Mathematical Programming
https://doi.org/10.1007/s10107-023-01935-7

FULL LENGTH PAPER

Series A

Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming

Sen Na[1,2] · Mihai Anitescu[3] · Mladen Kolar[4]

Received: 23 February 2022 / Accepted: 21 January 2023

Abstract

We study nonlinear optimization problems with a stochastic objective and deterministic equality and inequality constraints, which emerge in numerous applications including finance, manufacturing, power systems and, recently, deep neural networks. We propose an active-set stochastic sequential quadratic programming (StoSQP) algorithm that utilizes a differentiable exact augmented Lagrangian as the merit function. The algorithm adaptively selects the penalty parameters of the augmented Lagrangian, and performs a stochastic line search to decide the stepsize. The global convergence is established: for any initialization, the KKT residuals converge to zero *almost surely*. Our algorithm and analysis further develop the prior work of Na et al. (Math Program, 2022. https://doi.org/10.1007/s10107-022-01846-z). Specifically, we allow nonlinear inequality constraints *without* requiring the strict complementary condition; refine some of designs in Na et al. (2022) such as the feasibility error condition and the monotonically increasing sample size; strengthen the global convergence guarantee; and improve the sample complexity on the objective Hessian. We demonstrate the performance of the designed algorithm on a subset of nonlinear problems collected in CUTEst test set and on constrained logistic regression problems.

Keywords Inequality constraints · Stochastic optimization · Exact augmented Lagrangian · Sequential quadratic programming

✉ Sen Na
senna@berkeley.edu

Mihai Anitescu
anitescu@mcs.anl.gov

Mladen Kolar
mladen.kolar@chicagobooth.edu

[1] Department of Statistics, University of California, Berkeley, Berkeley, USA

[2] International Computer Science Institute, Berkeley, USA

[3] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, USA

[4] Booth School of Business, The University of Chicago, Chicago, USA

Published online: 02 March 2023

Mathematics Subject Classification 90-08 · 90C15 · 90C26 · 90C30 · 90C55 · 90C90

1 Introduction

We study stochastic nonlinear optimization problems with deterministic equality and inequality constraints:

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

where $f:\mathbb{R}^d\rightarrow\mathbb{R}$ is an expected objective, $c:\mathbb{R}^d\rightarrow\mathbb{R}^m$ are deterministic equality constraints, $g:\mathbb{R}^d\rightarrow\mathbb{R}^r$ are deterministic inequality constraints, $\xi\sim\mathcal{P}$ is a random variable following the distribution $\mathcal{P}$, and $F(\cdot;\xi):\mathbb{R}^d\rightarrow\mathbb{R}$ is a realized objective. In stochastic optimization regime, the direct evaluation of f and its derivatives is not accessible. Instead, it is assumed that one can generate independent and identically distributed samples $\{\xi_i\}_i$ from $\mathcal{P}$, and estimate f and its derivatives based on the realizations $\{F(\cdot\,;\xi_i)\}_i$.

Problem (1) widely appears in a variety of industrial applications including finance, transportation, manufacturing, and power systems [8, 56]. It includes constrained empirical risk minimization (ERM) as a special case, where $\mathcal{P}$ can be regarded as a uniform distribution over n data points $\{\xi_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)\}_{i=1}^n$, with $(\boldsymbol{y}_i, \boldsymbol{z}_i)$ being the feature-outcome pairs. Thus, the objective has a finite-sum form as

$$f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{x};\xi_i) = \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{x};\boldsymbol{y}_i,\boldsymbol{z}_i).$$

The goal of (1) is to find the optimal parameter $\boldsymbol{x}^\star$ that fits the data best. One of the most common choices of F is the negative log-likelihood of the underlying distribution of $(\boldsymbol{y}_i, \boldsymbol{z}_i)$. In this case, the optimizer $\boldsymbol{x}^\star$ is called the maximum likelihood estimator (MLE). Constraints on parameters are also common in practice, which are used to encode prior model knowledge or to restrict model complexity. For example, [30, 31] studied inequality constrained least-squares problems, where inequality constraints maintain structural consistency such as non-negativity of the elasticities. [42, 45] studied statistical properties of constrained MLE, where constraints characterize the parameters space of interest. More recently, a growing literature on training constrained neural networks has been reported [15, 25, 32, 33], where constraints are imposed to avoid weights either vanishing or exploding, and objectives are in the above finite-sum form.

This paper aims to develop a numerical procedure to solve (1) with a global convergence guarantee. When the objective f is deterministic, numerous nonlinear optimization methods with well-understood convergence results are applicable, such as exact penalty methods, augmented Lagrangian methods, sequential quadratic programming (SQP) methods, and interior-point methods [41]. However, methods to solve constrained *stochastic* nonlinear problems with satisfactory convergence guarantees have been developed only recently. In particular, with only equality constraints, [4] designed a very first stochastic SQP (StoSQP) scheme using an ℓ_1-penalized merit function, and showed that for any initialization, the KKT residuals $\{R_t\}_t$ converge in two different regimes, determined by a prespecified deterministic stepsize-related sequence $\{\alpha_t\}_t$:

(a) (constant sequence) if $\alpha_t = \alpha$ for some small $\alpha > 0$, then $\sum_{i=0}^{t-1} \mathbb{E}[R_i^2]/t \leq \Upsilon/(\alpha t) + \Upsilon\alpha$ for some $\Upsilon > 0$;

(b) (decaying sequence) if α_t satisfies $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$, then $\liminf_{t\to\infty} \mathbb{E}[R_t^2] = 0$.

Both convergence regimes are well known for unconstrained stochastic problems where $R_t = \|\nabla f(\boldsymbol{x}_t)\|$ (see [12] for a recent review), while [4] generalized the results to equality constrained problems. Within the algorithm of [4], the authors designed a stepsize selection scheme (based on the prespecified deterministic sequence) to bring some sort of adaptivity into the algorithm. However, it turns out that the prespecified sequence, which can be aggressive or conservative, still highly affects the performance. To address the adaptivity issue, [40] proposed an alternative StoSQP, which exploits a differentiable exact augmented Lagrangian merit function, and enables a stochastic line search procedure to adaptively select the stepsize. Under a different setup (where the model is precisely estimated with high probability), [40] proved a different guarantee: for any initialization, $\liminf_{t\to\infty} R_t = 0$ *almost surely*. Subsequently, a series of extensions have been reported. [3] designed a StoSQP scheme to deal with rank-deficient constraints. [18] designed a StoSQP that exploits inexact Newton directions. [6] designed an accelerated StoSQP via variance reduction for finite-sum problems. [5] further developed [4] to achieve adaptive sampling. [17] established the worst-case iteration complexity of StoSQP, and [39] established the asymptotic local rate of StoSQP and performed statistical inference. In addition, [43] investigated a deterministic SQP where the objective and constraints are evaluated with noise. However, all aforementioned literature does not include inequality constraints.

Our paper develops this line of research by designing a StoSQP method that works with nonlinear inequality constraints. In order to do so, we have to overcome a number of intrinsic difficulties that arise in dealing with inequality constraints, which were already noted in classical nonlinear optimization literature [7, 41]. Our work is built upon [40], where we exploited an augmented Lagrangian merit function under the SQP framework. We enhance some of designs in [40] (e.g., the feasibility error condition, the increasing batch size, and the complexity of Hessian sampling; more on these later), and the analysis of this paper is more involved. To generalize [40], we address the following two subtleties.

(a) With inequalities, SQP subproblems are inequality constrained (nonconvex) quadratic programs (IQPs), which themselves are difficult to solve in most cases. Some SQP literature (e.g., [10]) supposes to apply a QP solver to solve IQPs exactly, however, a practical scheme should embed a finite number of inner loop iterations of active-set methods or interior-point methods into the main SQP loop, to solve IQPs approximately. Then, the inner loop may lead to an approximation error for search direction in each iteration, which complicates the analysis.

(b) When applied to deterministic objectives with inequalities, the SQP search direction is a descent direction of the augmented Lagrangian only in a neighborhood of a KKT point [50, Propositions 8.3, 8.4]. This is in contrast to equality constrained problems, where the descent property of the SQP direction holds globally, provided the penalty parameters of the augmented Lagrangian are suitably chosen. Such a difference is indeed brought by inequality constraints: to make the (active-set) SQP direction informative, the estimated active set has to be close to the optimal active set (see Lemma 3 for details). Thus, simply changing the merit function in [40] does not work for Problem (1).

The existing literature on inequality constrained SQP has addressed (a) and (b) via various tools for deterministic objectives, while we provide new insights into stochastic objectives. To resolve (a), we design an active-set StoSQP scheme, where given the current iterate, we first identify an active set which includes all inequality constraints that are likely to be equalities. We then obtain the search direction by solving a SQP subproblem, where we include all inequality constraints in the identified active set but regard them as equalities. In this case, the subproblem is an equality constrained QP (EQP), and can be solved exactly provided the matrix factorization is within the computational budget. To resolve (b), we provide a safeguarding direction to the scheme. In each step, we check if the SQP subproblem is solvable and generates a descent direction of the augmented Lagrangian merit function. If yes, we maintain the SQP direction as it typically enjoys a fast local rate; if no, we switch to the safeguarding direction (e.g., one gradient/Newton step of the augmented Lagrangian), along which the iterates still decrease the augmented Lagrangian although the convergence may not be as effective as that of SQP.

Furthermore, to design a scheme that adaptively selects the penalty parameters and stepsizes for Problem (1), additional challenges have to be resolved. In particular, we know that there are *unknown deterministic* thresholds for penalty parameters to ensure one-to-one correspondence between a stationary point of the merit function and a KKT point of Problem (1). However, due to the scheme stochasticity, the stabilized penalty parameters are random. We are unsure if the stabilized values are above (or below, depending on the context) the thresholds or not. Thus, we cannot directly conclude that the iterates converge to a KKT point, even if we ensure a sufficient decrease on the merit function in each step, and enforce the iterates to converge to one of its stationary points.

The above difficulty has been resolved for the ℓ_1-penalized merit function in [4], where the authors imposed a probability condition on the noise (satisfied by symmetric noise; see [4, Proposition 3.16]). [40] resolved this difficulty for the augmented Lagrangian merit function by modifying the SQP scheme when selecting the penalty parameters. In particular, [40] required the feasibility error to be bounded by the gradient magnitude of the augmented Lagrangian in *each* step, and generated monotonically increasing samples to estimate the gradient. Although that analysis does not require noise conditions, adjusting the penalty parameters to enforce the feasibility error condition may not be necessary for the iterates that are far from stationarity. Also, generating increasing samples is not satisfactory since the sample size should be adaptively chosen based on the iterates. In this paper, we refine the techniques of [40] and generalize them to inequality constraints. We weaken the feasibility error condition by using a (large) multiplier to rescale the augmented Lagrangian gradient, and more significantly, enforcing it *only when the magnitude of the rescaled augmented Lagrangian gradient is smaller than the estimated KKT residual*. In other words, the feasibility error condition is imposed only when we have a stronger evidence that the iterate is approaching to a stationary point than approaching to a KKT point. Such a relaxation matches the motivation of the feasibility error condition, i.e., bridging the gap between stationary points and KKT points. We also get rid of the increasing sample size requirement by adaptively controlling the absolute deviation of the augmented Lagrangian gradient *for the new iterates only* (i.e. the previous step is a *successful* step; see Sect. 3). Following [40], we perform a stochastic line search procedure. However, instead of using the same sample set to estimate the gradient ∇f and Hessian $\nabla^2 f$ as in [40], we sharpen the analysis and realize that the needed samples for $\nabla^2 f$ are significantly less than ∇f.

With all above extensions from [40], we finally prove that the KKT residual R_t satisfies $\lim_{t\rightarrow\infty} R_t = 0$ *almost surely* for any initialization. Such a result is stronger than [44, Theorem 4.10] for unconstrained problems and [40, Theorem 4] for equality constrained problems, which only showed the "liminf" type of convergence. Our result also differs from the (liminf) convergence of the expected KKT residual $\mathbb{E}[R_t^2]$ established in [3–6, 18] (under a different setup).

