis of $row((\mathbf{S}^{k,j})^T \boldsymbol{A}^k)$. Let $\{j_l\}_{l\geq 0}$ be a subsequence of the inner iteration where $j_0 = 0$ and j_l be the l-th iteration such that*

$$col\left(\mathbf{Q}^{k,j_{l-1}+1}\right) + \cdots + col\left(\mathbf{Q}^{k,j_l}\right) = \mathbb{R}^{n+m}.$$

Then, for any iteration k, conditioned on the event that the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$, the event

$$\mathcal{A}_{k,L} = \left\{ \cap_{l=0}^{L} \{j_l < \infty\} \right\} \tag{B.1}$$

with any $L \geq 0$, happens almost surely.

Proof Let $k \geq 0$ and assume the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$. Then the updating rule of IRS (2.4) can be rewritten as

$$\begin{aligned}\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^{k,j+1}\\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j+1}\end{pmatrix} &= \begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^{k,j}\\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j}\end{pmatrix} - \boldsymbol{W}^{k,j}\left(\boldsymbol{A}^k\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^{k,j}\\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j}\end{pmatrix} + \nabla\mathcal{L}^k\right)\\ &= \begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^{k,j}\\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j}\end{pmatrix} - \boldsymbol{W}^{k,j}\boldsymbol{A}^k\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^{k,j} - \Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j} - \Delta\boldsymbol{\lambda}^k\end{pmatrix}.\end{aligned}$$

Now let $\mathbf{e}^{k,j} = \begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^{k,j} - \Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j} - \Delta\boldsymbol{\lambda}^k\end{pmatrix}$. Then the above display can be rewritten as

$$\mathbf{e}^{k,j+1} = \mathbf{e}^{k,j} - \boldsymbol{W}^{k,j}\boldsymbol{A}^k\mathbf{e}^{k,j}. \tag{B.2}$$

Using the fact that $\boldsymbol{W}^{k,j}\boldsymbol{A}^k = (\boldsymbol{A}^k)^T\mathbf{S}^{k,j+1}((\mathbf{S}^{k,j+1})^T\boldsymbol{A}^k(\boldsymbol{A}^k)^T\mathbf{S}^{k,j+1})^{-1}(\mathbf{S}^{k,j+1})^T\boldsymbol{A}^k$ forms an orthogonal projection onto $row((\mathbf{S}^{k,j+1})^T\boldsymbol{A}^k)$ and letting $\mathbf{Q}^{k,j+1}$ be a matrix with orthonormal columns that form a basis of $row((\mathbf{S}^{k,j+1})^T\boldsymbol{A}^k)$, (B.2) can be simplified as

$$\mathbf{e}^{k,j+1} = \mathbf{e}^{k,j} - \mathbf{Q}^{k,j+1}(\mathbf{Q}^{k,j+1})^T\mathbf{e}^{k,j} = \left(\boldsymbol{I} - \mathbf{Q}^{k,j+1}(\mathbf{Q}^{k,j+1})^T\right)\mathbf{e}^{k,j}. \tag{B.3}$$

Let $j_0 = 0$ and j_l be the l-th iteration such that

$$col\left(\mathbf{Q}^{k,j_{l-1}+1}\right) + \cdots + col\left(\mathbf{Q}^{k,j_l}\right) = \mathbb{R}^{n+m}.$$

Assumption 3 implies for any inner iteration j, there exists $\pi \in (0,1]$ such that $p\left((\mathbf{Q}^{k,j})^T\boldsymbol{z} \neq 0\right) \geq \pi$. Since $\{\mathbf{Q}^{k,j}\}_{j\geq 0}$ are independent and identically distributed,

$$\begin{aligned}&p\left((\mathbf{Q}^{k,j+1})^T\boldsymbol{z} \neq 0\right)\\ &= p\left(dim\left(\sum_{i=0}^{j+1} col\left(\mathbf{Q}^{k,i}\right)\right) > dim\left(\sum_{i=0}^{j} col\left(\mathbf{Q}^{k,i}\right)\right) \middle| dim\left(\sum_{i=0}^{j} col\left(\mathbf{Q}^{k,i}\right)\right) < n+m\right)\\ &\geq \pi.\end{aligned}$$

Let $L \geq 0$. Then for any $l \leq L$, we denote the probability that the dimension increases $n+m$ times in N iterations by $p(j_l = j_{l-1} + N | j_{l-1} < \infty)$. Then the number of failure of dimension

increment $(N-n-m)$ follows the negative binomial distribution and the probability of the success of dimension increment is larger than π. Then we get

$$p(j_l = j_{l-1} + N | j_{l-1} < \infty) \leq \binom{N-1}{n+m-1} (1-\pi)^{N-n-m} \pi^{n+m}.$$

This leads to

$$\begin{aligned} p(j_l = \infty | j_{l-1} < \infty) &= \lim_{N\to\infty} p\left(j_l = j_{l-1} + N | j_{l-1} < \infty\right) \\ &\leq \lim_{N\to\infty} \binom{N-1}{n+m-1} (1-\pi)^{N-n-m} \pi^{n+m} = 0. \end{aligned}$$

Finally, we have for any $l \leq L$, $p(j_l < \infty | j_{l-1} < \infty) = 1$. For the simplicity of notation, we let $\boldsymbol{E}_l = \{j_l < \infty\}$ and obtain

$$\begin{aligned} p\left(\cap_{l=0}^{L} \{j_l < \infty\}\right) &= p(\cap_{l=0}^{L} \boldsymbol{E}_l) \\ &= p(\boldsymbol{E}_0) \times p(\boldsymbol{E}_1 | \boldsymbol{E}_0) \times p(\boldsymbol{E}_2 | \boldsymbol{E}_1, \boldsymbol{E}_0) \times \cdots \times p(\boldsymbol{E}_L | \boldsymbol{E}_{L-1}, \ldots, \boldsymbol{E}_0) \\ &= p(\boldsymbol{E}_0) \times p(\boldsymbol{E}_1 | \boldsymbol{E}_0) \times p(\boldsymbol{E}_2 | \boldsymbol{E}_1) \times \cdots \times p(\boldsymbol{E}_L | \boldsymbol{E}_{L-1}) = 1. \end{aligned}$$

This ends proof of Lemma 8. ■

Lemma 9 (Subsequence of error linearly decays) *For any iteration k and for any $L \geq 0$, conditioned on the event that the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$, the event $\mathcal{A}_{k,L}$ implies*

$$\mathcal{B}_{k,L} = \left\{ \left\| \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^{k,j_l} - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j_l} - \Delta\boldsymbol{\lambda}^k \end{pmatrix} \right\| \leq \gamma \left\| \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^{k,j_{l-1}} - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^{k,j_{l-1}} - \Delta\boldsymbol{\lambda}^k \end{pmatrix} \right\| \right\},$$

where $\gamma \in (0,1)$ and $l \leq L$.

Proof Let $k, L \geq 0$ and assume the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$. We denote $\boldsymbol{q}^{k,j,h}$ be the h-th column of $\mathbf{Q}^{k,j}$. Using (B.3) and Lemma 8, we have for any $l \leq L$,

$$\mathbf{e}^{k,j_l} = \left(\Pi_{j=j_{l-1}+1}^{j_l} \left(\Pi_{h=1}^{p} \left(\mathbf{I} - \mathbf{q}^{k,j,h} (\mathbf{q}^{k,j,h})^T\right)\right)\right) \mathbf{e}^{k,j_{l-1}}.$$

Since $\mathbf{e}^{k,j_{l-1}} \in \mathbb{R}^{n+m}$, [14, Lemma 4.1, Corollary 4.2] and [12, Theorem 4] show that on the event $\mathcal{A}_{k,L}$ in Lemma 8, there exists $\{\gamma_l\}_{l \leq L} \subset (0,1)$ such that for any $l \leq L$, $\|\mathbf{e}^{k,j_l}\| \leq \gamma_l \|\mathbf{e}^{k,j_{l-1}}\|$. Moreover, [14, Lemma 4.3] implies that we can find a uniform bound $\gamma = \sup_{l \leq L}\{\gamma_l\} \in (0,1)$ such that for $l \leq L$, $\left\|\mathbf{e}^{k,j_l}\right\| \leq \gamma \left\|\mathbf{e}^{k,j_{l-1}}\right\|$. This ends proof of Lemma 9. ■

Lemma 10 (Error of inexact direction) *For any iteration k, conditioned on the event that the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$, we have almost surely, for any $\delta \in (0,1)$, there exists $L_k > 0$ such that*

$$\left\|\mathbf{r}^k\right\| = \left\|\mathbf{r}^{k,L_k}\right\| \leq \frac{\delta}{\|\boldsymbol{A}^k\| \, \Psi_k} \left\|\nabla \mathcal{L}^k\right\|. \tag{B.4}$$

Moreover, (B.4) *implies*

$$\left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k - \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\| \leq \delta \left\|\begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\| \quad \textit{almost surely}. \tag{B.5}$$

Proof Let $k \geq 0$ and assume the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$. Applying Lemma 9, we have for any $L \geq 0$,

$$\left\|\mathbf{e}^{k,j_L}\right\| \leq (\gamma)^L \left\|\mathbf{e}^{k,j_0}\right\| \quad \textit{almost surely}.$$

Then we can find the upper bound of $\left\|\mathbf{r}^{k,L}\right\|$,

$$\begin{aligned}\left\|\mathbf{r}^{k,L}\right\| &= \left\|\boldsymbol{A}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^{k,L} \\ \tilde{\Delta}\boldsymbol{\lambda}^{k,L} \end{pmatrix} + \nabla \mathcal{L}^k\right\| = \left\|\boldsymbol{A}^k \mathbf{e}^{k,j_L}\right\| \leq \left\|\boldsymbol{A}^k\right\| \left\|\mathbf{e}^{k,j_L}\right\| \leq (\gamma)^L \left\|\boldsymbol{A}^k\right\| \left\|\mathbf{e}^{k,j_0}\right\| \\ &\leq (\gamma)^L \left\|\boldsymbol{A}^k\right\| \left\|(\boldsymbol{A}^k)^{-1}\right\| \left\|\nabla \mathcal{L}^k\right\| \leq (\gamma)^L \left\|\boldsymbol{A}^k\right\| \Psi_k \left\|\nabla \mathcal{L}^k\right\|.\end{aligned}$$

Since $\left\|\boldsymbol{A}^k\right\|$, Ψ_k, $\left\|\nabla \mathcal{L}^k\right\|$ are all uniformly bounded above and $\gamma \in (0,1)$, large L makes the upper bound of $\|\mathbf{r}^{k,L}\|$ as small as possible. Now for any iteration k, we let L_k be the inner iteration such that

$$(\gamma)^{L_k} \left\|\boldsymbol{A}^k\right\| \Psi_k \left\|\nabla \mathcal{L}^k\right\| \leq \delta \frac{\left\|\nabla \mathcal{L}^k\right\|}{\|\boldsymbol{A}^k\| \, \Psi_k}.$$