Related work

A number of methods have been proposed to optimize stochastic objectives without constraints, varying from first-order methods to second-order methods [12]. For all methods, adaptively choosing the stepsize is particularly important for practical deployment. A line of literature selects the stepsize by adaptively controlling the batch size and embedding natural (stochastic) line search into the schemes [11, 13, 20, 22, 29]. Although empirical experiments suggest the validity of stochastic line search, a rigorous analysis is missing. Until recently, researchers revisited unconstrained stochastic optimization via the lens of classical nonlinear optimization methods, and were able to show promising convergence guarantees. In particular, [1, 9, 16, 28, 57] studied stochastic trust-region methods, and [2, 14, 19, 44] studied stochastic line search methods. Moreover, [3–6, 18, 40] designed a variety of StoSQP schemes to solve equality constrained stochastic problems. Our paper contributes to this line of works by proposing an active-set StoSQP scheme to handle inequality constraints.

There are numerous methods for solving deterministic problems with nonlinear constraints, varying from exact penalty methods, augmented Lagrangian methods, interior-point methods, and sequential quadratic programming (SQP) methods [41]. Our paper is based on SQP, which is a very effective (or at least competitive) approach for small or large problems. When inequality constraints are present, SQP can be classified into IQP and EQP approaches. The former solves inequality constrained subproblems; the latter, to which our method belongs, solves equality constrained subproblems. A clear advantage of EQP over IQP is that the subproblems are less expensive to solve, especially when the quadratic matrix is indefinite. See [41, Chapter 18.2] for a comparison. Within SQP schemes, an exact penalty function is used as the merit function to monitor the progress of the iterates towards a KKT point.

The ℓ_1-penalized merit function, $f(\boldsymbol{x}) + \mu\,(\|c(\boldsymbol{x})\|_1 + \|\max\{g(\boldsymbol{x}), \mathbf{0}\}\|_1)$, is always a plausible choice because of its simplicity. However, a disadvantage of such non-differentiable merit functions is their impedance of fast local rates. A nontrivial local modification of SQP has to be employed to relieve such an issue [10]. As a resolution, multiple differentiable merit functions have been proposed [7]. We exploit an augmented Lagrangian merit function, which was first proposed for equality constrained problems by [46, 51], and then extended to inequality constrained problems by [47, 48]. [50] further improved this series of works by designing a new augmented Lagrangian, and established the exact property under weaker conditions. Although not crucial for that exact property analysis, [50] did not include equality constraints. In this paper, we enhance the augmented Lagrangian in [50] by containing both equality and inequality constraints; and study the case where the objective is stochastic. When inequality constraints are suppressed, our algorithm and analysis naturally reduce to [40] (with refinements). We should mention that differentiable merit functions are often more expensive to evaluate, and their benefits are mostly revealed for local rates (see [38, Figure 1] for a comparison between the augmented Lagrangian and ℓ_1 merit functions on an optimal control problem). Thus, with only established global analysis, we do not aim to claim the benefits of the augmented Lagrangian over the popular ℓ_1 merit function. On the other hand, the augmented Lagrangian is a very common alternative of non-differentiable penalty functions, which has been widely utilized for inequality constrained problems and achieved promising performance [52–55, 60]. Also, our global analysis is the first step towards understanding the local rate of StoSQP when differentiable merit functions are employed.

Structure of the paper

We introduce the exploited augmented Lagrangian merit function and active-set SQP subproblems in Sect. 2. We propose our StoSQP scheme and analyze it in Sect. 3. The experiments and conclusions are in Sects. 4 and 5. Due to the space limit, we defer all proofs to Appendix.

Notation We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and spectrum norm for matrices. For two scalars a and b, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ with the same dimension, $\min\{\boldsymbol{a}, \boldsymbol{b}\}$ and $\max\{\boldsymbol{a}, \boldsymbol{b}\}$ are vectors by taking entrywise minimum and maximum, respectively. For $\boldsymbol{a} \in \mathbb{R}^r$, $\text{diag}(\boldsymbol{a}) \in \mathbb{R}^{r\times r}$ is a diagonal matrix whose diagonal entries are specified by $\boldsymbol{a}$ sequentially. I denotes the identity matrix whose dimension is clear from the context. For a set $\mathcal{A} \subseteq \{1, 2, \ldots, r\}$ and a vector $\boldsymbol{a} \in \mathbb{R}^r$ (or a matrix $A \in \mathbb{R}^{r\times d}$), $\boldsymbol{a}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ (or $A_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|\times d}$) is a sub-vector (or a sub-matrix) including only the indices in $\mathcal{A}$; $\Pi_{\mathcal{A}}(\cdot) : \mathbb{R}^r \rightarrow \mathbb{R}^r$ (or $\mathbb{R}^{r\times d} \rightarrow \mathbb{R}^{r\times d}$) is a projection operator with $[\Pi_{\mathcal{A}}(\boldsymbol{a})]_i = \boldsymbol{a}_i$ if $i \in \mathcal{A}$ and $[\Pi_{\mathcal{A}}(\boldsymbol{a})]_i = 0$ if $i \notin \mathcal{A}$ (for $A \in \mathbb{R}^{r\times d}$, $\Pi_{\mathcal{A}}(A)$ is applied column-wise); $\mathcal{A}^c = \{1, 2, \ldots, r\}\backslash\mathcal{A}$. Finally, we reserve the notation for the Jacobian matrices of constraints: $J(\boldsymbol{x}) = \nabla^T c(\boldsymbol{x}) = (\nabla c_1(\boldsymbol{x}), \ldots, \nabla c_m(\boldsymbol{x}))^T \in \mathbb{R}^{m\times d}$ and $G(\boldsymbol{x}) = \nabla^T g(\boldsymbol{x}) = (\nabla g_1(\boldsymbol{x}), \ldots, \nabla g_r(\boldsymbol{x}))^T \in \mathbb{R}^{r\times d}$.

2 Preliminaries

Throughout this section, we suppose f, c, g are twice continuously differentiable (i.e., $f, g, c \in C^2$). The Lagrangian function of Problem (1) is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\mu}^T c(\boldsymbol{x}) + \boldsymbol{\lambda}^T g(\boldsymbol{x}).$$

We denote by

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^d : c(\boldsymbol{x}) = \boldsymbol{0}, g(\boldsymbol{x}) \leq \boldsymbol{0}\} \tag{2}$$

the feasible set and

$$\mathcal{I}(\boldsymbol{x}) = \{i : 1 \leq i \leq r, g_i(\boldsymbol{x}) = \boldsymbol{0}\} \tag{3}$$

the active set. We aim to find a KKT point $(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)$ of (1) satisfying

$$\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star) = \boldsymbol{0},\ c(\boldsymbol{x}^\star) = \boldsymbol{0},\ g(\boldsymbol{x}^\star) \leq \boldsymbol{0},\ \boldsymbol{\lambda}^\star \geq \boldsymbol{0},\ (\boldsymbol{\lambda}^\star)^T g(\boldsymbol{x}^\star) = 0. \tag{4}$$

When a constraint qualification holds, existing a dual pair $(\boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)$ to satisfy (4) is a first-order necessary condition for $\boldsymbol{x}^\star$ being a local solution of (1). In most cases, it is difficult to have an initial iterate that satisfies all inequality constraints, and enforce inequality constraints to hold as the iteration proceeds. This motivates us to consider a perturbed set. For $\nu > 0$, we let

$$\Omega \subsetneq \mathcal{T}_\nu := \left\{\boldsymbol{x} \in \mathbb{R}^d : a(\boldsymbol{x}) \leq \nu/2\right\} \quad \text{where} \ \ a(\boldsymbol{x}) = \sum_{i=1}^{r} \max\{g_i(\boldsymbol{x}), 0\}^3. \tag{5}$$

Here, the perturbation radius $\nu/2$ is not essential and can be replaced by ν/κ for any $\kappa > 1$. Also, the cubic power in $a(\boldsymbol{x})$ can be replaced by any power s with $s > 2$, which ensures that $a(\boldsymbol{x}) \in C^2$ provided $g_i(\boldsymbol{x}) \in C^2$, $\forall i$. We also define a scaling function

$$q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{a_\nu(\boldsymbol{x})}{1 + \|\boldsymbol{\lambda}\|^2} \quad \text{with} \ \ a_\nu(\boldsymbol{x}) = \nu - a(\boldsymbol{x}), \tag{6}$$

where $a_\nu(\boldsymbol{x})$ measures the distance of $a(\boldsymbol{x})$ to the boundary ν, and $q_\nu(\boldsymbol{x}, \boldsymbol{\lambda})$ rescales $a_\nu(\boldsymbol{x})$ by penalizing $\boldsymbol{\lambda}$ that has a large magnitude. In the definitions of (5) and (6), $\nu > 0$ is a parameter to be chosen: given the current primal iterate $\boldsymbol{x}_t$, we choose $\nu = \nu_t$ large enough so that $\boldsymbol{x}_t \in \mathcal{T}_\nu$. Note that while it is difficult to have $\boldsymbol{x}_t \in \Omega$, it is easy to choose ν to have $\boldsymbol{x}_t \in \mathcal{T}_\nu$. We also note that

$$\frac{\nu}{2(1 + \|\boldsymbol{\lambda}\|^2)} \leq q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \leq \nu \ \ \forall (\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathcal{T}_\nu \times \mathbb{R}^r, \text{ and } q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \to 0 \ \text{ as } \|\boldsymbol{\lambda}\| \to \infty.$$

With (6) and a parameter $\epsilon > 0$, we define a function to measure the dual feasibility of inequality constraints:

$$\begin{aligned}\boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda}) &:= g(\boldsymbol{x}) - \boldsymbol{b}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda}) \\ &:= g(\boldsymbol{x}) - \min\{\boldsymbol{0}, g(\boldsymbol{x}) + \epsilon q_\nu(\boldsymbol{x},\boldsymbol{\lambda})\boldsymbol{\lambda}\} = \max\{g(\boldsymbol{x}), -\epsilon q_\nu(\boldsymbol{x},\boldsymbol{\lambda})\boldsymbol{\lambda}\}. \quad (7)\end{aligned}$$

The following lemma justifies the reasonability of the definition (7). The proof is immediate and omitted.

Lemma 1 *Let* $\epsilon, \nu > 0$. *For any* $(\boldsymbol{x},\boldsymbol{\lambda}) \in \mathcal{T}_\nu \times \mathbb{R}^r$, $\boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda}) = \boldsymbol{0} \Leftrightarrow g(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{\lambda}^T g(\boldsymbol{x}) = 0$.

An implication of Lemma 1 is that, when the iteration sequence converges to a KKT point, $\boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})$ converges to 0, i.e., $g(\boldsymbol{x}) = \boldsymbol{b}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})$. This motivates us to define the following augmented Lagrangian function:

$$\begin{aligned}\mathcal{L}_{\epsilon,\nu,\eta}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\lambda}) &+ \frac{1}{2\epsilon}\|c(\boldsymbol{x})\|^2 \\ &+ \frac{1}{2\epsilon q_\nu(\boldsymbol{x},\boldsymbol{\lambda})}\left(\|g(\boldsymbol{x})\|^2 - \|\boldsymbol{b}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})\|^2\right) \\ &+ \frac{\eta}{2}\left\|\begin{pmatrix} J(\boldsymbol{x})\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\lambda}) \\ G(\boldsymbol{x})\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\lambda}) + \mathrm{diag}^2(g(\boldsymbol{x}))\boldsymbol{\lambda}\end{pmatrix}\right\|^2, \quad (8)\end{aligned}$$

where $\eta > 0$ is a prespecified parameter, which can be any positive number throughout the paper. The augmented Lagrangian (8) generalizes the one in [50] by including equality constraints and introducing η to enhance flexibility ($\eta = 2$ in [50]). Without inequalities, (8) reduces to the augmented Lagrangian studied in [40]. The penalty in (8) consists of two parts. The first part characterizes the feasibility error and consists of $\|c(\boldsymbol{x})\|^2$ and $\|g(\boldsymbol{x})\|^2 - \|\boldsymbol{b}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})\|^2$. The latter term is rescaled by $1/q_\nu(\boldsymbol{x},\boldsymbol{\lambda})$ to penalize $\boldsymbol{\lambda}$ with a large magnitude. In fact, if $\|\boldsymbol{\lambda}\| \rightarrow \infty$, then $q_\nu(\boldsymbol{x},\boldsymbol{\lambda})\boldsymbol{\lambda} \rightarrow \boldsymbol{0}$ so that $b_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda}) \rightarrow \min\{\boldsymbol{0}, g(\boldsymbol{x})\}$ (cf. (7)). Thus, the penalty term $(\|g(\boldsymbol{x})\|^2 - \|b_\epsilon(\boldsymbol{x},\boldsymbol{\lambda})\|^2)/q_\nu(\boldsymbol{x},\boldsymbol{\lambda}) \rightarrow \infty$, which is impossible when the iterates decrease $\mathcal{L}_{\epsilon,\nu,\eta}$. The second part characterizes the optimality error and does not depend on the parameters ϵ and ν. We mention that there are alternative forms of the augmented Lagrangian, some of which transform nonlinear inequalities using (squared) slack variables [7, 60]. In that case, additional variables are involved and the strict complementarity condition is often needed to ensure the equivalence between the original and transformed problems [23].