This implies

$$(\gamma)^{L_k} \leq \frac{\delta}{\|\boldsymbol{A}^k\|^2 \, \Psi_k^2} \Rightarrow L_k \leq log_\gamma \left(\frac{\delta}{\|\boldsymbol{A}^k\|^2 \, \Psi_k^2}\right) \Rightarrow L_k \leq log_\gamma \left(\frac{\delta}{\Upsilon^2 \Psi^2}\right) < \infty.$$

Using (B.4) and the fact that $\left\|(\boldsymbol{A}^k)^{-1}\right\| \leq \Psi_k$, we get

$$\begin{aligned}\left\|\mathbf{r}^k\right\| = \left\|\mathbf{r}^{k,L_k}\right\| \leq \delta \frac{\left\|\nabla \mathcal{L}^k\right\|}{\|\boldsymbol{A}^k\| \, \Psi_k} &\Rightarrow \Psi_k \left\|\boldsymbol{A}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k - \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\| \leq \delta \frac{\left\|\nabla \mathcal{L}^k\right\|}{\|\boldsymbol{A}^k\|} \\ &\Rightarrow \left\|(\boldsymbol{A}^k)^{-1}\right\| \left\|\boldsymbol{A}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k - \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\| \leq \delta \frac{\left\|\nabla \mathcal{L}^k\right\|}{\|\boldsymbol{A}^k\|} \\ &\Rightarrow \left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k - \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\| \leq \delta \frac{\left\|\nabla \mathcal{L}^k\right\|}{\|\boldsymbol{A}^k\|} \\ &\Rightarrow \left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k - \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\| \leq \delta \frac{\|\boldsymbol{A}^k\| \left\|\begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\|}{\|\boldsymbol{A}^k\|} \\ &\Rightarrow \left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k - \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\| \leq \delta \left\|\begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix}\right\| \quad \textit{almost surely}.\end{aligned}$$

■

Lemma 11 (Descent direction of inexact step) *For any iteration k, conditioned on the event that the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$, if*

$$\eta_2 \leq \frac{\xi_{\boldsymbol{B}}}{12\Upsilon^2}, \quad \eta_1 \geq \frac{17}{\eta_2 \xi_{\boldsymbol{G}}}, \quad \delta \leq \frac{\eta_2 \xi_{\boldsymbol{G}}}{16\eta_1 \Upsilon^2},$$

then we have

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \leq -\frac{\eta_2}{2} \left\| \nabla \mathcal{L}^k \right\|^2 \quad \textit{almost surely}.$$

Proof This proof is based on [11]. Let $k \geq 0$ and assume the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$. Then we have

$$\begin{aligned}
&\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \\
&= \begin{pmatrix} (\boldsymbol{I} + \eta_2 \boldsymbol{H}^k)\nabla_{\boldsymbol{x}} \mathcal{L}^k + \eta_1 (\boldsymbol{G}^k)^T c^k \\ c^k + \eta_2 \boldsymbol{G}^k \nabla_{\boldsymbol{x}} \mathcal{L}^k \end{pmatrix}^T \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \\
&= \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix}^T \begin{pmatrix} \boldsymbol{I} + \eta_2 \boldsymbol{H}^k & \eta_1 (\boldsymbol{G}^k)^T \\ \eta_2 \boldsymbol{G}^k & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}^k \\ c^k \end{pmatrix} \\
&= -\begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix}^T \begin{pmatrix} \boldsymbol{I} + \eta_2 \boldsymbol{H}^k & \eta_1 (\boldsymbol{G}^k)^T \\ \eta_2 \boldsymbol{G}^k & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \\
&= -\begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix}^T \begin{pmatrix} (\boldsymbol{I} + \eta_2 \boldsymbol{H}^k)\boldsymbol{B}^k + \eta_1 (\boldsymbol{G}^k)^T \boldsymbol{G}^k & (\boldsymbol{I} + \eta_2 \boldsymbol{H}^k)(\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k (\boldsymbol{I} + \eta_2 \boldsymbol{B}^k) & \eta_2 \boldsymbol{G}^k (\boldsymbol{G}^k)^T \end{pmatrix} \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \\
&= -(\Delta\boldsymbol{x}^k)^T \left((\boldsymbol{I} + \eta_2 \boldsymbol{H}^k)\boldsymbol{B}^k + \frac{\eta_1}{2} (\boldsymbol{G}^k)^T \boldsymbol{G}^k \right) \Delta\boldsymbol{x}^k - \frac{\eta_1}{2} (\Delta\boldsymbol{x}^k)^T (\boldsymbol{G}^k)^T \boldsymbol{G}^k \Delta\boldsymbol{x}^k \\
&\quad - \eta_2 (\Delta\boldsymbol{\lambda}^k)^T \boldsymbol{G}^k (\boldsymbol{G}^k)^T \Delta\boldsymbol{\lambda}^k - (\Delta\boldsymbol{\lambda}^k)^T \boldsymbol{G}^k \left(2\boldsymbol{I} + \eta_2 (\boldsymbol{B}^k + \boldsymbol{H}^k) \right) \Delta\boldsymbol{x}^k \\
&= -(\Delta\boldsymbol{x}^k)^T \left((\boldsymbol{I} + \eta_2 \boldsymbol{H}^k)\boldsymbol{B}^k + \frac{\eta_1}{2} (\boldsymbol{G}^k)^T \boldsymbol{G}^k \right) \Delta\boldsymbol{x}^k - \frac{\eta_1}{2} \left\| \boldsymbol{G}^k \Delta\boldsymbol{x}^k \right\|^2 - \eta_2 \left\| (\boldsymbol{G}^k)^T \Delta\boldsymbol{\lambda}^k \right\|^2 \\
&\quad - (\Delta\boldsymbol{\lambda}^k)^T \boldsymbol{G}^k \left(2\boldsymbol{I} + \eta_2 (\boldsymbol{B}^k + \boldsymbol{H}^k) \right) \Delta\boldsymbol{x}^k.
\end{aligned}$$

Using (2.2) we have

$$\boldsymbol{G}^k \Delta\boldsymbol{x} = -c^k, \quad \boldsymbol{G}^k \Delta\boldsymbol{\lambda}^k = -\left(\boldsymbol{B}^k \Delta\boldsymbol{x}^k + \nabla_{\boldsymbol{x}} \mathcal{L}^k \right)$$

and obtain

$$\begin{aligned}
&\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \\
&= -(\Delta\boldsymbol{x}^k)^T \left((\boldsymbol{I} + \eta_2 \boldsymbol{H}^k)\boldsymbol{B}^k + \frac{\eta_1}{2} (\boldsymbol{G}^k)^T \boldsymbol{G}^k \right) \Delta\boldsymbol{x}^k - \frac{\eta_1}{2} \left\| c^k \right\|^2 - \eta_2 \left\| \boldsymbol{B}^k \Delta\boldsymbol{x}^k + \nabla_{\boldsymbol{x}} \mathcal{L}^k \right\|^2 \\
&\quad - (\Delta\boldsymbol{\lambda}^k)^T \boldsymbol{G}^k \left(2\boldsymbol{I} + \eta_2 (\boldsymbol{B}^k + \boldsymbol{H}^k) \right) \Delta\boldsymbol{x}^k \\
&= -(\Delta\boldsymbol{x}^k)^T \left((\boldsymbol{I} + \eta_2 \boldsymbol{H}^k)\boldsymbol{B}^k + \frac{\eta_1}{2} (\boldsymbol{G}^k)^T \boldsymbol{G}^k \right) \Delta\boldsymbol{x}^k - \frac{\eta_1}{2} \left\| c^k \right\|^2 - \frac{\eta_2}{2} \left\| \nabla_{\boldsymbol{x}} \mathcal{L}^k \right\|^2 + \frac{\eta_2}{2} \left\| \nabla_{\boldsymbol{x}} \mathcal{L}^k \right\|^2 \\
&\quad - \eta_2 \left\| \boldsymbol{B}^k \Delta\boldsymbol{x}^k + \nabla_{\boldsymbol{x}} \mathcal{L}^k \right\|^2 - (\Delta\boldsymbol{\lambda}^k)^T \boldsymbol{G}^k \left(2\boldsymbol{I} + \eta_2 (\boldsymbol{B}^k + \boldsymbol{H}^k) \right) \Delta\boldsymbol{x}^k.
\end{aligned}$$

Now we have

$$
\begin{aligned}
&\frac{\eta_2}{2}\left\|\nabla_{\boldsymbol{x}}\mathcal{L}^k\right\|^2-\eta_2\left\|\boldsymbol{B}^k\Delta\boldsymbol{x}^k+\nabla_{\boldsymbol{x}}\mathcal{L}^k\right\|^2\\
&=-\eta_2\left\|\boldsymbol{B}^k\Delta\boldsymbol{x}^k\right\|^2-2\eta_2(\Delta\boldsymbol{x}^k)^T\boldsymbol{B}^k\nabla_{\boldsymbol{x}}\mathcal{L}^k-\frac{\eta_2}{2}\left\|\nabla_{\boldsymbol{x}}\mathcal{L}^k\right\|^2\\
&=-\eta_2\left\|\boldsymbol{B}^k\Delta\boldsymbol{x}^k\right\|^2+2\eta_2(\Delta\boldsymbol{x}^k)^T\boldsymbol{B}^k\left(\boldsymbol{B}^k\Delta\boldsymbol{x}^k+(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right)-\frac{\eta_2}{2}\left\|\boldsymbol{B}^k\Delta\boldsymbol{x}^k+(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|^2\\
&=\eta_2(\Delta\boldsymbol{x}^k)^T\boldsymbol{B}^k\boldsymbol{B}^k\Delta\boldsymbol{x}^k+\eta_2(\Delta\boldsymbol{x}^k)^T\boldsymbol{B}^k(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k-\frac{\eta_2}{2}\left\|\boldsymbol{B}^k\Delta\boldsymbol{x}^k\right\|^2-\frac{\eta_2}{2}\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|^2\\
&\leq\eta_2(\Delta\boldsymbol{x}^k)^T\boldsymbol{B}^k\boldsymbol{B}^k\Delta\boldsymbol{x}^k+\eta_2(\Delta\boldsymbol{x}^k)^T\boldsymbol{B}^k(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k-\frac{\eta_2}{2}\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|^2.
\end{aligned}
$$

This leads to

$$
\begin{aligned}
\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}^k_{\boldsymbol{\eta}}\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}^k_{\boldsymbol{\eta}}\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\leq&-(\Delta\boldsymbol{x}^k)^T\left((\boldsymbol{I}+\eta_2\boldsymbol{H}^k)\boldsymbol{B}^k+\frac{\eta_1}{2}(\boldsymbol{G}^k)^T\boldsymbol{G}^k\right)\Delta\boldsymbol{x}^k-\frac{\eta_1}{2}\left\|c^k\right\|^2-\frac{\eta_2}{2}\left\|\nabla_{\boldsymbol{x}}\mathcal{L}^k\right\|^2\\
&+\eta_2(\Delta\boldsymbol{x}^k)^T\boldsymbol{B}^k\boldsymbol{B}^k\Delta\boldsymbol{x}^k+\eta_2(\Delta\boldsymbol{\lambda}^k)^T\boldsymbol{G}^k\boldsymbol{B}^k\Delta\boldsymbol{x}^k-\frac{\eta_2}{2}\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\|_2^2\\
&-(\Delta\boldsymbol{\lambda}^k)^T\boldsymbol{G}^k\left(2\boldsymbol{I}+\eta_2(\boldsymbol{B}^k+\boldsymbol{H}^k)\right)\Delta\boldsymbol{x}^k.
\end{aligned}
$$

Using the above display, assuming $\eta_1\geq\eta_2$ at the moment we get

$$
\begin{aligned}
&\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}^k_{\boldsymbol{\eta}}\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}^k_{\boldsymbol{\eta}}\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\\
&\leq-\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2-(\Delta\boldsymbol{x}^k)^T\left((\boldsymbol{I}+\eta_2(\boldsymbol{H}^k-\boldsymbol{B}^k))\boldsymbol{B}^k+\frac{\eta_1}{2}(\boldsymbol{G}^k)^T\boldsymbol{G}^k\right)\Delta\boldsymbol{x}^k-\frac{\eta_2}{2}\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|^2\\
&\quad-(\Delta\boldsymbol{\lambda}^k)^T\boldsymbol{G}^k\left(2\boldsymbol{I}+\eta_2\boldsymbol{H}^k\right)\Delta\boldsymbol{x}^k\\
&\leq-\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2-\eta_2(\Delta\boldsymbol{x}^k)^T(\boldsymbol{H}^k-\boldsymbol{B}^k)\boldsymbol{B}^k\Delta\boldsymbol{x}^k-(\Delta\boldsymbol{x}^k)^T\boldsymbol{B}^k\Delta\boldsymbol{x}^k-\frac{\eta_1}{2}(\Delta\boldsymbol{x}^k)^T(\boldsymbol{G}^k)^T\boldsymbol{G}^k\Delta\boldsymbol{x}^k\\
&\quad-\frac{\eta_2}{2}\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|^2-(\Delta\boldsymbol{\lambda}^k)^T\boldsymbol{G}^k\left(2\boldsymbol{I}+\eta_2\boldsymbol{H}^k\right)\Delta\boldsymbol{x}^k\\
&\leq-\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2+2\eta_2\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2-(\Delta\boldsymbol{x}^k)^T\left(\boldsymbol{B}^k+\frac{\eta_1}{2}(\boldsymbol{G}^k)^T\boldsymbol{G}^k\right)\Delta\boldsymbol{x}^k-\frac{\eta_2}{2}\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|^2\\
&\quad+2\left\|\Delta\boldsymbol{\lambda}^k\right\|\left\|\boldsymbol{G}^k\Delta\boldsymbol{x}^k\right\|+\eta_2\Upsilon\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|\left\|\Delta\boldsymbol{x}^k\right\|.
\end{aligned}
$$

Applying Young's inequality: Take $\eta_2\Upsilon\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\|\|\Delta\boldsymbol{x}^k\|=\left(\sqrt{\frac{\eta_2}{2}}\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\|\right)\left(\sqrt{2\eta_2}\Upsilon\|\Delta\boldsymbol{x}^k\|\right)$ and $2\|\Delta\boldsymbol{\lambda}^k\|\|\boldsymbol{G}^k\Delta\boldsymbol{x}^k\|=\left(\frac{\sqrt{\eta_2\xi_{\boldsymbol{G}}}}{2}\|\Delta\boldsymbol{\lambda}^k\|\right)\left(\frac{4}{\sqrt{\eta_2\xi_{\boldsymbol{G}}}}\|\boldsymbol{G}^k\Delta\boldsymbol{x}^k\|\right)$. Then we get

$$
\eta_2\Upsilon\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|\left\|\Delta\boldsymbol{x}^k\right\|\leq\frac{\eta_2}{4}\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|^2+\eta_2\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2.
$$

and

$$
2\left\|\Delta\boldsymbol{\lambda}^k\right\|\left\|\boldsymbol{G}^k\Delta\boldsymbol{x}^k\right\|\leq\frac{\eta_2\xi_{\boldsymbol{G}}}{8}\left\|\Delta\boldsymbol{\lambda}^k\right\|^2+\frac{8}{\eta_2\xi_{\boldsymbol{G}}}\left\|\boldsymbol{G}^k\Delta\boldsymbol{x}^k\right\|^2.
$$

This leads to

$$\begin{aligned}
&\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\\
&\le -\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2+3\eta_2\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2+2\left\|\Delta\boldsymbol{\lambda}^k\right\|\left\|\boldsymbol{G}^k\Delta\boldsymbol{x}^k\right\|-\frac{\eta_2}{4}\left\|(\boldsymbol{G}^k)^T\Delta\boldsymbol{\lambda}^k\right\|^2\\
&\quad-(\Delta\boldsymbol{x}^k)^T\left(\boldsymbol{B}^k+\frac{\eta_1}{2}(\boldsymbol{G}^k)^T\boldsymbol{G}^k\right)\Delta\boldsymbol{x}^k\\
&\le -\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2-\frac{\eta_2}{8}\xi_{\boldsymbol{G}}\left\|\Delta\boldsymbol{\lambda}^k\right\|^2+3\eta_2\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2-(\Delta\boldsymbol{x}^k)^T\left(\boldsymbol{B}^k+\left(\frac{\eta_1}{2}-\frac{8}{\eta_2\xi_{\boldsymbol{G}}}\right)(\boldsymbol{G}^k)^T\boldsymbol{G}^k\right)\Delta\boldsymbol{x}^k.
\end{aligned}$$

Using the fact that $\Delta\boldsymbol{x}^k=\Delta\boldsymbol{u}^k+\Delta\boldsymbol{v}^k$ where $\Delta\boldsymbol{u}^k\in Null(\boldsymbol{G}^k)$ and $\Delta\boldsymbol{v}^k\in Range((\boldsymbol{G}^k)^T)$, we get $\|\Delta\boldsymbol{x}^k\|^2=\|\Delta\boldsymbol{u}^k\|^2+\|\Delta\boldsymbol{v}^k\|^2$ and $\Delta\boldsymbol{v}^k=(\boldsymbol{G}^k)^T\Delta\bar{\boldsymbol{v}}^k$ for some $\Delta\bar{\boldsymbol{v}}^k$, and obtain

$$\begin{aligned}
&3\eta_2\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2-(\Delta\boldsymbol{x}^k)^T\left(\boldsymbol{B}^k+\left(\frac{\eta_1}{2}-\frac{8}{\eta_2\xi_{\boldsymbol{G}}}\right)(\boldsymbol{G}^k)^T\boldsymbol{G}^k\right)\Delta\boldsymbol{x}^k\\
&=3\eta_{2,k}\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2-(\Delta\boldsymbol{u}^k)^T\boldsymbol{B}^k\Delta\boldsymbol{u}^k-2(\Delta\boldsymbol{u}^k)^T\boldsymbol{B}^k\Delta\boldsymbol{v}^k-(\Delta\boldsymbol{v}^k)^T\boldsymbol{B}^k\Delta\boldsymbol{v}^k-\left(\frac{\eta_1}{2}-\frac{8}{\eta_2\xi_{\boldsymbol{G}}}\right)\left\|\boldsymbol{G}^k\Delta\boldsymbol{x}^k\right\|^2\\
&\le 3\eta_{2,k}\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2-\xi_{\boldsymbol{B}}\left\|\Delta\boldsymbol{u}^k\right\|^2+2\Upsilon\left\|\Delta\boldsymbol{u}^k\right\|\left\|\Delta\boldsymbol{v}^k\right\|+\Upsilon\left\|\Delta\boldsymbol{v}^k\right\|^2-\left(\frac{\eta_1}{2}-\frac{8}{\eta_2\xi_{\boldsymbol{G}}}\right)\left\|\boldsymbol{G}^k(\boldsymbol{G}^k)^T\Delta\bar{\boldsymbol{v}}^k\right\|^2\\
&\le\left(3\eta_2\Upsilon^2-\xi_{\boldsymbol{B}}\right)\left\|\Delta\boldsymbol{x}^k\right\|^2+2\Upsilon\left\|\Delta\boldsymbol{u}^k\right\|\left\|\Delta\boldsymbol{v}^k\right\|+(\xi_{\boldsymbol{B}}+\Upsilon)\left\|\Delta\boldsymbol{v}^k\right\|^2-\left(\frac{\eta_1}{2}-\frac{8}{\eta_2\xi_{\boldsymbol{G}}}\right)\frac{1}{\xi_{\boldsymbol{G}}}\left\|(\boldsymbol{G}^k)^T\Delta\bar{\boldsymbol{v}}^k\right\|^2.
\end{aligned}$$

Applying Young's inequality: Take $2\Upsilon\|\Delta\boldsymbol{u}^k\|\|\Delta\boldsymbol{v}^k\|=(\sqrt{\xi_{\boldsymbol{B}}}\|\Delta\boldsymbol{u}^k\|)\left(\frac{2\Upsilon}{\sqrt{\xi_{\boldsymbol{B}}}}\|\Delta\boldsymbol{v}^k\|\right)$. Then we get

$$2\Upsilon\left\|\Delta\boldsymbol{u}^k\right\|\left\|\Delta\boldsymbol{v}^k\right\|\le\frac{\xi_{\boldsymbol{B}}}{2}\left\|\Delta\boldsymbol{u}^k\right\|^2+\frac{2\Upsilon^2}{\xi_{\boldsymbol{B}}}\left\|\Delta\boldsymbol{v}^k\right\|^2\le\frac{\xi_{\boldsymbol{B}}}{2}\left\|\Delta\boldsymbol{x}^k\right\|^2+\frac{2\Upsilon^2}{\xi_{\boldsymbol{B}}}\left\|\Delta\boldsymbol{v}^k\right\|^2.$$