The exact property of (8) can be studied similarly as in [50], however this is incremental and not crucial for our analysis. We will only use (a stochastic version of) (8) to monitor the progress of the iterates. By direct calculation, we obtain the gradient $\nabla\mathcal{L}_{\epsilon,\nu,\eta}$. We first suppress the evaluation point for conciseness, and define the following matrices

$$Q_{11} = (\nabla_{\boldsymbol{x}}^2\mathcal{L})J^T, \quad Q_{12} = \sum_{i=1}^m (\nabla^2 c_i)(\nabla_{\boldsymbol{x}}\mathcal{L})\boldsymbol{e}_{i,m}^T, \quad Q_1 = Q_{11} + Q_{12} \in \mathbb{R}^{d\times m},$$

$$Q_{21} = (\nabla_{\boldsymbol{x}}^2 \mathcal{L}) G^T, \quad Q_{22} = \sum_{i=1}^{r} (\nabla^2 g_i)(\nabla_{\boldsymbol{x}} \mathcal{L}) \boldsymbol{e}_{i,r}^T, \quad Q_{23} = 2G^T \mathrm{diag}(g)\mathrm{diag}(\boldsymbol{\lambda}),$$

$$Q_2 = \sum_{i=1}^{3} Q_{2i} \in \mathbb{R}^{d\times r},$$

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} = \begin{pmatrix} JJ^T & JG^T \\ GJ^T & GG^T + \mathrm{diag}^2(g) \end{pmatrix} \in \mathbb{R}^{(m+r)\times(m+r)}, \tag{9}$$

where $\boldsymbol{e}_{i,m} \in \mathbb{R}^m$ is the i-th canonical basis of $\mathbb{R}^m$ (similar for $\boldsymbol{e}_{i,r} \in \mathbb{R}^r$). Then,

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\epsilon,\nu,\eta} \\ \nabla_{\boldsymbol{\mu}} \mathcal{L}_{\epsilon,\nu,\eta} \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\epsilon,\nu,\eta} \end{pmatrix} = \begin{pmatrix} I & \frac{1}{\epsilon} J^T & \frac{1}{\epsilon q_\nu} G^T \\ & I & \\ & & I \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L} \\ c \\ \boldsymbol{w}_{\epsilon,\nu} \end{pmatrix} + \begin{pmatrix} \frac{3\|\boldsymbol{w}_{\epsilon,\nu}\|^2}{2\epsilon q_\nu a_\nu} G^T \boldsymbol{l} \\ \boldsymbol{0} \\ \frac{\|\boldsymbol{w}_{\epsilon,\nu}\|^2}{\epsilon a_\nu} \boldsymbol{\lambda} \end{pmatrix}$$
$$+\eta \begin{pmatrix} Q_1 & Q_2 \\ M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} J \nabla_{\boldsymbol{x}} \mathcal{L} \\ G \nabla_{\boldsymbol{x}} \mathcal{L} + \mathrm{diag}^2(g)\boldsymbol{\lambda} \end{pmatrix}, \tag{10}$$

where $\boldsymbol{l} = \boldsymbol{l}(\boldsymbol{x}) = \mathrm{diag}(\max\{g(\boldsymbol{x}), \boldsymbol{0}\}) \max\{g(\boldsymbol{x}), \boldsymbol{0}\}$. Clearly, the evaluation of $\nabla \mathcal{L}_{\epsilon,\nu,\eta}$ requires ∇f and $\nabla^2 f$, which have to be replaced by their stochastic counterparts $\bar{\nabla} f$ and $\bar{\nabla}^2 f$ for Problem (1). Based on (10), we note that, if the feasibility error vanishes, then $\nabla \mathcal{L}_{\epsilon,\nu,\eta} = \boldsymbol{0}$ implies the KKT conditions (4) hold for any $\epsilon, \nu, \eta > 0$. We summarize this observation in the next lemma. The result holds without any constraint qualifications.

Lemma 2 *Let $\epsilon, \nu, \eta > 0$ and let $(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star) \in \mathcal{T}_\nu \times \mathbb{R}^m \times \mathbb{R}^r$ be a primal-dual triple. If $\|c(\boldsymbol{x}^\star)\| = \|\boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)\| = \|\nabla \mathcal{L}_{\epsilon,\nu,\eta}(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)\| = 0$, then $(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)$ satisfies (4) and, hence, is a KKT point of Problem* (1).

Proof See Appendix A.1 □

In the next subsection, we introduce an active-set SQP direction that is motivated by the augmented Lagrangian (8).

2.1 An active-set SQP direction via EQP

Let $\epsilon, \nu, \eta > 0$ be fixed parameters. Suppose we have the t-th iterate $(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t) \in \mathcal{T}_\nu \times \mathbb{R}^m \times \mathbb{R}^r$, let us denote $J_t = J(\boldsymbol{x}_t)$, $G_t = G(\boldsymbol{x}_t)$ (similar for $\nabla f_t, c_t, g_t, q_\nu^t$ etc.) to be the quantities evaluated at the t-th iterate. We generally use index t as subscript, except for the quantities (e.g., q_ν^t) that depend on ϵ, ν, or η, which have been used as subscript. For an active set $\mathcal{A} \subseteq \{1, \ldots, r\}$, we denote $\boldsymbol{\lambda}_{t_a} = (\boldsymbol{\lambda}_t)_{\mathcal{A}}$, $\boldsymbol{\lambda}_{t_c} = (\boldsymbol{\lambda}_t)_{\mathcal{A}^c}$ (similar for $g_{t_a}, g_{t_c}, G_{t_a}, G_{t_c}$ etc.) to be the sub-vectors (or sub-matrices), and denote $\Pi_a(\cdot) = \Pi_{\mathcal{A}}(\cdot)$, $\Pi_c(\cdot) = \Pi_{\mathcal{A}^c}(\cdot)$ for shorthand.

With the t-th iterate $(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$ and the above notation, we first define the identified active set as

$$\mathcal{A}_{\epsilon,\nu}^t := \mathcal{A}_{\epsilon,\nu}(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) := \{i : 1 \le i \le r, \ (g_t)_i \ge -\epsilon q_\nu^t \cdot (\boldsymbol{\lambda}_t)_i\}. \tag{11}$$

We then solve the following coupled linear system

$$\overbrace{\begin{pmatrix} B_t & J_t^T & G_{t_a}^T \\ J_t & & \\ G_{t_a} & & \end{pmatrix}}^{K_{t_a}} \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \tilde{\Delta}\boldsymbol{\mu}_t \\ \tilde{\Delta}\boldsymbol{\lambda}_{t_a} \end{pmatrix} = -\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t - G_{t_c}^T \boldsymbol{\lambda}_{t_c} \\ c_t \\ g_{t_a} \end{pmatrix}, \tag{12a}$$

$$\underbrace{\begin{pmatrix} J_t J_t^T & J_t G_t^T \\ G_t J_t^T & G_t G_t^T + \text{diag}^2(g_t) \end{pmatrix}}_{M_t} \begin{pmatrix} \Delta \boldsymbol{\mu}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix}$$

$$= -\left\{ \begin{pmatrix} J_t \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ G_t \nabla_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\text{diag}^2(g_t)\boldsymbol{\lambda}_t) \end{pmatrix} + \begin{pmatrix} Q_{1,t}^T \\ Q_{2,t}^T \end{pmatrix} \Delta \boldsymbol{x}_t \right\}, \tag{12b}$$

for some B_t that approximates the Hessian $\nabla_{\boldsymbol{x}}^2 \mathcal{L}_t$. Our active-set SQP direction is then $\Delta_t := (\Delta \boldsymbol{x}_t, \Delta \boldsymbol{\mu}_t, \Delta \boldsymbol{\lambda}_t)$. Finally, we update the iterate as

$$\begin{pmatrix} \boldsymbol{x}_{t+1} \\ \boldsymbol{\mu}_{t+1} \\ \boldsymbol{\lambda}_{t+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_t \\ \boldsymbol{\mu}_t \\ \boldsymbol{\lambda}_t \end{pmatrix} + \alpha_t \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\mu}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix}$$

with α_t chosen to ensure a certain sufficient decrease on the merit function (8).

The definition of active set was introduced in [50, (8.5)] and has been utilized, e.g., in [53]. Intuitively, for the i-th inequality constraint, if $g_i^\star = (g(\boldsymbol{x}^\star))_i = 0$ and $\boldsymbol{\lambda}_i^\star > 0$, then i will be identified when $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ is close to $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$; if $g_i^\star < 0$ and $\boldsymbol{\lambda}_i^\star = 0$, then i will not be identified. The stepsize α_t is usually chosen by line search. In Sect. 3, we will design a stochastic line search scheme to select α_t adaptively. Compared to fully stochastic SQP schemes [3, 4, 18], we need a more precise model estimation. We explain the SQP direction (12) in the next remark.

Remark 1 Our dual direction $(\Delta \boldsymbol{\mu}_t, \Delta \boldsymbol{\lambda}_t)$ differs from the usual SQP direction introduced, for example, in [50, (8.9)]. In particular, the system (12a) is nothing but the KKT conditions of EQP:

$$\begin{aligned} \min_{\Delta \boldsymbol{x}_t} \quad & \frac{1}{2}(\Delta \boldsymbol{x}_t)^T B_t \Delta \boldsymbol{x}_t + (\nabla f_t)^T \Delta \boldsymbol{x}_t, \\ \text{s.t.} \quad & c_t + J_t \Delta \boldsymbol{x}_t = \boldsymbol{0}, \\ & g_{t_a} + G_{t_a} \Delta \boldsymbol{x}_t = \boldsymbol{0}. \end{aligned}$$

Thus, $(\Delta \boldsymbol{x}_t, \boldsymbol{\mu}_t + \tilde{\Delta}\boldsymbol{\mu}_t, \boldsymbol{\lambda}_{t_a} + \tilde{\Delta}\boldsymbol{\lambda}_{t_a})$ solved from (12a) is also the primal-dual solution of the above EQP. However, instead of using $(\tilde{\Delta}\boldsymbol{\mu}_t, \tilde{\Delta}\boldsymbol{\lambda}_{t_a}, -\boldsymbol{\lambda}_{t_c})$, we solve the dual direction $(\Delta \boldsymbol{\mu}_t, \Delta \boldsymbol{\lambda}_t)$ for both active and inactive constraints from (12b). As B_t converges to $\nabla_{\boldsymbol{x}}^2 \mathcal{L}_t$ and $(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$ converges to a KKT point $(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)$, it is fairly easy to see that $(\Delta \boldsymbol{\mu}_t, \Delta \boldsymbol{\lambda}_t)$ converges to $(\tilde{\Delta}\boldsymbol{\mu}_t, \tilde{\Delta}\boldsymbol{\lambda}_t)$ (where we denote $\tilde{\Delta}\boldsymbol{\lambda}_{t_c} = -\boldsymbol{\lambda}_{t_c}$) in a higher order by noting that

$$\begin{pmatrix} J_tJ_t^T & J_tG_t^T \\ G_tJ_t^T & G_tG_t^T + \text{diag}^2(g_t) \end{pmatrix} \begin{pmatrix} \tilde{\Delta}\boldsymbol{\mu}_t \\ \tilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \overset{(12a)}{=} \begin{pmatrix} \mathbf{0} \\ \Pi_a(\text{diag}^2(g_t)\tilde{\Delta}\boldsymbol{\lambda}_t) \end{pmatrix} \\ - \left\{ \begin{pmatrix} J_t\nabla_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\nabla_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\text{diag}^2(g_t)\boldsymbol{\lambda}_t) \end{pmatrix} + \begin{pmatrix} J_t \\ G_t \end{pmatrix} B_t\Delta\boldsymbol{x}_t \right\}.$$

Thus, the fast *local* rate of the SQP direction $(\Delta\boldsymbol{x}_t, \tilde{\Delta}\boldsymbol{\mu}_t, \tilde{\Delta}\boldsymbol{\lambda}_t)$ is preserved by Δ_t. However, it turns out that the adjustment of Δ_t is crucial for the merit function (8) when B_t is far from $\nabla_{\boldsymbol{x}}^2\mathcal{L}_t$. A similar, coupled SQP system is employed for equality constrained problems [35, 40], while we extend to inequality constraints here. In fact, [50, Proposition 8.2] showed that $(\Delta\boldsymbol{x}_t, \tilde{\Delta}\boldsymbol{\mu}_t, \tilde{\Delta}\boldsymbol{\lambda}_t)$ is a descent direction of $\mathcal{L}_{\epsilon,\nu,\eta}^t$ if $(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$ is near a KKT point and $B_t = \nabla_{\boldsymbol{x}}^2\mathcal{L}_t$. However, $B_t = \nabla_{\boldsymbol{x}}^2\mathcal{L}_t$ (i.e., no Hessian modification) is restrictive even for a deterministic line search, and that descent result does not hold if $B_t \neq \nabla_{\boldsymbol{x}}^2\mathcal{L}_t$. In contrast, as shown in Lemma 3, Δ_t is a descent direction even if B_t is not close to $\nabla_{\boldsymbol{x}}^2\mathcal{L}_t$.

2.2 The descent property of Δ_t

In this subsection, we present a descent property of Δ_t. We focus on the term $(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^t)^T\Delta_t$. Different from SQP for equality constrained problems, Δ_t may not be a descent direction of $\mathcal{L}_{\epsilon,\nu,\eta}^t$ for some points even if ϵ is chosen small enough. To see it clearly, we suppress the iteration index, denote $g_a = g_{t_a}$ (similar for $\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_c$ etc.), and divide $\nabla\mathcal{L}_{\epsilon,\nu,\eta}$ (cf. (10)) into two terms: a dominating term that depends on $(g_a, \boldsymbol{\lambda}_c)$ *linearly*, and a higher-order term that depends on $(g_a, \boldsymbol{\lambda}_c)$ at least *quadratically*. In particular, we write $\nabla\mathcal{L}_{\epsilon,\nu,\eta} = \nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(1)} + \nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(2)}$ where

$$\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\epsilon,\nu,\eta}^{(1)} \\ \nabla_{\boldsymbol{\mu}}\mathcal{L}_{\epsilon,\nu,\eta}^{(1)} \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\epsilon,\nu,\eta}^{(1)} \end{pmatrix} = \begin{pmatrix} I & \frac{1}{\epsilon}J^T & \frac{1}{\epsilon q_\nu}G^T \\ & I & \\ & & I \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L} \\ c \\ \boldsymbol{w}_{\epsilon,\nu} \end{pmatrix} \\ + \eta \begin{pmatrix} Q_1 & Q_2 \\ M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix},$$

$$\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\epsilon,\nu,\eta}^{(2)} \\ \nabla_{\boldsymbol{\mu}}\mathcal{L}_{\epsilon,\nu,\eta}^{(2)} \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\epsilon,\nu,\eta}^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{3\|\boldsymbol{w}_{\epsilon,\nu}\|^2}{2\epsilon q_\nu a_\nu}G^T\boldsymbol{l} \\ \mathbf{0} \\ \frac{\|\boldsymbol{w}_{\epsilon,\nu}\|^2}{\epsilon a_\nu}\boldsymbol{\lambda} \end{pmatrix} + \eta \begin{pmatrix} Q_{2,a} \\ M_{12,a} \\ M_{22,a} \end{pmatrix} \text{diag}^2(g_a)\boldsymbol{\lambda}_a. \tag{13}$$

Loosely speaking (see Lemma 3 for a rigorous result), $(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(1)})^T\Delta$ provides a sufficient decrease provided the penalty parameters are suitably chosen, while $(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(2)})^T\Delta$ has no such guarantee in general. Since $\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(2)}$ depends on $(g_a, \boldsymbol{\lambda}_c)$ *quadratically*, to ensure $\nabla\mathcal{L}_{\epsilon,\nu,\eta}^T\Delta < 0$, we require $\|g_a\| \vee \|\boldsymbol{\lambda}_c\|$ to be small enough to let the linear term $(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(1)})^T\Delta$ dominate. This essentially requires the iterate to be close to a KKT point, since $\|g_a\| = \|\boldsymbol{\lambda}_c\| = 0$ at a KKT point. With this discussion in mind, if the iterate is far from a KKT point, Δ may not be a descent direction of $\mathcal{L}_{\epsilon,\nu,\eta}$. In fact, for an iterate that is far from a KKT point, the KKT matrix K_a (and its component G_a) is likely to be singular due to the imprecisely identified active set. Thus, Newton system (12) is not solvable at this iterate at all, let alone it generates a descent direction. Without inequalities, the quadratic term $\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(2)}$ disappears and our analysis reduces to the one in [40]. We realize that the existence of $\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(2)}$ results in a very different augmented Lagrangian to the one in [40]; and brings difficulties in designing a global algorithm to deal with inequality constraints.

We point out that requiring a local iterate is not an artifact of the proof technique. Such a requirement is imposed for different search directions in related literature. For example, [50] showed that the SQP direction obtained by either EQP or IQP is a descent direction of $\mathcal{L}_{\epsilon,\nu,\eta}$ in a *neighborhood* of a KKT point (cf. Propositions 8.2 and 8.4). That work also required $B_t = \nabla_{\boldsymbol{x}}^2\mathcal{L}_t$, which we relax by considering a coupled Newton system. Subsequently, [53, 55] studied truncated Newton directions, whose descent properties hold only *locally* as well (cf. [53, Proposition 3.7], [55, Proposition 10]).

Now, we introduce two assumptions and formalize the descent property.

Assumption 1 (LICQ) We assume at $\boldsymbol{x}^\star$ that $(J^T(\boldsymbol{x}^\star)\quad G_{\mathcal{I}(\boldsymbol{x}^\star)}^T(\boldsymbol{x}^\star))$ has full column rank, where $\mathcal{I}(\boldsymbol{x}^\star)$ is the active inequality set defined in (3).

Assumption 2 For $\boldsymbol{z} \in \{\boldsymbol{z} \in \mathbb{R}^d : J_t\boldsymbol{z} = \boldsymbol{0}, G_{t_a}\boldsymbol{z} = \boldsymbol{0}\}$, we have $\boldsymbol{z}^T B_t \boldsymbol{z} \geq \gamma_B\|\boldsymbol{z}\|^2$ and $\|B_t\| \leq \Upsilon_B$ for constants $\Upsilon_B \geq 1 \geq \gamma_B > 0$.

The above condition on B_t is standard in nonlinear optimization literature [7]. In fact, $B_t = I$ with $\gamma_B = \Upsilon_B = 1$ is sufficient for the analysis in this paper. The condition $\Upsilon_B \geq 1 \geq \gamma_B > 0$ (similar for other constants defined later) is inessential, which is only for simplifying the presentation. Without such a requirement, our analyses hold by replacing γ_B with $\gamma_B \wedge 1$ and Υ_B with $\Upsilon_B \vee 1$.

Lemma 3 *Let $\nu, \eta > 0$ and suppose Assumptions 1 and 2 hold. There exist a constant $\Upsilon > 0$ depending on Υ_B but not on (ν, η, γ_B), and a compact set $\mathcal{X}_{\epsilon,\nu} \times \mathcal{M} \times \Lambda_{\epsilon,\nu}$ around $(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)$ depending on (ϵ, ν) but not on η,[1] such that if $(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t) \in \mathcal{X}_{\epsilon,\nu} \times \mathcal{M} \times \Lambda_{\epsilon,\nu}$ with ϵ satisfying $\epsilon \leq \gamma_B^2(\gamma_B \wedge \eta)/\{(1 \vee \nu)\Upsilon\}$, then*

$$(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{t\ (1)})\Delta_t \leq -\frac{\gamma_B \wedge \eta}{2}\left\|\begin{pmatrix}\Delta\boldsymbol{x}_t\\ J_t\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\nabla_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2.$$

[1] Here, we mean $\mathcal{X}_{\epsilon,\nu}$ and $\Lambda_{\epsilon,\nu}$ only *directly depend* on ϵ, ν but not η, which are in contrast to neighborhoods $\mathcal{X}_{\epsilon,\nu,\eta}$ and $\Lambda_{\epsilon,\nu,\eta}$. However, since the threshold of ϵ, $\gamma_B^2(\gamma_B \wedge \eta)/\{(1 \vee \nu)\Upsilon\}$, is also determined by η, the final local neighborhoods $\mathcal{X}_{\epsilon,\nu}$ and $\Lambda_{\epsilon,\nu}$ with ϵ below the threshold also *indirectly depend* on η. Recall that η can be any positive constant throughout the paper.

Furthermore, there exists a compact subset $\mathcal{X}_{\epsilon,\nu,\eta}\times\mathcal{M}\times\Lambda_{\epsilon,\nu,\eta}\subseteq\mathcal{X}_{\epsilon,\nu}\times\mathcal{M}\times\Lambda_{\epsilon,\nu}$ *depending additionally on* η*, such that if* $(\boldsymbol{x}_t,\boldsymbol{\mu}_t,\boldsymbol{\lambda}_t)\in\mathcal{X}_{\epsilon,\nu,\eta}\times\mathcal{M}\times\Lambda_{\epsilon,\nu,\eta}$*, then*

$$(\nabla\mathcal{L}^{t\,(2)}_{\epsilon,\nu,\eta})\Delta_t\le\frac{\gamma_B\wedge\eta}{4}\left\|\begin{pmatrix}\Delta\boldsymbol{x}_t\\ J_t\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\nabla_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2.$$

Proof See Appendix A.2 □

Similar arguments for other directions can be found in [53, Proposition 3.5] and [55, Proposition 9]. By the proof of Lemma 3, we know that as long as M_t and $(J_t^T\;\;G_{t_a}^T)$ in the SQP system (12) have full (column) rank, $(\nabla\mathcal{L}^{t\,(1)}_{\epsilon,\nu,\eta})^T\Delta_t$ ensures a sufficient decrease provided ϵ is small enough. However, from (A.11) in the proof, we also see that $(\nabla\mathcal{L}^{t\,(2)}_{\epsilon,\nu,\eta})^T\Delta_t$ is only bounded by

$$(\nabla\mathcal{L}^{t\,(2)}_{\epsilon,\nu,\eta})\Delta_t\le\Upsilon'\left(\frac{1\vee\nu}{\epsilon(1\wedge\nu^2)}\vee\eta\right)(\|g_{t_a}\|+\|\boldsymbol{\lambda}_{t_c}\|)\left\|\begin{pmatrix}\Delta\boldsymbol{x}_t\\ J_t\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\nabla_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2,$$

where $\Upsilon'>0$ is a constant independent of (ϵ,ν,η). Thus, to ensure $(\nabla\mathcal{L}^t_{\epsilon,\nu,\eta})^T\Delta_t$ to be negative, we have to restrict to a neighborhood, in which $\|g_{t_a}\|\vee\|\boldsymbol{\lambda}_{t_c}\|$ is small enough so that $\Upsilon'(\frac{1\vee\nu}{\epsilon(1\wedge\nu^2)}\vee\eta)(\|g_{t_a}\|+\|\boldsymbol{\lambda}_{t_c}\|)\le(\gamma_B\wedge\eta)/4$. This requirement is achievable near a KKT pair $(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)$, where the active set is correctly identified (implying that $\|g_{t_a}\|\le\|(g_t)_{\mathcal{I}(\boldsymbol{x}^\star)}\|$ and $\|\boldsymbol{\lambda}_{t_c}\|\le\|(\boldsymbol{\lambda}_t)_{\{i:1\le i\le r,\lambda_i^\star=0\}}\|$); and the radius of the neighborhood clearly depends on (ϵ,ν,η).