This leads to

$$\begin{aligned}
&3\eta_2\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2-(\Delta\boldsymbol{x}^k)^T\left(\boldsymbol{B}^k+\left(\frac{\eta_1}{2}-\frac{8}{\eta_2\xi_{\boldsymbol{G}}}\right)(\boldsymbol{G}^k)^T\boldsymbol{G}^k\right)\Delta\boldsymbol{x}^k\\
&\le\left(3\eta_2\Upsilon^2-\frac{\xi_{\boldsymbol{B}}}{2}\right)\left\|\Delta\boldsymbol{x}^k\right\|^2+\frac{2\Upsilon^2}{\xi_{\boldsymbol{B}}}\left\|\Delta\boldsymbol{v}^k\right\|^2+(\xi_{\boldsymbol{B}}+\Upsilon)\left\|\Delta\boldsymbol{v}^k\right\|^2-\left(\frac{\eta_1}{2\xi_{\boldsymbol{G}}}-\frac{8}{\eta_2\xi_{\boldsymbol{G}}^2}\right)\left\|\Delta\boldsymbol{v}^k\right\|^2\\
&\le\left(3\eta_2\Upsilon^2-\frac{\xi_{\boldsymbol{B}}}{2}\right)\left\|\Delta\boldsymbol{x}^k\right\|^2+\left(\frac{2\Upsilon^2}{\xi_{\boldsymbol{B}}}+\xi_{\boldsymbol{B}}+\Upsilon+\frac{8}{\eta_2\xi_{\boldsymbol{G}}^2}-\frac{\eta_1}{2\xi_{\boldsymbol{G}}}\right)\left\|\Delta\boldsymbol{v}^k\right\|^2.
\end{aligned}$$

Without loss of generality, assuming $\Upsilon/2\ge 1\ge\xi_{\boldsymbol{B}}\vee\xi_{\boldsymbol{G}}$, we have

$$\frac{2\Upsilon^2}{\xi_{\boldsymbol{B}}}+\xi_{\boldsymbol{B}}+\Upsilon+\frac{8}{\eta_2\xi_{\boldsymbol{G}}^2}\le\frac{2\Upsilon^2}{\xi_{\boldsymbol{B}}}+\frac{3\Upsilon}{2}+\frac{8}{\eta_2\xi_{\boldsymbol{G}}^2}\le\frac{5\Upsilon^2}{\xi_{\boldsymbol{B}}}+\frac{8}{\eta_2\xi_{\boldsymbol{G}}^2}\le\frac{1}{2\eta_2}+\frac{8}{\eta_2\xi_{\boldsymbol{G}}^2}\le\frac{8.5}{\eta_2\xi_{\boldsymbol{G}}^2}.$$

This leads to

$$\begin{aligned}
&3\eta_2\Upsilon^2\left\|\Delta\boldsymbol{x}^k\right\|^2-(\Delta\boldsymbol{x}^k)^T\left(\boldsymbol{B}^k+\left(\frac{\eta_1}{2}-\frac{8}{\eta_2\xi_{\boldsymbol{G}}}\right)(\boldsymbol{G}^k)^T\boldsymbol{G}^k\right)\Delta\boldsymbol{x}^k\\
&\le\left(3\eta_2\Upsilon^2-\frac{\xi_{\boldsymbol{B}}}{2}\right)\left\|\Delta\boldsymbol{x}^k\right\|^2+\left(\frac{8.5}{\eta_2\xi_{\boldsymbol{G}}^2}-\frac{\eta_1}{2\xi_{\boldsymbol{G}}}\right)\left\|\Delta\boldsymbol{v}^k\right\|^2.
\end{aligned}$$

Combining with the previous display, we get

$$\begin{aligned}&\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\\ &\leq -\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2-\frac{\eta_2}{8}\xi_{\boldsymbol{G}}\left\|\Delta\boldsymbol{\lambda}^k\right\|^2+\left(3\eta_2\Upsilon^2-\frac{\xi_{\boldsymbol{B}}}{2}\right)\left\|\Delta\boldsymbol{x}^k\right\|^2+\left(\frac{8.5}{\eta_2\xi_{\boldsymbol{G}}^2}-\frac{\eta_1}{2\xi_{\boldsymbol{G}}}\right)\left\|\Delta\boldsymbol{v}^k\right\|^2.\end{aligned}\tag{B.6}$$

To make (B.6) negative we let

$$\eta_2\leq\frac{\xi_{\boldsymbol{B}}}{12\Upsilon^2}\quad\text{and}\quad\eta_1\geq\frac{17}{\eta_2\xi_{\boldsymbol{G}}}.\tag{B.7}$$

Then we have $\eta_1\geq\eta_2$ that justifies our previous assumption. Using (B.7) we get

$$\begin{aligned}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}&\leq-\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2-\frac{\eta_2\xi_{\boldsymbol{G}}}{8}\left\|\Delta\boldsymbol{\lambda}^k\right\|^2-\frac{\xi_{\boldsymbol{B}}}{4}\left\|\Delta\boldsymbol{x}^k\right\|^2\\ &\leq-\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2-\frac{\eta_2\xi_{\boldsymbol{G}}}{8}\left\|\Delta\boldsymbol{\lambda}^k\right\|^2-\frac{\eta_2\xi_{\boldsymbol{G}}}{8}\left\|\Delta\boldsymbol{x}^k\right\|^2\\ &\leq-\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2-\frac{\eta_2\xi_{\boldsymbol{G}}}{8}\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2.\end{aligned}$$

Using the above display, we get

$$\begin{aligned}&\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}\\ &=\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}+\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k-\Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k-\Delta\boldsymbol{\lambda}^k\end{pmatrix}\\ &=\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}+\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k-\Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k-\Delta\boldsymbol{\lambda}^k\end{pmatrix}^T\begin{pmatrix}\boldsymbol{I}+\eta_2\boldsymbol{H}^k & \eta_1(\boldsymbol{G}^k)^T\\ \eta_2\boldsymbol{G}^k & \boldsymbol{I}\end{pmatrix}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}^k\\ c^k\end{pmatrix}\\ &=\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}-\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k-\Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k-\Delta\boldsymbol{\lambda}^k\end{pmatrix}^T\begin{pmatrix}(\boldsymbol{I}+\eta_2\boldsymbol{H}^k)\boldsymbol{B}^k+\eta_1(\boldsymbol{G}^k)^T\boldsymbol{G}^k & (\boldsymbol{I}+\eta_2\boldsymbol{H}^k)(\boldsymbol{G}^k)^T\\ \boldsymbol{G}^k(\boldsymbol{I}+\eta_2\boldsymbol{B}^k) & \eta_2\boldsymbol{G}^k(\boldsymbol{G}^k)^T\end{pmatrix}\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\\ &\leq\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}+\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k-\Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k-\Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left((1+\eta_2\Upsilon)\Upsilon+(\eta_1+\eta_2)\Upsilon^2+2(1+\eta_2\Upsilon)\Upsilon\right)\\ &\leq\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}+\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k-\Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k-\Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left(3\Upsilon+4\eta_2\Upsilon^2+\eta_1\Upsilon^2\right)\\ &\leq\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}+\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k-\Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k-\Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left(3\Upsilon+\frac{\xi_{\boldsymbol{B}}}{3}+\eta_1\Upsilon^2\right).\end{aligned}$$

Without loss of generality, assuming $\Upsilon/2\geq 1\geq\xi_{\boldsymbol{B}}\vee\xi_{\boldsymbol{G}}$, we have

$$3\Upsilon+\frac{\xi_{\boldsymbol{B}}}{3}+\eta_1\Upsilon^2\leq\frac{19\Upsilon}{6}+\eta_1\Upsilon^2\leq 2\eta_1\Upsilon^2$$

and using (B.5), we obtain

$$\begin{aligned}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix} &\leq \begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix} + \left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k-\Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k-\Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|(2\eta_1\Upsilon^2)\\ &\leq -\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2 - \frac{\eta_2\xi_{\boldsymbol{G}}}{8}\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2 + 2\delta\eta_1\Upsilon^2\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\\ &\leq -\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2 - \left(\frac{\eta_2\xi_{\boldsymbol{G}}}{8} - 2\delta\eta_1\Upsilon^2\right)\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2 \quad \textit{almost surely}.\end{aligned} \tag{B.8}$$

To make (B.8) negative, we let

$$\delta \leq \frac{\eta_2\xi_{\boldsymbol{G}}}{16\eta_1\Upsilon^2}. \tag{B.9}$$

Finally, we have

$$\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix} \leq -\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2 \quad \textit{almost surely}.$$

■

Lemma 12 (Stability of penalty parameters) *For sufficiently large k, conditioned on the event that the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$, all adaptivity parameters (η_1, η_2, δ) are stabilized almost surely.*

Proof Let $k \geq 0$ and assume the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$. Applying Lemma 11, which says that if

$$\eta_1\eta_2 \geq \frac{17}{\xi_{\boldsymbol{G}}}, \quad \eta_2 \leq \frac{\xi_{\boldsymbol{B}}}{12\Upsilon^2}, \quad \frac{\delta\eta_1}{\eta_2} \leq \frac{\xi_{\boldsymbol{G}}}{16\Upsilon^2}, \tag{B.10}$$

then almost surely, the inexact direction is a descent direction. Since the updating rule of the parameters (2.7) implies

$$\eta_1\eta_2 \leftarrow \nu\eta_1\eta_2, \quad \eta_2 \leftarrow \frac{1}{\nu}\eta_2, \quad \frac{\delta\eta_1}{\eta_2} \leftarrow \frac{1}{\nu}\frac{\delta\eta_1}{\eta_2},$$

where $\nu > 1$, we have for any iteration k, (B.10) must hold. Moreover, using the fact that the upper bound of $\eta_2, \frac{\delta\eta_1}{\eta_2}$ and the lower bound of $\eta_1\eta_2$ do not depend on k, we have for sufficiently large k, all parameters are stabilized almost surely. ■

Lemma 13 (Armijo condition) *For any iteration k, conditioned on the event that the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$, almost surely, the Armijo condition (2.9) is satisfied. Moreover, there exists a uniform constant $\alpha_{\min} > 0$, independent of k, such that for any k, $0 < \alpha_{\min} \leq \alpha_k$.*

Proof Let $k \geq 0$ and assume the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$. Using (B.5) we have almost surely,

$$\left\| \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \right\| \leq \delta \left\| \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \right\| + \left\| \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \right\| \leq (\delta + 1) \left\| \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \right\| \leq 2 \left\| \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \right\|. \tag{B.11}$$

Using this fact and applying Taylor expansion for the augmented Lagrangian merit function about $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$ yields

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\eta}}^{k+1} &\leq \mathcal{L}_{\boldsymbol{\eta}}^k + \alpha_k \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + \frac{\alpha_k^2 \Upsilon_\eta}{2} \left\| \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \right\|^2 \\
&\leq \mathcal{L}_{\boldsymbol{\eta}}^k + \alpha_k \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + 2\alpha_k^2 \Upsilon_\eta \left\| \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \right\|^2 \\
&\leq \mathcal{L}_{\boldsymbol{\eta}}^k + \alpha_k \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + 2\alpha_k^2 \Upsilon_\eta \Psi^2 \left\| \nabla \mathcal{L}^k \right\|^2 \\
&\leq \mathcal{L}_{\boldsymbol{\eta}}^k + \alpha_k \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} - \frac{4\alpha_k^2 \Upsilon_\eta \Psi^2}{\eta_2} \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \\
&\leq \mathcal{L}_{\boldsymbol{\eta}}^k + \left(1 - \frac{4\Upsilon_\eta \Psi^2 \alpha_k}{\eta_2}\right) \alpha_k \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \quad \textit{almost surely}.
\end{aligned}$$