In the next section, we exploit the introduced augmented Lagrangian merit function (8) and the active-set SQP direction (12) to design a StoSQP scheme for Problem (1). We will adaptively choose proper ϵ and ν (recall that $\eta>0$ can be any positive number in this paper), incorporate stochastic line search to select the stepsize, and globalize the scheme by utilizing a safeguarding direction (e.g., Newton or steepest descent step) of the merit function $\mathcal{L}_{\epsilon,\nu,\eta}$. If the system (12) is not solvable, or is solvable but does not generate a descent direction, we search along the alternative direction to decrease the merit function. However, since Δ_t usually enjoys a fast local rate (see [50, Proposition 8.3] for a local analysis of $(\Delta\boldsymbol{x}_t,\tilde{\Delta}\boldsymbol{\mu}_t,\tilde{\Delta}\boldsymbol{\lambda}_t)$ and Remark 1), we prefer to preserve Δ_t as much as possible.

3 An adaptive active-set StoSQP scheme

We design an adaptive scheme for Problem (1) that embeds stochastic line search, originally designed and analyzed for unconstrained problems in [14, 44], into an active-set StoSQP. There are two challenges to design adaptive schemes for constrained problems. First, the merit function has penalty parameters that are random and adaptively specified; while for unconstrained problems one simply uses the objective function in line search. To show the global convergence, it is crucial that the stochastic penalty parameters are stabilized *almost surely*. Thus, for each run, after few iterations we always target a stabilized merit function. Otherwise, if each iteration decreases a different merit function, the decreases across iterations may not accumulate. Second, since the stabilized parameters are random, they may not be below *unknown deterministic* thresholds. Such a condition is critical to ensure the equivalence between the stationary points of the merit function and the KKT points of Problem (1). Thus, even if we converge to a stationary point of the (stabilized) merit function, it is not necessarily true that the stationary point is a KKT point of Problem (1).

With only equality constraints, [4, 40] addressed the first challenge under a boundedness condition, and our paper follows the same type of analysis. Similar boundedness condition is also required for deterministic analyses to have the penalty parameters stabilized [7, Chapter 4.3.3]. [4] resolved the second challenge by introducing a noise condition (satisfied by symmetric noise), while [40] resolved it by adjusting the SQP scheme when selecting the penalty parameters. As introduced in Sect. 1, the technique of [40] has multiple flaws: (i) it requires generating increasing samples to estimate the gradient of the augmented Lagrangian (cf. [40, Step 1]); (ii) it imposes a feasibility error condition for each step (cf. [40, (19)]). In this paper, we refine the technique of [40] and enable inequality constraints. As revealed by Sect. 2, the present analysis of inequality constraints is much more involved; and more importantly, our "lim" convergence guarantee strengthens the existing "liminf" convergence of the stochastic line search in [40, 44]. In what follows, we use $\bar{(\cdot)}$ to denote random quantities, except for the iterate $(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$. For example, $\bar{\alpha}_t$ denotes a random stepsize.

3.1 The proposed scheme

Let $\eta, \alpha_{max}, \kappa_{grad}, \chi_{grad}, \chi_f, \chi_{err} > 0$; $\rho > 1$; $\gamma_B \in (0, 1]$; $\beta, p_{grad}, p_f \in (0, 1)$; $\kappa_f \in (0, \beta/(4\alpha_{max})]$ be fixed tuning parameters. Given quantities $(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, \bar{\nu}_t, \bar{\epsilon}_t, \bar{\alpha}_t, \bar{\delta}_t)$ at the t-th iteration with $\boldsymbol{x}_t \in \mathcal{T}_{\bar{\nu}_t}$, we perform the following five steps to derive quantities at the $(t+1)$-th iteration.

Step 1: Estimate objective derivatives

We generate a batch of independent samples ξ_1^t to estimate the gradient ∇f_t and Hessian $\nabla^2 f_t$. The estimators $\bar{\nabla} f_t$ and $\bar{\nabla}^2 f_t$ may not be computed with the same amount of samples, since they have different sample complexities. For example, we can compute $\bar{\nabla} f_t$ using ξ_1^t while compute $\bar{\nabla}^2 f_t$ using a fraction of ξ_1^t (more on this in Sect. 3.4). With $\bar{\nabla} f_t$, $\bar{\nabla}^2 f_t$, we then compute $\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t$, $\bar{Q}_{1,t}$, and $\bar{Q}_{2,t}$ used in the system (12).

We require the batch size $|\xi_1^t|$ to be large enough to make the gradient error of the merit function small. In particular, we define

$$\bar{\Delta}(\nabla \mathcal{L}_\eta^t) := \|\bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t - \nabla \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t\|.$$

A simple observation from (10) is that $\bar{\Delta}(\nabla \mathcal{L}_\eta^t)$ is independent of $\bar{\epsilon}_t$ (and $\bar{\nu}_t$), which will be selected later (Step 2). We require $|\xi_1^t|$ to satisfy two conditions:

(a) the event $\mathcal{E}_1^t$,

$$\mathcal{E}_1^t = \left\{ \bar{\Delta}(\nabla \mathcal{L}_\eta^t) \le \kappa_{grad} \bar{\alpha}_t \underbrace{\left\| \begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t \\ c_t \\ \max\{g_t, -\boldsymbol{\lambda}_t\} \end{pmatrix} \right\|}_{\bar{R}_t} \right\}, \tag{14}$$

satisfies

$$P_{\xi_1^t}\left(\mathcal{E}_1^t\right) \ge 1 - p_{grad}; \tag{15}$$

(b) if $t-1$ is a *successful* step (see Step 5 for the meaning), then

$$\mathbb{E}_{\xi_1^t}[\bar{\Delta}(\nabla \mathcal{L}_\eta^t)] \le \chi_{grad} \cdot (\bar{\delta}_t / \bar{\alpha}_t)^{1/2}. \tag{16}$$

The sample complexities to ensure (15) and (16) will be discussed in Sect. 3.4. Compared to [40], we do not let $|\xi_1^t|$ increase monotonically, while we impose an expectation condition (16) when we arrive at a new iterate. By our analysis, it is easy to see that (16) can also be replaced by requiring the *subsequence* $\{|\xi_1^t| : t - 1 \text{ is a successful step}\}$ to increase to the infinity (e.g., increase by at least one each time), which is still weaker than [40]. The right hand side of (16) will be clear when we utilize $\bar{\delta}_t$ later in Step 5 (cf. (27)). We use $P_{\xi_1^t}(\cdot)$ and $\mathbb{E}_{\xi_1^t}[\cdot]$ to denote the probability and expectation that are evaluated over the randomness of sampling ξ_1^t only, while other random quantities are conditioned on, such as $(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$ and $\bar{\alpha}_t$. More precisely, we mean $P_{\xi_1^t}(\mathcal{E}_1^t) = P(\mathcal{E}_1^t \mid \mathcal{F}_{t-1})$ (similar for $\mathbb{E}_{\xi_1^t}[\cdot]$) where the σ-algebra $\mathcal{F}_{t-1}$ is defined in (28) below.

Step 2: Set parameter $\bar{\epsilon}_t$. With current $\bar{\nu}_t$, we decrease $\bar{\epsilon}_t \leftarrow \bar{\epsilon}_t/\rho$ until $\bar{\epsilon}_t$ is small enough to satisfy the following two conditions simultaneously:
(a) the feasibility error is proportionally bounded by the gradient of the merit function, whenever the iterate is closer to a stationary point than a KKT point:

$$\left\| (c_t, \boldsymbol{w}_{\bar{\epsilon}_t, \bar{\nu}_t}^t) \right\| \le \chi_{err} \| \bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t \| \qquad \text{if } \chi_{err} \| \bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t \| \le \bar{R}_t; \tag{17}$$

(we use the same multiplier χ_{err} only for simplifying the notation.)
(b) if the SQP system (12) with $\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t$, $\bar{Q}_{1,t}$, and $\bar{Q}_{2,t}$ is solvable, then we obtain $\bar{\Delta}_t = (\bar{\Delta} \boldsymbol{x}_t, \bar{\Delta} \boldsymbol{\mu}_t, \bar{\Delta} \boldsymbol{\lambda}_t)$ and require

$$(\bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^{t\ (1)})^T \bar{\Delta}_t \le -\frac{(\gamma_B \wedge \eta)}{2} \left\| \begin{pmatrix} \bar{\Delta} \boldsymbol{x}_t \\ J_t \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t \\ G_t \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t + \Pi_c(\operatorname{diag}^2(g_t) \boldsymbol{\lambda}_t) \end{pmatrix} \right\|^2. \tag{18}$$

We prove in Lemma 4 and Lemma 5 that both (17) and (18) can be satisfied for sufficiently small $\bar{\epsilon}_t$. In fact, Lemma 3 has already established (18) for the deterministic case. Even though $\bar{\Delta}_t$ is not always used as the search direction, we still enforce (18) to hold for $(\bar{\nabla}\mathcal{L}^{t\,(1)}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\bar{\Delta}_t$. The reason for this is to avoid ruling out $\bar{\Delta}_t$ just because $\bar{\epsilon}_t$ is not small enough, which would result in a positive dominating term $(\bar{\nabla}\mathcal{L}^{t\,(1)}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\bar{\Delta}_t$. If (12) is not solvable (e.g., the active set is imprecisely identified so that K_{t_a} is singular), then (18) is not needed.

The condition (17) is the key to ensure that the stationary point of the merit function that we converge to is a KKT point of (1). Motivated by Lemma 2, we know that "the stationarity of the merit function plus vanishing feasibility error" implies vanishing KKT residual. (17) states that the feasibility error is roughly controlled by the gradient of the merit function. (17) relaxes [40, (19)] from two aspects. First, [40] had no multiplier while we allow any (large) multiplier χ_{err}. Second, [40] enforced (17) for each step, while we enforce it only when we observe a stronger evidence that the scheme is approaching to a stationary point than to a KKT point. The above relaxations are driven by the intention of imposing the condition. When adjusting $\bar{\epsilon}_t$, if $\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\|$ first exceeds $\bar{R}_t$ before $\|(c_t, \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t})\|$ (which easily happens for a large $\bar{\nu}_t$), then one can immediately stop the adjustment of $\bar{\epsilon}_t$. Compared to [40] where the SQP system is supposed to be always solvable, (17) has extra usefulness: when $\bar{\Delta}_t$ is not available, (17) ensures that the safeguarding direction can be computed using the samples in Step 1. Such a desire is not easily achieved, and further relaxations of (17) can be designed if we generate new samples for the safeguarding direction (in Step 3). The subtlety lies in the fact that no penalty parameters are involved when we generate ξ^t_1 in Step 1, while (17) builds a connection between ξ^t_1 and the penalty parameters. It implies that the set ξ^t_1 satisfying (15) and (16) also satisfies the corresponding conditions for the safeguarding direction.

Step 3: Decide the search direction.

We may obtain a stochastic SQP direction $\bar{\Delta}_t$ from Step 2. However, if (12) is not solvable, or it is solvable but $\bar{\Delta}_t$ is not a sufficient descent direction because

$$(\bar{\nabla}\mathcal{L}^{t\,(2)}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\bar{\Delta}_t > \frac{(\gamma_B\wedge\eta)}{4}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2, \tag{19}$$

then an alternative safeguarding direction $\hat{\Delta}_t$ must be employed to ensure the decrease of the merit function. In that case, we follow [53, 55] and regard $\mathcal{L}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$ as a penalized objective. We require $\hat{\Delta}_t$ to satisfy

$$(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\hat{\Delta}_t \le -1/\chi_u\cdot\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\|^2 \quad\text{and}\quad \|\hat{\Delta}_t\|\le\chi_u\cdot\left\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right\| \tag{20}$$

for a constant $\chi_u\ge 1$. Similar to (17), we use the same constant χ_u for the two multipliers to simplify the notation. When using two different constants $\chi_{1,u}$ and $\chi_{2,u}$, we can always set $\chi_u = 1/\chi_{1,u}\vee\chi_{2,u}$ to let (20) hold. The condition (20) is standard in the literature [53, (60a,b)] [55, (52a,b)]. One example that satisfies (20) and is computationally cheap is the steepest descent direction $\hat{\Delta}_t = -\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$ with $\chi_u = 1$. Such a direction can be computed (almost) without any extra cost since the two components of $\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$, $\bar{\nabla}\mathcal{L}^{t\,(1)}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$ and $\bar{\nabla}\mathcal{L}^{t\,(2)}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$, have been computed when checking (18) and (19). Another example that is more computationally expensive is the regularized Newton step $\hat{H}_t \hat{\Delta}_t = -\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}$, where $\hat{H}_t$ captures second-order information of $\mathcal{L}^t_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}$ and satisfies $1/\chi_u I \preceq \hat{H}_t \preceq \chi_u I$. In particular, $\hat{H}_t$ can be obtained by regularizing the (generalized) Hessian matrix H_t, which is provided and discussed in [50, 53], and has the form[2]

$$\begin{aligned} H_{t,\boldsymbol{x}\boldsymbol{x}} &= B_t + \eta B_t \left\{ J_t^T J_t + G_t^T G_t \right\} B_t + \frac{1}{\bar{\epsilon}_t} J_t^T J_t + \frac{1}{\bar{\epsilon}_t q^t_{\bar{\nu}_t}} G_{t_a}^T G_{t_a}, \\ H_{t,(\boldsymbol{\mu},\boldsymbol{\lambda})\boldsymbol{x}} &= \begin{pmatrix} J_t \\ \Pi_a(G_t) \end{pmatrix} + \eta \begin{pmatrix} J_t J_t^T & J_t G_t^T \\ G_t J_t^T & G_t G_t^T + \mathrm{diag}^2(\Pi_c(g_t)) \end{pmatrix} \begin{pmatrix} J_t \\ G_t \end{pmatrix} B_t, \\ H_{t,(\boldsymbol{\mu},\boldsymbol{\lambda})(\boldsymbol{\mu},\boldsymbol{\lambda})} &= \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\bar{\epsilon}_t q^t_{\bar{\nu}_t} \mathrm{diag}(\Pi_c(\mathbf{1})) \end{pmatrix} + \eta \begin{pmatrix} J_t J_t^T & J_t G_t^T \\ G_t J_t^T & G_t G_t^T + \mathrm{diag}^2(\Pi_c(g_t)) \end{pmatrix}^2 . \end{aligned} \tag{21}$$

Here, $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^r$ is the all one vector. Other examples that improve upon the regularized Newton step include the choices in [21, 54], where a truncated conjugate gradient method is applied to an *indefinite* Newton system [54, Proposition 3.3, (14)]. We will numerically implement the regularized Newton and the steepest descent steps in Sect. 4.

Step 4: Estimate the merit function. Let $\check{\Delta}_t$ denote the adopted search direction; thus $\check{\Delta}_t = \bar{\Delta}_t$ from Step 2 or $\check{\Delta}_t = \hat{\Delta}_t$ from Step 3. We aim to perform stochastic line search by checking the Armijo condition (26) at the trial point

$$\boldsymbol{x}_{s_t} = \boldsymbol{x}_t + \bar{\alpha}_t \check{\Delta}\boldsymbol{x}_t, \qquad \boldsymbol{\mu}_{s_t} = \boldsymbol{\mu}_t + \bar{\alpha}_t \check{\Delta}\boldsymbol{\mu}_t, \qquad \boldsymbol{\lambda}_{s_t} = \boldsymbol{\lambda}_t + \bar{\alpha}_t \check{\Delta}\boldsymbol{\lambda}_t.$$

We estimate the merit function in this step and perform line search in Step 5.

First, we check if the trial primal point $\boldsymbol{x}_{s_t}$ is in $\mathcal{T}_{\bar{\nu}_t}$. In particular, if $\boldsymbol{x}_{s_t} \notin \mathcal{T}_{\bar{\nu}_t}$, that is $a_{s_t} = a(\boldsymbol{x}_{s_t}) > \bar{\nu}_t/2$ (cf. (5)), then we stop the current iteration and reject the trial point by letting $(\boldsymbol{x}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$, $\bar{\epsilon}_{t+1} = \bar{\epsilon}_t$, $\bar{\alpha}_{t+1} = \bar{\alpha}_t$, and $\bar{\delta}_{t+1} = \bar{\delta}_t$. We also increase $\bar{\nu}_t$ by letting

$$\bar{\nu}_{t+1} = \rho^j \bar{\nu}_t \quad \text{with} \quad j = \lceil \log(2a_{s_t}/\bar{\nu}_t)/\log \rho \rceil, \tag{22}$$

where $\lceil y \rceil$ denotes the least integer that exceeds y. The definition of $j \geq 1$ in (22) ensures $\boldsymbol{x}_{s_t} \in \mathcal{T}_{\bar{\nu}_{t+1}}$. However, $j = 1$ works as well, since $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t \in \mathcal{T}_{\bar{\nu}_t} \subseteq \mathcal{T}_{\bar{\nu}_{t+1}}$, as required for performing the next iteration. In the case of $\boldsymbol{x}_{s_t} \notin \mathcal{T}_{\bar{\nu}_t}$, particularly if $a_{s_t} \geq \bar{\nu}_t$, evaluating the merit function $\mathcal{L}^{s_t}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}$ is not informative since the penalty term in $\mathcal{L}^{s_t}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}$ may be rescaled by a negative multiplier. Thus, we increase $\bar{\nu}_t$ and rerun the iteration at the current point.

Otherwise $\boldsymbol{x}_{s_t} \in \mathcal{T}_{\bar{\nu}_t}$, then we generate a batch of independent samples ξ^t_2, that are independent from ξ^t_1 as well, and estimate $f_t, f_{s_t}, \nabla f_t, \nabla f_{s_t}$. Similar to Step 1,

[2] See (6.1)–(6.3) in [50] for a similar expression to (21). Our H_t generalizes that definition by including equality constraints and approximating the Hessian $\nabla^2_{\boldsymbol{x}}\mathcal{L}_t$ by B_t.

the estimators $\bar{f}_t$, $\bar{f}_{s_t}$ and $\bar{\bar{\nabla}} f_t$, $\bar{\bar{\nabla}} f_{s_t}$ may not be computed with the same amount of samples. For example, $\bar{f}_t$ and $\bar{f}_{s_t}$ can be computed using ξ_2^t while $\bar{\bar{\nabla}} f_t$ and $\bar{\bar{\nabla}} f_{s_t}$ can be computed using a fraction of ξ_2^t. The sample complexities are discussed in Sect. 3.4. Here, we distinguish $\bar{\bar{\nabla}} f_t$ from $\bar{\nabla} f_t$ in Step 1. While both of them are estimates of ∇f_t, the former is computed based on ξ_2^t and the latter is computed based on ξ_1^t. Using $\bar{f}_t$, $\bar{f}_{s_t}$, $\bar{\bar{\nabla}} f_t$, $\bar{\bar{\nabla}} f_{s_t}$, we compute $\bar{\mathcal{L}}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$ and $\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$ according to (8).

We require $|\xi_2^t|$ is large enough such that the event $\mathcal{E}_2^t$,

$$\mathcal{E}_2^t = \left\{\left|\bar{\mathcal{L}}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} - \mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right| \vee \left|\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right| \le -\kappa_f \bar{\alpha}_t^2 (\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T \check{\Delta}_t\right\}, \tag{23}$$

satisfies

$$P_{\xi_2^t}\left(\mathcal{E}_2^t\right) \ge 1 - p_f \tag{24}$$

and

$$\mathbb{E}_{\xi_2^t}[|\bar{\mathcal{L}}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} - \mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}|^2] \vee \mathbb{E}_{\xi_2^t}[|\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}|^2] \le \chi_f \cdot \bar{\delta}_t^2. \tag{25}$$

Similar to (15) and (16), $P_{\xi_2^t}(\cdot)$ and $\mathbb{E}_{\xi_2^t}[\cdot]$ denote that the randomness is taken over sampling ξ_2^t only, while other random quantities are conditioned on. That is, $P_{\xi_2^t}(\mathcal{E}_2^t) = P(\mathcal{E}_2^t \mid \mathcal{F}_{t-0.5})$ (similar for $\mathbb{E}_{\xi_2^t}[\cdot]$) where the σ-algebra $\mathcal{F}_{t-0.5} = \mathcal{F}_{t-1} \cup \sigma(\xi_1^t)$ is defined in (28) below.

Step 5: Perform line search. With the merit function estimates, we check the Armijo condition next.
(a) If the Armijo condition holds,

$$\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} \le \bar{\mathcal{L}}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} + \beta\bar{\alpha}_t(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\check{\Delta}_t, \tag{26}$$

then the trial point is accepted by letting $(\boldsymbol{x}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_{s_t}, \boldsymbol{\mu}_{s_t}, \boldsymbol{\lambda}_{s_t})$ and the stepsize is increased by $\bar{\alpha}_{t+1} = \rho\bar{\alpha}_t \wedge \alpha_{max}$. Furthermore, we check if the decrease of the merit function is reliable. In particular, if

$$-\beta\bar{\alpha}_t(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\check{\Delta}_t \ge \bar{\delta}_t, \tag{27}$$

then we increase $\bar{\delta}_t$ by $\bar{\delta}_{t+1} = \rho\bar{\delta}_t$; otherwise, we decrease $\bar{\delta}_t$ by $\bar{\delta}_{t+1} = \bar{\delta}_t/\rho$.
(b) If the Armijo condition (26) does not hold, then the trial point is rejected by letting $(\boldsymbol{x}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$, $\bar{\alpha}_{t+1} = \bar{\alpha}_t/\rho$ and $\bar{\delta}_{t+1} = \bar{\delta}_t/\rho$.

Finally, for both cases **(a)** and **(b)**, we let $\bar{\epsilon}_{t+1} = \bar{\epsilon}_t$, $\bar{\nu}_{t+1} = \bar{\nu}_t$ and repeat the procedure from Step 1. From (27), we can see that $\bar{\delta}_t$ (roughly) has the order $\bar{\alpha}_t\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\|^2$, which justifies the definition of the right hand side of (16).

The proposed scheme is summarized in Algorithm 1. We define three types of iterations for line search. If the Armijo condition (26) holds, we call the iteration a ***successful step***, otherwise we call it an ***unsuccessful step***. For a successful step, if the sufficient decrease in (27) is satisfied, we call it a ***reliable step***, otherwise we call it an ***unreliable step***. Same notion is used in [14, 40, 44].

To end this section, let us introduce the filtration induced by the randomness of the algorithm. Given a random sample sequence $\{\xi_1^t, \xi_2^t\}_{t=0}^{\infty}$,[3] we let $\mathcal{F}_t = \sigma(\{\xi_1^j, \xi_2^j\}_{j=0}^t)$, $t \geq 0$, be the σ-algebra generated by all the samples till t; $\mathcal{F}_{t-0.5} = \sigma(\{\xi_1^j, \xi_2^j\}_{j=0}^{t-1} \cup \xi_1^t)$, $t \geq 0$, be the σ-algebra generated by all the samples till $t-1$ and the sample ξ_1^t; and $\mathcal{F}_{-1}$ be the trivial σ-algebra generated by the initial iterate (which is deterministic). Throughout the presentation, we let $\bar{\epsilon}_t$ be the quantity obtained after Step 2; that is, $\bar{\epsilon}_t$ satisfies (17) and (18). With this setup, it is easy to see that

$$\begin{aligned} \sigma(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t) \cup \sigma(\bar{\nu}_t) \cup \sigma(\bar{\alpha}_t) \cup \sigma(\bar{\delta}_t) &\subseteq \mathcal{F}_{t-1}, \\ \sigma(\boldsymbol{x}_{s_t}, \boldsymbol{\mu}_{s_t}, \boldsymbol{\lambda}_{s_t}) \cup \sigma(\bar{\Delta}_t, \hat{\Delta}_t, \check{\Delta}_t) \cup \sigma(\bar{\epsilon}_t) &\subseteq \mathcal{F}_{t-0.5}. \end{aligned} \tag{28}$$

We analyze Algorithm 1 in the next subsection.