Using the above display, if

$$\left(1 - \frac{4\Upsilon_\eta \Psi^2 \alpha_k}{\eta_2}\right) \geq \beta \leftrightarrow \alpha_k \leq \frac{(1-\beta)\eta_2}{4\Upsilon_\eta \Psi^2},$$

then the Armijo condition is satisfied almost surely. Moreover, since the upper bound of α_k does not depend on k, we can find $l \geq 0$, independent of k, such that for any k, $0 < (\rho)^l \leq \dfrac{(1-\beta)\eta_2}{4\Upsilon_\eta \Psi^2}$. Finally, if we let $\alpha_{\min} = (\rho)^l$, then for any iteration k, we have $0 < \alpha_{\min} \leq \alpha_k$. ■

Lemma 14 (Boundedness of dual variable) *Almost surely, $\{\boldsymbol{\lambda}^k\}_{k\geq 0}$ produced by Algorithm 1 is bounded.*

Proof Let $k \geq 1$ and assume the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$. Using Lemma 11 and 13, we have almost surely, $\mathcal{L}_\eta(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) < \mathcal{L}_\eta(\boldsymbol{x}^0, \boldsymbol{\lambda}^0)$. Let $\left|f^k\right| \leq k_f$, $\left\|c^k\right\| \leq k_c$, and $\left\|\nabla_{\boldsymbol{x}} f^k\right\| \leq k_g$. Then we get

$$
\begin{aligned}
&\mathcal{L}_{\boldsymbol{\eta}}(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\
&= \mathcal{L}^k + \frac{\eta_1}{2}\left\|c^k\right\|^2 + \frac{\eta_2}{2}\left\|\nabla_{\boldsymbol{x}} f^k + (\boldsymbol{G}^k)^T\boldsymbol{\lambda}^k\right\|^2 \\
&= f^k + (\boldsymbol{\lambda}^k)^T c^k + \frac{\eta_1}{2}\left\|c^k\right\|^2 + \frac{\eta_2}{2}\left\|\nabla_{\boldsymbol{x}} f^k + (\boldsymbol{G}^k)^T\boldsymbol{\lambda}^k\right\|^2 \\
&= f^k + (\boldsymbol{\lambda}^k)^T c^k + \frac{\eta_1}{2}\left\|c^k\right\|^2 + \frac{\eta_2}{2}\left\|\nabla_{\boldsymbol{x}} f^k\right\|^2 + \eta_2(\boldsymbol{\lambda}^k)^T\boldsymbol{G}^k\nabla_{\boldsymbol{x}} f^k + \frac{\eta_2}{2}(\boldsymbol{\lambda}^k)^T\boldsymbol{G}^k(\boldsymbol{G}^k)^T\boldsymbol{\lambda}^k \\
&\geq f^k + (\boldsymbol{\lambda}^k)^T(c^k + \eta_2\boldsymbol{G}^k\nabla_{\boldsymbol{x}} f^k) + \frac{\eta_2}{2}\xi_{\boldsymbol{G}}\left\|\boldsymbol{\lambda}^k\right\|^2 \\
&\geq f^k - \left\|\boldsymbol{\lambda}^k\right\|\left\|c^k + \eta_2\boldsymbol{G}^k\nabla_{\boldsymbol{x}} f^k\right\| + \frac{\eta_2}{2}\xi_{\boldsymbol{G}}\left\|\boldsymbol{\lambda}^k\right\|^2 \\
&\geq -\left|f^k\right| - \left\|\boldsymbol{\lambda}^k\right\|\left(\left\|c^k\right\| + \eta_2\left\|\boldsymbol{G}^k\right\|\left\|\nabla_{\boldsymbol{x}} f^k\right\|\right) + \frac{\eta_2}{2}\xi_{\boldsymbol{G}}\left\|\boldsymbol{\lambda}^k\right\|^2 \\
&\geq -k_f - \left\|\boldsymbol{\lambda}^k\right\|(k_c + \eta_2\Upsilon_{\boldsymbol{G}}k_g) + \frac{\eta_2}{2}\xi_{\boldsymbol{G}}\left\|\boldsymbol{\lambda}^k\right\|^2.
\end{aligned}
$$

This leads to

$$
\frac{\eta_2}{2}\xi_{\boldsymbol{G}}\left\|\boldsymbol{\lambda}^k\right\|^2 - \left\|\boldsymbol{\lambda}^k\right\|(k_c + \eta_2\Upsilon_{\boldsymbol{G}}k_g) < \mathcal{L}_{\boldsymbol{\eta}}(\boldsymbol{x}^0, \boldsymbol{\lambda}^0) + k_f \leq \left|\mathcal{L}_{\boldsymbol{\eta}}(\boldsymbol{x}^0, \boldsymbol{\lambda}^0)\right| + k_f. \tag{B.12}
$$

If we let $K_1 = \frac{\eta_2}{2}\xi_{\boldsymbol{G}} > 0$, $K_2 = k_c + \eta_2\Upsilon_{\boldsymbol{G}}k_g > 0$, and $K_3 = \left|\mathcal{L}_{\boldsymbol{\eta}}(\boldsymbol{x}^0, \boldsymbol{\lambda}^0)\right| + k_f > 0$, then (B.12) can be rewritten as

$$
K_1\left\|\boldsymbol{\lambda}^k\right\|^2 - K_2\left\|\boldsymbol{\lambda}^k\right\| < K_3.
$$

This implies $\boldsymbol{\lambda}^k$ is bounded. For the simplicity of notation, we let $\boldsymbol{E}_k = \left\{\boldsymbol{\lambda}^k \text{ is bounded}\right\}$ and obtain

$$
\begin{aligned}
p\left(\cap_{k=0}^{\infty}\boldsymbol{E}_k\right) &= 1 - p(\cup_{k=0}^{\infty}\boldsymbol{E}_k^c) \geq 1 - \sum_{k=0}^{\infty} p(\boldsymbol{E}_k^c) \\
&\geq 1 - \sum_{k=0}^{\infty}\iint_{\mathcal{X}\times\Lambda} p\left(\boldsymbol{E}_k^c|(\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) p\left((\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) d(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\
&\geq 1 - \sum_{k=0}^{\infty}\iint_{\mathcal{X}\times\Lambda} 0 \cdot p\left((\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) d(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\
&\geq 1.
\end{aligned}
$$

Finally, we have $p\left(\cap_{k\geq 0}\left\{\boldsymbol{\lambda}^k \text{ is bounded}\right\}\right) = 1$. This ends proof of Lemma 14. ■

Appendix C.

C.1. Proof of Theorem 1

Proof Let $k \geq 0$ and assume the algorithm reaches $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$. Using Lemma 11 and 13, we have almost surely,

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta} \boldsymbol{x}^k \\ \tilde{\Delta} \boldsymbol{\lambda}^k \end{pmatrix} \leq -\frac{\eta_2}{2} \left\| \nabla \mathcal{L}^k \right\|^2,$$

and

$$\mathcal{L}_{\boldsymbol{\eta}}^{k+1} \leq \mathcal{L}_{\boldsymbol{\eta}}^k + \alpha_k \beta \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta} \boldsymbol{x}^k \\ \tilde{\Delta} \boldsymbol{\lambda}^k \end{pmatrix}.$$

Combining the above displays, we have

$$\mathcal{L}_{\boldsymbol{\eta}}^{k+1} - \mathcal{L}_{\boldsymbol{\eta}}^k \leq -\frac{\eta_2 \alpha_k \beta}{2} \left\| \nabla \mathcal{L}^k \right\|^2 \quad \textit{almost surely}.$$

For the simplicity of notation, we let $\boldsymbol{E}_k = \left\{ \mathcal{L}_{\boldsymbol{\eta}}^{k+1} - \mathcal{L}_{\boldsymbol{\eta}}^k \leq -\frac{\eta_2 \alpha_k \beta}{2} \left\| \nabla \mathcal{L}^k \right\|^2 \right\}$ and obtain

$$\begin{aligned} p\left(\cap_{k=0}^{\infty} \boldsymbol{E}_k\right) &= 1 - p(\cup_{k=0}^{\infty} \boldsymbol{E}_k^c) \geq 1 - \sum_{k=0}^{\infty} p(\boldsymbol{E}_k^c) \\ &\geq 1 - \sum_{k=0}^{\infty} \iint_{\mathcal{X} \times \Lambda} p\left(\boldsymbol{E}_k^c | (\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) p\left((\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) d(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\ &\geq 1 - \sum_{k=0}^{\infty} \iint_{\mathcal{X} \times \Lambda} 0 \cdot p\left((\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) d(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\ &\geq 1. \end{aligned}$$

This implies $p\left(\cap_{k \geq 0} \left\{ \mathcal{L}_{\boldsymbol{\eta}}^{k+1} - \mathcal{L}_{\boldsymbol{\eta}}^k \leq -\frac{\eta_2 \alpha_k \beta}{2} \left\| \nabla \mathcal{L}^k \right\|^2 \right\}\right) = 1$. Now using the fact that

$$\begin{aligned} \left\{ \cap_{k \geq 0} \left\{ \mathcal{L}_{\boldsymbol{\eta}}^{k+1} - \mathcal{L}_{\boldsymbol{\eta}}^k \leq -\frac{\eta_2 \alpha_k \beta}{2} \left\| \nabla \mathcal{L}^k \right\|^2 \right\} \right\} &\Rightarrow \left\{ \cap_{k \geq 0} \left\{ \mathcal{L}_{\boldsymbol{\eta}}^{k+1} - \mathcal{L}_{\boldsymbol{\eta}}^k \leq -\frac{\eta_2 \alpha_{\min} \beta}{2} \left\| \nabla \mathcal{L}^k \right\|^2 \right\} \right\} \\ &\Rightarrow \left\{ \frac{\eta_2 \beta \alpha_{\min}}{2} \sum_{k=0}^{\infty} \left\| \nabla \mathcal{L}^k \right\|^2 \leq \mathcal{L}_{\boldsymbol{\eta}}^0 - \min_{\mathcal{X} \times \Lambda} \{ \mathcal{L}_{\eta}(\boldsymbol{x}, \boldsymbol{\lambda}) \} \right\} \\ &\Rightarrow \left\{ \sum_{k=0}^{\infty} \left\| \nabla \mathcal{L}^k \right\|^2 < \infty \right\} \\ &\Rightarrow \left\{ \left\| \nabla \mathcal{L}^k \right\| \to 0 \text{ as } k \to \infty \right\}, \end{aligned}$$

we have $p\left(\left\{ \left\| \nabla \mathcal{L}^k \right\| \to 0 \text{ as } k \to \infty \right\}\right) = 1$. This ends proof of Theorem 1. ■