3.2 Assumptions and stability of parameters

We study the stability of the parameter sequence $\{\bar{\epsilon}_t, \bar{\nu}_t\}_t$. We will show that, for each run of the algorithm, the sequence is stabilized after a finite number of iterations. Thus, Lines 5 and 14 of Algorithm 1 will not be performed when the iteration index t is large enough. We begin by introducing the assumptions.

Assumption 3 (*Regularity condition*) We assume the iterate $\{(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)\}$ and trial point $\{(\boldsymbol{x}_{s_t}, \boldsymbol{\mu}_{s_t}, \boldsymbol{\lambda}_{s_t})\}$ are contained in a convex compact region $\mathcal{X} \times \mathcal{M} \times \Lambda$. Further, if $\boldsymbol{x}_{s_t} \in \mathcal{T}_{\bar{\nu}_t}$, then the segment $\{\zeta \boldsymbol{x}_t + (1-\zeta)\boldsymbol{x}_{s_t} : \zeta \in (0,1)\} \subseteq \mathcal{T}_{\theta\bar{\nu}_t}$ for some $\theta \in [1,2)$. We also assume the functions f, g, c are thrice continuously differentiable over $\mathcal{X}$, and realizations $|F(\boldsymbol{x}, \xi)|$, $\|\nabla F(\boldsymbol{x}, \xi)\|$, $\|\nabla^2 F(\boldsymbol{x}, \xi)\|$ are uniformly bounded over $\boldsymbol{x} \in \mathcal{X}$ and $\xi \sim \mathcal{P}$.

Assumption 4 (*Constraint qualification*) For any $\boldsymbol{x} \in \Omega$, we assume that $(J^T(\boldsymbol{x})\ G^T_{\mathcal{I}(\boldsymbol{x})}(\boldsymbol{x}))$ has full column rank, where Ω is the feasible set in (2) and $\mathcal{I}(\boldsymbol{x})$ is the active set in (3). For any $\boldsymbol{x} \in \mathcal{X} \backslash \Omega$, we assume the linear system

$$\begin{aligned} c_i(\boldsymbol{x}) + \nabla^T c_i(\boldsymbol{x}) \boldsymbol{z} = \boldsymbol{0}, &\qquad i : c_i(\boldsymbol{x}) \neq 0, \\ g_i(\boldsymbol{x}) + \nabla^T g_i(\boldsymbol{x}) \boldsymbol{z} \leq \boldsymbol{0}, &\qquad i : g_i(\boldsymbol{x}) > 0, \end{aligned} \tag{29}$$

has a solution for $\boldsymbol{z} \in \mathbb{R}^d$.

The boundedness condition on realizations in Assumption 3 is widely used in StoSQP analysis to have a well-behaved stochastic penalty parameter sequence [3, 4, 18, 40]. The third derivatives of f, g, c are only required in the analysis and not needed

[3] We note that ξ_2^t may not be generated if Lines 13 and 14 of Algorithm 1 are performed. However, for simplicity we suppose a sample ξ_2^t is still generated in this case, although no quantity is determined by this sample.

Algorithm 1 An Adaptive Active-Set StoSQP with Augmented Lagrangian

```
Input: $(\boldsymbol{x}_0, \boldsymbol{\mu}_0, \boldsymbol{\lambda}_0)$, $\bar{\alpha}_0 = \alpha_{max} > 0$, $\bar{\epsilon}_0, \bar{\delta}_0, \eta, \kappa_{grad}, \chi_{grad}, \chi_f, \chi_{err} > 0$, $\rho > 1$, $\gamma_B \in (0, 1]$, $\beta$,
    $p_{grad}, p_f \in (0, 1)$, $\kappa_f \in (0, \beta/(4\alpha_{max})]$, $\bar{\nu}_0 = 2\sum_{i=1}^r \max\{(g_0)_i, 0\}^3 + 1$;
for $t = 0, 1, 2 \ldots$ do
    Generate $\xi_1^t$ so that (a) (15) holds; (b) (16) holds if $t - 1$ is a successful step; compute $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t$, $\bar{Q}_{1,t}$,
    $\bar{Q}_{2,t}$ as in (9);                                   ▷ Step 1: estimate derivatives

    while {(17) does not hold} OR {(12) is solvable AND (18) does not hold} do
        $\bar{\epsilon}_t \leftarrow \bar{\epsilon}_t/\rho$;                                   ▷ Step 2: set $\bar{\epsilon}_t$
    end while

    if {(12) is not solvable} OR {(12) is solvable AND (19) holds} then
        Obtain a backup direction $\hat{\Delta}_t$ and let $\check{\Delta}_t = \hat{\Delta}_t$;          ▷ Step 3: decide $\check{\Delta}_t$
    else
        $\check{\Delta}_t = \bar{\Delta}_t$;
    end if

    if $\boldsymbol{x}_{s_t} \notin \mathcal{T}_{\bar{\nu}_t}$ then                        ▷ Step 4: estimate merit function
        $(\boldsymbol{x}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$, $\bar{\alpha}_{t+1} = \bar{\alpha}_t$, $\bar{\delta}_{t+1} = \bar{\delta}_t$, $\bar{\epsilon}_{t+1} = \bar{\epsilon}_t$;
        $\bar{\nu}_{t+1} = \rho^j \bar{\nu}_t$ with $j = \lceil \log(2a_{s_t}/\bar{\nu}_t)/\log\rho \rceil$;
    else
        Generate $\xi_2^t$ and compute $\bar{\mathcal{L}}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$, $\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$ so that (24) and (25) hold;
        if $\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} \leq \bar{\mathcal{L}}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} + \beta\bar{\alpha}_t(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\check{\Delta}_t$ then      ▷ Step 5: line search
            $(\boldsymbol{x}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_{s_t}, \boldsymbol{\mu}_{s_t}, \boldsymbol{\lambda}_{s_t})$, $\bar{\alpha}_{t+1} = \rho\bar{\alpha}_t \wedge \alpha_{max}$;     ▷ successful step
            if $-\beta\bar{\alpha}_t(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\check{\Delta}_t \geq \bar{\delta}_t$ then                 ▷ reliable step
                $\bar{\delta}_{t+1} = \rho\bar{\delta}_t$;
            else                                                  ▷ unreliable step
                $\bar{\delta}_{t+1} = \bar{\delta}_t/\rho$;
            end if
        else                                                      ▷ unsuccessful step
            $(\boldsymbol{x}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$, $\bar{\alpha}_{t+1} = \bar{\alpha}_t/\rho$, $\bar{\delta}_{t+1} = \bar{\delta}_t/\rho$;
        end if
        $\bar{\epsilon}_{t+1} = \bar{\epsilon}_t$, $\bar{\nu}_{t+1} = \bar{\nu}_t$;
    end if
end for
```

in the implementation. They are required since the existence of the (generalized) Hessian of the augmented Lagrangian needs the third derivatives. See, for example, [50, Section 6] for the same requirement. For deterministic schemes, the compactness condition on the iterates is typical for the augmented Lagrangian and SQP analyses [7, Chapter 4] [41, Chapter 18]. Some literature relaxed it by assuming all quantities (e.g., the objective gradient and constraints Jacobian, etc.) are uniformly upper bounded with a lower bounded objective (so as the merit function). However, either condition is rather restrictive for StoSQP due to the underlying randomness of the scheme. That said, given the StoSQP iterates presumably contract to a deterministic feasible set, we believe that an unbounded iteration sequence is rare in general. Furthermore, compared to fully stochastic schemes in [3, 4, 18], we generate a batch of samples to have a more precise estimation of the true model in each iteration; thus, our stochastic iterates have a higher chance to closely track the underlying deterministic iterates.

The convexity of $\mathcal{M}\times\Lambda$ can be removed by defining a closed convex hull $\overline{\mathrm{conv}(\mathcal{M})}\times\overline{\mathrm{conv}(\mathcal{M})}$. However, the convexity of the set for the primal iterates is essential to enable a valid Taylor expansion. See [54, Proposition 2.2 and Section 4] [52, Proposition 2.4 and (14)] and references therein for the same requirement for doing line search with (8) and applying its Taylor expansion.

In particular, by the design of Algorithm 1, we have $\boldsymbol{x}_t\in\mathcal{T}_{\bar{\nu}_t}$ for any t, while the trial step $\boldsymbol{x}_{s_t}$ may be outside $\mathcal{T}_{\bar{\nu}_t}$. If $\boldsymbol{x}_{s_t}\notin\mathcal{T}_{\bar{\nu}_t}$, we enlarge $\bar{\nu}_t$ (Line 14) and rerun the iteration from the beginning. Assumption 3 states that if it turns out that $\boldsymbol{x}_{s_t}\in\mathcal{T}_{\bar{\nu}_t}$, then the whole segment $\zeta\boldsymbol{x}_t+(1-\zeta)\boldsymbol{x}_{s_t}$, which may not completely lie in $\mathcal{T}_{\bar{\nu}_t}$ as $\mathcal{T}_{\bar{\nu}_t}$ may be nonconvex, is supposed to lie in a larger space $\mathcal{T}_{\theta\bar{\nu}_t}$ with $\theta\in[1,2)$. Since $\mathcal{L}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$ is SC^1 in $\mathcal{T}^{\circ}_{2\bar{\nu}_t}\times\mathbb{R}^m\times\mathbb{R}^r$ and $\mathcal{T}_{\theta\bar{\nu}_t}\subseteq\mathcal{T}^{\circ}_{2\bar{\nu}_t}$, where $\mathcal{T}^{\circ}_{2\bar{\nu}_t}$ denotes the interior of $\mathcal{T}_{2\bar{\nu}_t}$, the second-order Taylor expansion at $(\boldsymbol{x}_t,\boldsymbol{\mu}_t,\boldsymbol{\lambda}_t)$ is allowed [50]. Note that the range of θ is inessential. If we replace $\nu/2$ in (5) by ν/κ for any $\kappa>1$, then we would allow the existence of θ in $[1,\kappa)$. In other words, θ can be as large as any κ. In fact, the condition on the segment always holds when the input α_{max}, the upper bound of $\bar{\alpha}_t$ (cf. Line 18), is suitably upper bounded. Specifically, supposing $\sup_{\mathcal{X}}\|\nabla a(\boldsymbol{x})\|\vee\sup_t\|\check{\Delta}\boldsymbol{x}_t\|\leq\Upsilon$ (ensured by compactness of iterates), for any $\theta>1$ and $\zeta\in(0,1)$, as long as $\alpha_{max}\leq(\theta-1)\bar{\nu}_0/(2\Upsilon^2)$, we have $\zeta\boldsymbol{x}_t+(1-\zeta)\boldsymbol{x}_{s_t}\in\mathcal{T}_{\theta\bar{\nu}_t}$ by noting that

$$
\begin{aligned}
a(\zeta\boldsymbol{x}_t+(1-\zeta)\boldsymbol{x}_{s_t}) &= a(\boldsymbol{x}_t+\bar{\alpha}_t(1-\zeta)\check{\Delta}\boldsymbol{x}_t)\leq a(\boldsymbol{x}_t)+\bar{\alpha}_t(1-\zeta)\Upsilon^2\\
&\leq\frac{\bar{\nu}_t}{2}+\alpha_{max}\Upsilon^2\leq\frac{\bar{\nu}_t}{2}+\frac{(\theta-1)\bar{\nu}_0}{2}\leq\frac{\bar{\nu}_t}{2}+\frac{(\theta-1)\bar{\nu}_t}{2}=\frac{\theta\bar{\nu}_t}{2}.
\end{aligned}
$$

Clearly, the condition on the segment is not required if $\mathcal{T}_\nu$ in (5) is a convex set, which is the case, for example, if we have linear inequality constraints $\boldsymbol{x}\leq\boldsymbol{0}$; or more generally, each $g_i(\cdot)$ is a convex function. We further investigate the effect of the range of θ by varying κ ($\kappa=2$ by default; cf. (5)) in the experiments.

By the compactness condition and noting that $\bar{\nu}_t$ is increased by at least a factor of ρ each time in (22), we immediately know that $\bar{\nu}_t$ stabilizes when t is large. Moreover, if we let

$$
\tilde{\nu}=\rho^{\tilde{j}}\bar{\nu}_0\quad\text{with}\quad\tilde{j}=\lceil\log(2\max_{\mathcal{X}}a(\boldsymbol{x})/\bar{\nu}_0)/\log\rho\rceil,\tag{30}
$$

then $\bar{\nu}_t\leq\tilde{\nu}$, $t\geq0$, almost surely. We will show a similar result for $\bar{\epsilon}_t$.