C.2. Proof of Theorem 2

Proof We first show for all sufficiently large k, almost surely, unit stepsizes are admissible. It suffices to show that for all sufficiently large k,

$$\mathcal{L}_{\boldsymbol{\eta}}(\boldsymbol{x}^k + \tilde{\Delta}\boldsymbol{x}^k, \boldsymbol{\lambda}^k + \tilde{\Delta}\boldsymbol{\lambda}^k) \leq \mathcal{L}_{\boldsymbol{\eta}}^k + \beta \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \quad \textit{almost surely}. \tag{C.1}$$

Using Lemma 5 we have

$$\begin{aligned}
\nabla^2_{\boldsymbol{xx}}\mathcal{L}_{\boldsymbol{\eta}}^k &= \nabla_{\boldsymbol{x}}\left(\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_2\boldsymbol{H}^k\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_1(\boldsymbol{G}^k)^T c^k\right) \\
&= \boldsymbol{H}^k + \eta_2\left(\nabla_{\boldsymbol{x}}\mathcal{L}^k \cdot \nabla_{\boldsymbol{x}}\boldsymbol{H}^k + (\boldsymbol{H}^k)^2\right) + \eta_1\left(c^k \cdot \nabla_{\boldsymbol{x}}((\boldsymbol{G}^k)^T) + (\boldsymbol{G}^k)^T\boldsymbol{G}^k\right), \\
\nabla^2_{\boldsymbol{\lambda\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k &= \nabla_{\boldsymbol{\lambda}}\left(c^k + \eta_2\boldsymbol{G}^k\nabla_{\boldsymbol{x}}\mathcal{L}^k\right) = \eta_2\boldsymbol{G}^k\nabla^2_{\boldsymbol{x\lambda}}\mathcal{L}^k = \eta_2\boldsymbol{G}^k(\boldsymbol{G}^k)^T, \\
\nabla^2_{\boldsymbol{x\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k &= \nabla_{\boldsymbol{\lambda}}\left(\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_2\boldsymbol{H}^k\nabla_{\boldsymbol{x}}\mathcal{L}^k + \eta_1(\boldsymbol{G}^k)^T c^k\right) = (\boldsymbol{G}^k)^T + \eta_2\left(\nabla_{\boldsymbol{x}}\mathcal{L}^k \cdot \nabla_{\boldsymbol{\lambda}}\boldsymbol{H}^k + \boldsymbol{H}^k(\boldsymbol{G}^k)^T\right),
\end{aligned}$$

where $\nabla_{\boldsymbol{x}}\boldsymbol{H}^k = \nabla^3_{\boldsymbol{xxx}}f^k + \sum_{i=1}^m \lambda_i^k\nabla^3_{\boldsymbol{xxx}}c_i^k$ and $\nabla_{\boldsymbol{\lambda}}\boldsymbol{H}^k = \nabla^2_{\boldsymbol{xx}}c^k$. Then using Assumption 1, we have that $\nabla^2\mathcal{L}_{\boldsymbol{\eta}}^k$ is continuous. Now we let

$$\mathcal{H}^k = \begin{pmatrix} \boldsymbol{H}^k + \eta_2\boldsymbol{H}_k^2 + \eta_1(\boldsymbol{G}^k)^T\boldsymbol{G}^k & (\boldsymbol{G}^k)^T + \eta_2\boldsymbol{H}^k(\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k + \eta_2\boldsymbol{G}^k\boldsymbol{H}^k & \eta_2\boldsymbol{G}^k(\boldsymbol{G}^k)^T \end{pmatrix}.$$

Then using the fact that $\|\nabla\mathcal{L}^k\| = \left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}^k \\ c^k \end{pmatrix}\right\| = o(1)$, we have $\|\nabla^2\mathcal{L}_{\boldsymbol{\eta}}^k - \mathcal{H}^k\| = o(1)$. Now we let $\Upsilon_k = \|\boldsymbol{B}^k\| \vee \|\boldsymbol{H}^k\| \vee \|\boldsymbol{G}^k\|$ and applying Taylor expansion for the augmented Lagrangian merit function about $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$ yields

$$\begin{aligned}
&\mathcal{L}_{\eta}(\boldsymbol{x}^k + \tilde{\Delta}\boldsymbol{x}^k, \boldsymbol{\lambda}^k + \tilde{\Delta}\boldsymbol{\lambda}^k) \\
&\leq \mathcal{L}_{\boldsymbol{\eta}}^k + \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}^T \nabla^2\mathcal{L}_{\boldsymbol{\eta}}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + o\left(\left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}\right\|^2\right) \\
&= \mathcal{L}_{\boldsymbol{\eta}}^k + \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}^T \mathcal{H}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + o\left(\left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}\right\|^2\right) \\
&= \mathcal{L}_{\boldsymbol{\eta}}^k + \frac{1}{2}\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}^T \mathcal{H}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + o\left(\left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}\right\|^2\right) \\
&= \mathcal{L}_{\boldsymbol{\eta}}^k + \frac{1}{2}\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}^T \mathcal{H}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \\
&\quad - \frac{1}{2}\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}^T \begin{pmatrix} (\boldsymbol{I} + \eta_2\boldsymbol{H}^k)\boldsymbol{B}^k + \eta_1(\boldsymbol{G}^k)^T\boldsymbol{G}^k & (\boldsymbol{I} + \eta_2\boldsymbol{H}^k)(\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k(\boldsymbol{I} + \eta_2\boldsymbol{B}^k) & \eta_2\boldsymbol{G}^k(\boldsymbol{G}^k)^T \end{pmatrix} \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} + o\left(\left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}\right\|^2\right) \\
&= \mathcal{L}_{\boldsymbol{\eta}}^k + \frac{1}{2}\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}^T \mathcal{H}^k \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k - \Delta\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k - \Delta\boldsymbol{\lambda}^k \end{pmatrix} \\
&\quad + \frac{1}{2}\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}^T \left(\mathcal{H}^k - \begin{pmatrix} (\boldsymbol{I} + \eta_2\boldsymbol{H}^k)\boldsymbol{B}^k + \eta_1(\boldsymbol{G}^k)^T\boldsymbol{G}^k & (\boldsymbol{I} + \eta_2\boldsymbol{H}^k)(\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k(\boldsymbol{I} + \eta_2\boldsymbol{B}^k) & \eta_2\boldsymbol{G}^k(\boldsymbol{G}^k)^T \end{pmatrix}\right) \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} + o\left(\left\|\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix}\right\|^2\right).
\end{aligned}$$

This leads to

$$\mathcal{L}_\eta(\boldsymbol{x}^k+\tilde{\Delta}\boldsymbol{x}^k,\boldsymbol{\lambda}^k+\tilde{\Delta}\boldsymbol{\lambda}^k) \le \mathcal{L}_{\boldsymbol{\eta}}^k + \frac{1}{2}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix} + \frac{1}{2}\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}^T\mathcal{H}^k\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k-\Delta\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k-\Delta\boldsymbol{\lambda}^k\end{pmatrix}$$
$$+\frac{1}{2}\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}^T\begin{pmatrix}(\boldsymbol{I}+\eta_2\boldsymbol{H}^k)(\boldsymbol{H}^k-\boldsymbol{B}^k) & \boldsymbol{0}\\ \eta_2\boldsymbol{G}^k(\boldsymbol{H}^k-\boldsymbol{B}^k) & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix} + o\left(\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\right).$$

Using (B.5), (B.11) and the fact that $o\left(\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\right) = o\left(\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\right)$, we get

$$\begin{aligned}
&\mathcal{L}_\eta(\boldsymbol{x}^k+\tilde{\Delta}\boldsymbol{x}^k,\boldsymbol{\lambda}^k+\tilde{\Delta}\boldsymbol{\lambda}^k)\\
&\le \mathcal{L}_{\boldsymbol{\eta}}^k + \frac{1}{2}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix} + \frac{1}{2}\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left\|\mathcal{H}^k\right\|\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}-\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\\
&\quad+\frac{1}{2}\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}\right\|\left(\left\|\boldsymbol{I}+\eta_2\boldsymbol{H}^k\right\|\left\|(\boldsymbol{H}^k-\boldsymbol{B}^k)\Delta\boldsymbol{x}^k\right\|+\left\|\eta_2\boldsymbol{G}^k\right\|\left\|(\boldsymbol{H}^k-\boldsymbol{B}^k)\Delta\boldsymbol{x}^k\right\|\right)+o\left(\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\right)\\
&\le \mathcal{L}_{\boldsymbol{\eta}}^k + \frac{1}{2}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix} + \frac{1}{2}\left\|\mathcal{H}^k\right\|\left(2\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\right)\left(\delta\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|\right)\\
&\quad+(1+2\eta_2\Upsilon_k)\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|o\left(\left\|\Delta\boldsymbol{x}^k\right\|\right)+o\left(\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\right)\\
&\le \mathcal{L}_{\boldsymbol{\eta}}^k + \frac{1}{2}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix} + \delta(3\Upsilon_k+4\eta_2\Upsilon_k^2+\eta_1\Upsilon_k^2)\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2+o\left(\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\right)\quad \textit{almost surely}.
\end{aligned}$$

Now if we let

$$\delta \le \left(\frac{1}{2}-\beta\right)\frac{\eta_2}{2\Psi_k^2(3\Upsilon_k+4\eta_2\Upsilon_k^2+\eta_1\Upsilon_k^2)},$$

then we have

$$\begin{aligned}
&\delta(3\Upsilon_k+4\eta_2\Upsilon_k^2+\eta_1\Upsilon_k^2) \le \left(\frac{1}{2}-\beta\right)\frac{\eta_2}{2\Psi_k^2}\\
&\Rightarrow \delta(3\Upsilon_k+4\eta_2\Upsilon_k^2+\eta_1\Upsilon_k^2)\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2 \le \left(\frac{1}{2}-\beta\right)\frac{\eta_2}{2\Psi_k^2}\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\\
&\Rightarrow \delta(3\Upsilon_k+4\eta_2\Upsilon_k^2+\eta_1\Upsilon_k^2)\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2 \le \left(\frac{1}{2}-\beta\right)\frac{\eta_2}{2}\left\|\nabla\mathcal{L}^k\right\|^2\\
&\Rightarrow \delta(3\Upsilon_k+4\eta_2\Upsilon_k^2+\eta_1\Upsilon_k^2)\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2 \le -\left(\frac{1}{2}-\beta\right)\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}\\
&\Rightarrow \frac{1}{2}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix} + \delta(3\Upsilon_k+4\eta_2\Upsilon_k^2+\eta_1\Upsilon_k^2)\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2 \le \beta\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}.
\end{aligned}$$

We let $K_1 \ge 0$ be such that for any iteration $k \ge K_1$,

$$\frac{1}{2}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}+\delta(3\Upsilon_k+4\eta_2\Upsilon_k^2+\eta_1\Upsilon_k^2)\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2+o\left(\left\|\begin{pmatrix}\Delta\boldsymbol{x}^k\\ \Delta\boldsymbol{\lambda}^k\end{pmatrix}\right\|^2\right) \le \beta\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{x}^k\\ \tilde{\Delta}\boldsymbol{\lambda}^k\end{pmatrix}.$$

Then if $\delta \leq \left(\frac{1}{2} - \beta\right) \dfrac{\eta_2}{2\Psi_k^2(3\Upsilon_k + 4\eta_2\Upsilon_k^2 + \eta_1\Upsilon_k^2)}$ where $0 < \beta < \dfrac{1}{2}$, we have for any iteration $k \geq K_1$,

$$\mathcal{L}_\eta(\boldsymbol{x}^k + \tilde{\Delta}\boldsymbol{x}^k, \boldsymbol{\lambda}^k + \tilde{\Delta}\boldsymbol{\lambda}^k) \leq \mathcal{L}_{\boldsymbol{\eta}}^k + \beta \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \quad \textit{almost surely}.$$

For the simplicity of notation, we let

$$\boldsymbol{E}_k = \left\{ \mathcal{L}_\eta(\boldsymbol{x}^k + \tilde{\Delta}\boldsymbol{x}^k, \boldsymbol{\lambda}^k + \tilde{\Delta}\boldsymbol{\lambda}^k) \leq \mathcal{L}_{\boldsymbol{\eta}}^k + \beta \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \right\}$$

and obtain

$$\begin{aligned} p\left(\cap_{k=K_1}^\infty \boldsymbol{E}_k\right) &= 1 - p(\cup_{k=K_1}^\infty \boldsymbol{E}_k^c) \geq 1 - \sum_{k=K_1}^\infty p(\boldsymbol{E}_k^c) \\ &\geq 1 - \sum_{k=K_1}^\infty \iint_{\mathcal{X}\times\Lambda} p\left(\boldsymbol{E}_k^c | (\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) p\left((\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) d(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\ &\geq 1 - \sum_{k=K_1}^\infty \iint_{\mathcal{X}\times\Lambda} 0 \cdot p\left((\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) d(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\ &\geq 1. \end{aligned}$$

Finally, we have

$$p\left(\cap_{k=K_1}^\infty \left\{ \mathcal{L}_\eta(\boldsymbol{x}^k + \tilde{\Delta}\boldsymbol{x}^k, \boldsymbol{\lambda}^k + \tilde{\Delta}\boldsymbol{\lambda}^k) \leq \mathcal{L}_{\boldsymbol{\eta}}^k + \beta \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{\eta}}^k \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\boldsymbol{\eta}}^k \end{pmatrix}^T \begin{pmatrix} \tilde{\Delta}\boldsymbol{x}^k \\ \tilde{\Delta}\boldsymbol{\lambda}^k \end{pmatrix} \right\}\right) = 1.$$

■

Proof Next, we show for all sufficiently large k

$$\left\| \begin{pmatrix} \boldsymbol{x}^k + \tilde{\Delta}\boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k + \tilde{\Delta}\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \leq \delta \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \quad \textit{almost surely},$$

where $0 < \delta < 1$. We first consider

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{x}^k + \Delta\boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k + \Delta\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} &= \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} + \begin{pmatrix} \Delta\boldsymbol{x}^k \\ \Delta\boldsymbol{\lambda}^k \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} - \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix}^{-1} \nabla\mathcal{L}^k \\
&= \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix}^{-1} \left(\begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} - \nabla\mathcal{L}^k \right) \\
&= \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix}^{-1} \left(\begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} - \left(\nabla\mathcal{L}^k - \nabla\mathcal{L}^\star \right) \right).
\end{aligned}$$

Using Taylor's theorem and obtain

$$\begin{aligned}
\nabla\mathcal{L}^k - \nabla\mathcal{L}^\star &= \int_0^1 \nabla^2\mathcal{L}\left(\boldsymbol{x}^k + t(\boldsymbol{x}^\star - \boldsymbol{x}^k), \boldsymbol{\lambda}^k + t(\boldsymbol{\lambda}^\star - \boldsymbol{\lambda}^k) \right) \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} dt \\
&= \int_0^1 \begin{pmatrix} \boldsymbol{H}\left(\boldsymbol{x}^k + t(\boldsymbol{x}^\star - \boldsymbol{x}^k), \boldsymbol{\lambda}^k + t(\boldsymbol{\lambda}^\star - \boldsymbol{\lambda}^k) \right) & \boldsymbol{G}^T(\boldsymbol{x}^k + t(\boldsymbol{x}^\star - \boldsymbol{x}^k)) \\ \boldsymbol{G}(\boldsymbol{x}^k + t(\boldsymbol{x}^\star - \boldsymbol{x}^k)) & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} dt.
\end{aligned}$$

Now we let $\boldsymbol{H}(t) = \boldsymbol{H}\left(\boldsymbol{x}^k + t(\boldsymbol{x}^\star - \boldsymbol{x}^k), \boldsymbol{\lambda}^k + t(\boldsymbol{\lambda}^\star - \boldsymbol{\lambda}^k) \right)$ and $\boldsymbol{G}(t) = \boldsymbol{G}(\boldsymbol{x}^k + t(\boldsymbol{x}^\star - \boldsymbol{x}^k))$. Then we have

$$\begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} - \left(\nabla\mathcal{L}^k - \nabla\mathcal{L}^\star \right) = \int_0^1 \begin{pmatrix} \boldsymbol{B}^k - \boldsymbol{H}(t) & (\boldsymbol{G}^k)^T - (\boldsymbol{G}(t))^T \\ \boldsymbol{G}^k - \boldsymbol{G}(t) & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} dt.$$

From the fact that f^k and c^k are thrice continuously differentiable and $\boldsymbol{H}^k = \nabla^2_{\boldsymbol{xx}} f^k + \sum_{i=1}^m \lambda_i^k \nabla^2_{\boldsymbol{xx}} c_i^k$, we know $\boldsymbol{H}^k$ and $\boldsymbol{G}^k$ are locally Lipschitz continuous. Using this fact and Assumption 4, and taking ℓ_2 norm on both sides yields,

$$\begin{aligned}
&\left\| \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} - \left(\nabla\mathcal{L}^k - \nabla\mathcal{L}^\star \right) \right\| \\
&= \left\| \int_0^1 \begin{pmatrix} \boldsymbol{B}^k - \boldsymbol{H}(t) & (\boldsymbol{G}^k)^T - (\boldsymbol{G}(t))^T \\ \boldsymbol{G}^k - \boldsymbol{G}(t) & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} dt \right\| \\
&\leq \int_0^1 \left\| \begin{pmatrix} \boldsymbol{B}^k - \boldsymbol{H}(t) & (\boldsymbol{G}^k)^T - (\boldsymbol{G}(t))^T \\ \boldsymbol{G}^k - \boldsymbol{G}(t) & \boldsymbol{0} \end{pmatrix} \right\| \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| dt \\
&\leq \int_0^1 \left(\|\boldsymbol{B}^k - \boldsymbol{H}^k\| + \|\boldsymbol{H}^k - \boldsymbol{H}(t)\| + 2\|\boldsymbol{G}^k - \boldsymbol{G}(t)\| \right) \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| dt \\
&\leq o\left(\left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \right) + O\left(\left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\|^2 \right) \\
&\leq o\left(\left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \right).
\end{aligned}$$

Using Lemma 7, we have $\left\| \begin{pmatrix} \boldsymbol{B}^k & (\boldsymbol{G}^k)^T \\ \boldsymbol{G}^k & \boldsymbol{0} \end{pmatrix}^{-1} \right\|$ is bounded above. Finally, we get

$$\left\| \begin{pmatrix} \boldsymbol{x}^k + \Delta \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k + \Delta \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \leq o\left(\left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \right). \tag{C.2}$$

Using (C.2) we have

$$\begin{aligned} \left\| \begin{pmatrix} \Delta \boldsymbol{x}^k \\ \Delta \boldsymbol{\lambda}^k \end{pmatrix} \right\| &\leq \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| + \left\| \begin{pmatrix} \boldsymbol{x}^k + \Delta \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k + \Delta \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \\ &\leq \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| + o\left(\left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \right). \end{aligned} \tag{C.3}$$

Now applying (B.5), (C.2), and (C.3), we get

$$\begin{aligned} \left\| \begin{pmatrix} \boldsymbol{x}^k + \tilde{\Delta} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k + \tilde{\Delta} \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| &\leq \left\| \begin{pmatrix} \tilde{\Delta} \boldsymbol{x}^k - \Delta \boldsymbol{x}^k \\ \tilde{\Delta} \boldsymbol{\lambda}^k - \Delta \boldsymbol{\lambda}^k \end{pmatrix} \right\| + \left\| \begin{pmatrix} \boldsymbol{x}^k + \Delta \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k + \Delta \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \\ &\leq \delta \left\| \begin{pmatrix} \Delta \boldsymbol{x}^k \\ \Delta \boldsymbol{\lambda}^k \end{pmatrix} \right\| + o\left(\left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \right) \\ &\leq \delta \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| + o\left(\left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \right) \quad \textit{almost surely}. \end{aligned}$$

We let $\bar{K} \geq 0$ be the iteration such that $\left\| \begin{pmatrix} \boldsymbol{x}^k + \tilde{\Delta} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k + \tilde{\Delta} \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \leq \delta \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\|$ holds. Now we let $K_2 = \bar{K} \vee K_1$. Then, we have for any iteration $k \geq K_2$,

$$\left\| \begin{pmatrix} \boldsymbol{x}^{k+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| = \left\| \begin{pmatrix} \boldsymbol{x}^k + \tilde{\Delta} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k + \tilde{\Delta} \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \leq \delta \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \quad \textit{almost surely}.$$

For the simplicity of notation, we let

$$\boldsymbol{E}_k = \left\{ \left\| \begin{pmatrix} \boldsymbol{x}^{k+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \leq \delta \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \right\}$$

and obtain

$$\begin{aligned} p\left(\cap_{k=K_2}^{\infty} \boldsymbol{E}_k\right) &= 1 - p(\cup_{k=K_2}^{\infty} \boldsymbol{E}_k^c) \geq 1 - \sum_{k=K_2}^{\infty} p(\boldsymbol{E}_k^c) \\ &\geq 1 - \sum_{k=K_2}^{\infty} \iint_{\mathcal{X} \times \Lambda} p\left(\boldsymbol{E}_k^c | (\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) p\left((\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) d(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\ &\geq 1 - \sum_{k=K_2}^{\infty} \iint_{\mathcal{X} \times \Lambda} 0 \cdot p\left((\boldsymbol{X}^k, \boldsymbol{\Lambda}^k) = (\boldsymbol{x}^k, \boldsymbol{\lambda}^k)\right) d(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) \\ &\geq 1. \end{aligned}$$

Finally, we have

$$p\left(\cap_{k=K_2}^{\infty} \left\{ \left\| \begin{pmatrix} \boldsymbol{x}^{k+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \leq \delta \left\| \begin{pmatrix} \boldsymbol{x}^k - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^\star \end{pmatrix} \right\| \right\} \right) = 1.$$

■

Appendix D.

D.1. CUTEst

Among all the problems in the CUTEst test set, we selected the problems for which f is not a constant objective with $n < 1000$, containing only equality constraints, positive definiteness of $\boldsymbol{G}^k(\boldsymbol{G}^k)^T$ at all iterates of all algorithm that we ran. This selection scheme yields a total of 47 problems.

Implementation Details.

(a) Algorithm 1: Adaptive inexact SQP with the augmented Lagrangian merit function. The parameters are set as $\eta_{1,0} = 1$, $\eta_{2,0} = 0.1$, $\delta_0 = 0.1$, $\xi_B = 0.1$, $\beta = 0.1$, $\nu = 1.4$, $\rho = 0.5$.

(b) Algorithm 2: Byrd et al. [4] with the ℓ_1 penalized merit function. We let $\pi_0 = 1$, $\kappa = 1$, $\kappa_1 = 0.1$, $\epsilon = 0.1$, $\tau = 0.1$, $\eta = 10^{-8}$. Likewise, the remaining parameters are set as $\xi_B = 0.1$, $\sigma = \tau(1-\epsilon)$, and $\kappa_2 = \beta = \dfrac{\|\nabla \mathcal{L}^0\|_1}{\|c^0\|_1 + 1} \vee 1$.

(c) Algorithm 3: Adaptive version of Algorithm 2. The parameters are set as $\pi_0 = 1$, $\kappa_0 = 0.1$, $\eta = 10^{-8}$, $\xi_B = 0.1$, $\beta = 0.1$, $\nu = 1.4$.

It is worth mentioning that the step acceptance condition of Algorithm 1 is directly correspondent to the Termination Test 1 in Byrd et al. [4]. However, the algorithmic differences lie in two points. First, the bounds on the residual norm in (2.5) uses an adaptive parameter δ, however, Byrd et al. [4] used a fixed constant κ. Second, Algorithm 1 updates the penalty parameters until (2.6) is satisfied, on the other hand, Byrd et al. [4] utilizes additional Termination Test to update the penalty parameters. Therefore, we can view Algorithm 1 equipped with the ℓ_1 penalized merit function as an adaptive version of Byrd et al. [4] when only the Termination Test 1 is used.

Throughout the experiments, we use the initial value of primal-dual variables which are provided by the CUTEst package. For the Hessian modification, we regularize the Hessian $\boldsymbol{H}^k$ by $\boldsymbol{B}^k = \boldsymbol{H}^k + (\xi_{\boldsymbol{B}} + \|\boldsymbol{H}^k\|)\boldsymbol{I}_n$ whenever $\boldsymbol{H}^k$ does not satisfy Assumption 2. For each algorithm, we average over 10 independent runs. The stopping criterion is set as:

$$\left\|\nabla \mathcal{L}^k\right\| \leq 10^{-4} \quad \text{OR} \quad k \geq 10^4.$$

If the algorithm terminates from the former stopping criterion, we say the algorithm converges, otherwise the latter stopping criterion would be satisfied.

Comparisons. We compare the performance of Algorithm 2 and Algorithm 3 on CUTEst set in Figure 2. From Figure 2, we observe that Algorithm 3 is superior than Algorithm 2 in all three criteria. This is because Algorithm 3 adaptively controls the accuracy of the inexact solution and this adaptive scheme yields tighter bounds on the residuals of the iterative solver. This results in steeper decrease in the merit function at each iteration and smaller number of outer iterations. Since both algorithms use the ℓ_1 penalized merit function, both algorithm do not involve gradient and Jacobian evaluations when we find a stepsize α_k.

D.2. Constrained Logistic Regression

We follow the experiment details in Berahas et al. [1]. Among all datasets in the LIBSVM collection, we consider 7 binary classification datasets for which $12 \leq n \leq 1000$, $256 \leq N \leq 100000$, and positive definiteness of $\boldsymbol{G}^k(\boldsymbol{G}^k)^T$ at all iterates of all algorithm we ran. For the linear constraints, we fix $m = 10$ and randomly generate each entry of $\mathbf{A}$ and $\mathbf{b}$ from a standard normal

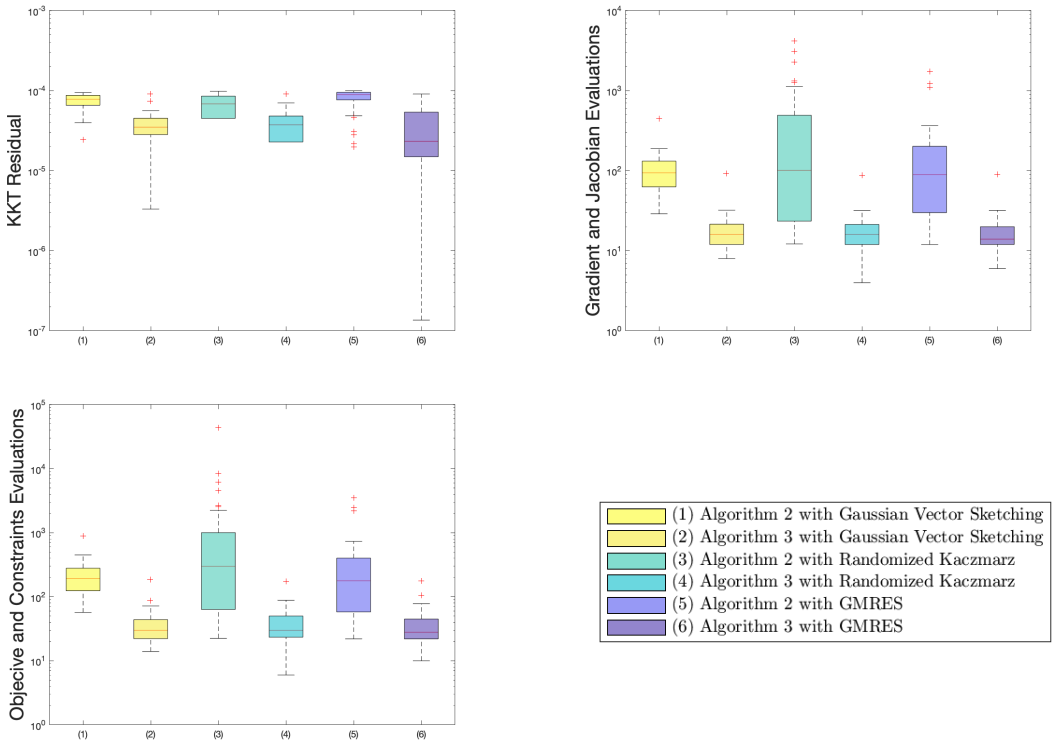

Figure 2: KKT residual, number of gradient and Jacobian evaluations, and number of objective and constraints evaluations boxplots for Algorithm 2 and Algorithm 3 on CUTEst problems.

distribution for each problem. Combining with the norm constraint, we use total of 11 number of constraints. For all problems and algorithms, we set the initial primal and dual iterates as the vector of all ones. Details of the datasets are in Table 1.

Table 1: Details of the datasets used in experiments on the LIBSVM collection.

Dataset	Primal dimension	Number of data points
a9a	123	32,561
ionosphere	34	351
mushrooms	112	8,124
phishing	68	11,055
sonar	60	208
splice	60	1,000
w8a	300	49,749

Implementation Details. Same as in D.1

Comparison. We evaluate Algorithm 1 and Algorithm 3 with the three criteria used in D.1 on the LIBSVM datasets. The boxplots for the criteria are shown in Figure 3. From Figure 3, we

observe that Algorithm 1 outperforms Algorithm 3 in terms of the KKT residual and number of objective and constraints evaluations, but Algorithm 3 has lower number of gradient and Jacobian evaluations than Algorithm 1 as we observed in D.1.

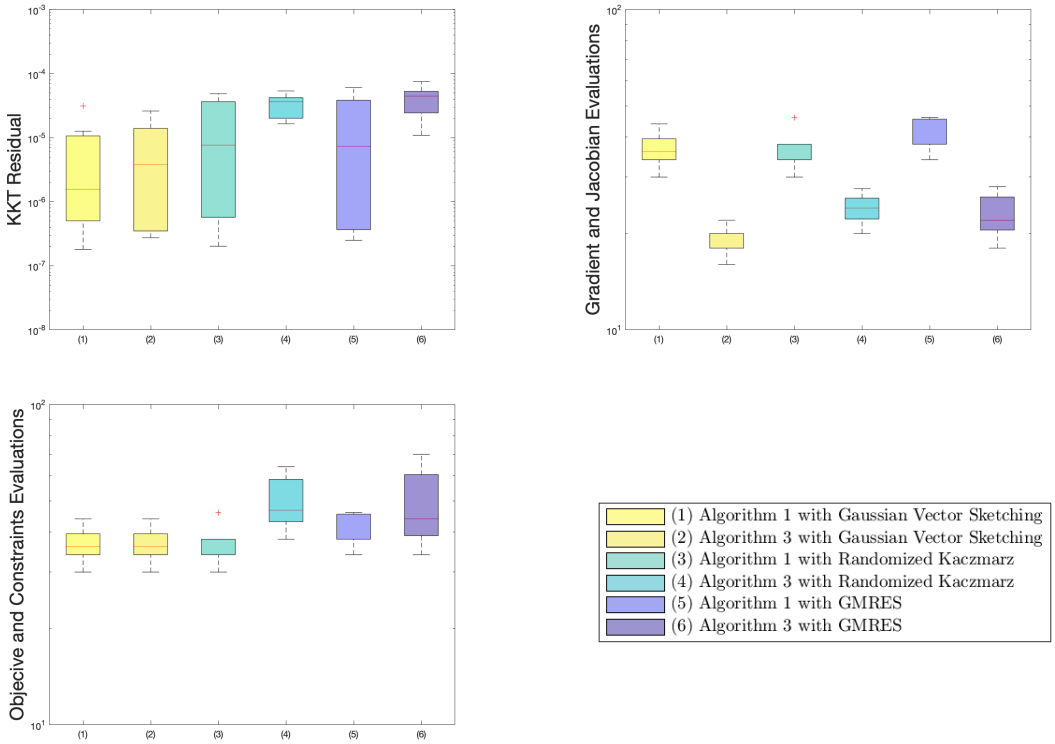

Figure 3: KKT residual, number of gradient and Jacobian evaluations, and number of objective and constraints evaluations boxplots for Algorithm 1 and Algorithm 3 on the LIBSVM datasets.