Assumption 4 imposes the constraint qualifications. In particular, for feasible points Ω, we assume the linear independence constraint qualification (LICQ), which is a standard condition to ensure the existence and uniqueness of the Lagrangian multiplier [41]. For infeasible points $\mathcal{X}\backslash\Omega$, we assume that the solution set of the linear system (29) is nonempty. The condition (29) restricts the behavior of the constraint functions outside the feasible set, which, together with the compactness condition, implies $\Omega\neq\emptyset$ (cf. [36, Proposition 2.5]). In fact, the condition (29) weakens the generalized Mangasarian-Fromovitz constraint qualification (MFCQ) [59, Definition 2.5]; and relates to the weak MFCQ, which is proposed for problems with only inequalities in [36, Definition 1] and adopted in [50, Assumption A3] and [53, Assumption 3.2]. However, [36] requires the weak MFCQ to hold for feasible points in addition to

LICQ; while [50, 53] and this paper remove such a condition. The condition (29) simplifies and generalizes the weak MFCQ in [36, 50, 53] by including equality constraints. We note that the weak MFCQ is slightly weaker than (29). By the Gordan's theorem [26], (29) implies that $\{c_i \cdot \nabla c_i\}_{i:c_i\neq 0} \cup \{\nabla g_i\}_{i:g_i>0}$ are positively linearly independent:

$$\sum_{i:c_i\neq 0} a_i c_i \nabla c_i + \sum_{i:g_i>0} b_i \nabla g_i \neq \mathbf{0},$$

for any coefficients $a_i, b_i \geq 0$ and $\sum_i a_i^2 + b_i^2 > 0$. In contrast, the weak MFCQ only requires that the above linear combination is nonzero for a particular set of coefficients. However, we adopt the simplified but a bit stronger condition only because (29) has a cleaner form and a clearer connection to SQP subproblems. The coefficients of the weak MFCQ in [36, 50, 53] are relatively hard to interpret. Instead of regarding the constraint qualification as the essence of constraints, those coefficients depend on particular choice of the merit function, although that assumption statement is sharper. That said, (29) is still weaker than other literature on the augmented Lagrangian [34, 47, 49]; and weaker than what is widely assumed in SQP analysis [10], where the IQP system, $c_i + \nabla^T c_i z = \mathbf{0}$, $1 \leq i \leq m$, $g_i + \nabla^T g_i z \leq \mathbf{0}$, $1 \leq i \leq r$, is supposed to have a solution. Moreover, we do not require the strict complementary condition, which is often imposed for the merit functions that apply (squared) slack variables to transform nonlinear inequality constraints [60, A2], [23, Proposition 3.8].

The first lemma shows that (17) is satisfied for a sufficiently small $\bar{\epsilon}_t$. Although (17) is inspired by [40, (19)] for equalities, the proof is quite different from that paper (cf. Lemma 4 there).

Lemma 4 *Under Assumptions 3 and 4, there exists a deterministic threshold $\tilde{\epsilon}_1 > 0$ such that (17) holds for any $\bar{\epsilon}_t \leq \tilde{\epsilon}_1$.*

Proof See Appendix B.1. □

The second lemma shows that (18) is satisfied for small $\bar{\epsilon}_t$. The analysis is similar to Lemma 3. We need the following condition on the SQP system (12).

Assumption 5 We assume that, whenever (12) is solvable, $(J_t^T \;\; G_{t_a}^T)$ has full column rank, and there exist positive constants $\Upsilon_B \geq 1 \geq \gamma_B \vee \gamma_H$ such that

$$B_t \preceq \Upsilon_B I, \qquad M_t \succeq \gamma_H I, \qquad \begin{pmatrix} J_t \\ G_{t_a} \end{pmatrix} \begin{pmatrix} J_t^T & G_{t_a}^T \end{pmatrix} \succeq \gamma_H I,$$

and $z^T B_t z \geq \gamma_B \|z\|^2$, $\forall z \in \{z \in \mathbb{R}^d : J_t z = \mathbf{0}, G_{t_a} z = \mathbf{0}\}$.

Assumption 5 summarizes Assumptions 1 and 2. As shown in Lemma 3, the conditions on M_t and $(J_t^T \;\; G_{t_a}^T)$ hold locally. For the presented global analysis, the Hessian approximation B_t is easy to construct to satisfy the condition, e.g., $B_t = I$; however, such a choice is not proper for fast local rates. In practice, given a lower bound $\gamma_B > 0$, B_t is constructed by doing a regularization on a subsampled Hessian (e.g., for finite-sum objectives) or a sketched Hessian (e.g., for regression objectives), which can preserve certain second-order information and be obtained with less expense. With Assumption 5, we have the following result.

Lemma 5 *Under Assumptions 3 and 5, there exists a deterministic threshold $\tilde{\epsilon}_2 > 0$ such that* (18) *holds for any $\bar{\epsilon}_t \leq \tilde{\epsilon}_2$.*

Proof See Appendix B.2. □

We summarize (30), Lemmas 4 and 5 in the next theorem.

Theorem 1 *Under Assumptions 3, 4, and 5, there exist deterministic thresholds $\tilde{\nu}$, $\tilde{\epsilon} > 0$ such that $\{\bar{\nu}_t, \bar{\epsilon}_t\}_t$ generated by Algorithm 1 satisfy $\bar{\nu}_t \leq \tilde{\nu}$, $\bar{\epsilon}_t \geq \tilde{\epsilon}$. Moreover, almost surely, there exists an iteration threshold $\bar{t} < \infty$, such that $\bar{\epsilon}_t = \bar{\epsilon}_{\bar{t}}$, $\bar{\nu}_t = \bar{\nu}_{\bar{t}}$, $t \geq \bar{t}$.*

Proof The existence of $\tilde{\nu}$ is showed in (30). By Lemmas 4 and 5, and defining $\tilde{\epsilon} = (\tilde{\epsilon}_1 \wedge \tilde{\epsilon}_2)/\rho$, we show the existence of $\tilde{\epsilon}$. The existence of the iteration threshold $\bar{t}$ is ensured by noting that $\{\bar{\nu}_t, 1/\bar{\epsilon}_t\}_t$ are bounded from above; and each update increases the parameters by at least a factor of $\rho > 1$. □

We mention that the iteration threshold $\bar{t}$ is random for stochastic schemes and it changes between different runs. However, it always exists. The following analysis supposes t is large enough such that $t \geq \bar{t}$ and $\bar{\epsilon}_t, \bar{\nu}_t$ have stabilized. We condition our analysis on the σ-algebra $\mathcal{F}_{\bar{t}}$, which means that we only consider the randomness of the generated samples after $\bar{t} + 1$ iterations and, by (28), the parameters $\bar{\epsilon}_{\bar{t}}$, $\bar{\nu}_{\bar{t}}$ are fixed. We should point out that, although it is standard to focus only on the tail of the iteration sequence to show the global convergence (even for the deterministic case [41, Theorem 18.3]), an important aspect that is missed by such an analysis is the non-asymptotic guarantees. In particular, we know the scheme changes the merit parameters for at most $\log(\tilde{\nu}\bar{\epsilon}_0/(\bar{\nu}_0\tilde{\epsilon}))/\log(\rho)$ times; however, how many iterations it spans for all the changes is not answered by our analysis. Establishing a bound on $\bar{t}$ in expectation or high probability sense would help us further understand the efficiency of the scheme. However, since any characterization of $\bar{t}$ is difficult even for deterministic schemes, we leave such a study to the future. Another missing aspect is the iteration complexity, where we are interested in the number of iterations to attain an ϵ-first- or second-order stationary point (we abuse ϵ notation here to refer to the accuracy level). The iteration complexity is recently studied for two StoSQP schemes under very particular setups [5, 17]; none of the existing works allow either stochastic line search or inequality constraints. We leave the iteration complexity of our scheme to the future as well.

3.3 Convergence analysis

We conduct the global convergence analysis for Algorithm 1. We prove that $\lim_{t\rightarrow\infty} R_t = 0$ *almost surely*, where $R_t = \|(\nabla_{\boldsymbol{x}}\mathcal{L}_t, c_t, \max\{g_t, -\boldsymbol{\lambda}_t\})\|$ is the KKT residual. We suppose the line search conditions (15), (16), (24), (25) hold. We will discuss the sample complexities that ensure these generic conditions in Sect. 3.4. It is fairly easy to see that all conditions hold for large batch sizes.

Our proof structure closely follows [40]. The analyses are more involved in Lemmas 7, 9, 10, 11 and Theorem 3, which account for the differences between equality and inequality constraints, and account for our relaxations of the feasibility error condition and the increasing sample size requirement of [40]. The analysis in Theorem 5 is new, which strengthens the “liminf" convergence in [40]. The analyses are slightly adjusted in Theorem 4, and the same in Lemma 8 and Theorem 2. The adopted potential function (or Lyapunov function) is

$$\Theta^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta,\omega} = \omega\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \frac{1-\omega}{2}\bar{\alpha}_t\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 + \frac{1-\omega}{2}\bar{\delta}_t, \quad t \geq \bar{t}+1, \tag{31}$$

where $\omega \in (0, 1)$ is a coefficient to be specified later. We note that using $\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ by itself (i.e., $\omega = 1$) to monitor the iteration progress is not suitable for the stochastic setting; it is possible that $\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ increases while $\bar{\mathcal{L}}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ decreases. In contrast, $\Theta^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta,\omega}$ linearly combines different components and has a composite measure of the progress. For example, the decrease of $\Theta^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta,\omega}$ may come from $\bar{\delta}_t$ (Lines 22 and 25 of Algorithm 1).

Since parameters $\bar{\epsilon}_{\bar{t}}$, $\bar{\nu}_{\bar{t}}$, η in $\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ are fixed (conditional on $\mathcal{F}_{\bar{t}}$), we denote $\Theta^t_\omega = \Theta^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta,\omega}$ for notational simplicity. In the presentation of theoretical results, we only track the parameters $(\beta, \alpha_{max}, \kappa_{grad}, \kappa_f, p_{grad}, p_f, \chi_{grad}, \chi_f)$ that relate to the line search conditions. In particular, we use $C_1, C_2 \ldots$ and $\Upsilon_1, \Upsilon_2 \ldots$ to denote deterministic constants that are independent from these parameters, but may depend on $(\gamma_B, \gamma_H, \Upsilon_B, \chi_u, \chi_{err}, \rho, \eta, \bar{\epsilon}_0, \bar{\nu}_0)$, and thus depend on the deterministic thresholds $\tilde{\epsilon}$ and $\tilde{\nu}$. Recall that $(\gamma_B, \gamma_H, \Upsilon_B, \chi_u)$ come from Assumption 5 and (20), while $(\chi_{err}, \rho, \eta, \bar{\epsilon}_0, \bar{\nu}_0)$ are any algorithm inputs.

The first lemma presents a preliminary result.

Lemma 6 *Under Assumptions 3, 4, 5, the following results hold deterministically conditional on* $\mathcal{F}_{t-1}$.

(a) There exists $C_1 > 0$ *such that the following two inequalities hold for any iteration* $t \geq 0$ *((a2) also holds for* s_t*), any parameters* ϵ, ν, *and any generated sample set* ξ*:*

(a1) $\left\|\bar{\nabla}\mathcal{L}^t_{\epsilon,\nu,\eta} - \nabla\mathcal{L}^t_{\epsilon,\nu,\eta}\right\| \leq C_1\left\{\|\bar{\nabla}f_t - \nabla f_t\| \vee (\bar{R}_t \wedge 1)\right\} \cdot \|\bar{\nabla}^2 f_t - \nabla^2 f_t\|$;

(a2) $\left|\bar{\mathcal{L}}^t_{\epsilon,\nu,\eta} - \mathcal{L}^t_{\epsilon,\nu,\eta}\right| \leq C_1\{|\bar{f}_t - f_t| \vee [(\bar{R}_t \vee \|\bar{\nabla}f_t - \nabla f_t\|) \wedge 1] \cdot \|\bar{\nabla}f_t - \nabla f_t\|\}$.

(b) There exists $C_2 > 0$ *such that for any* $t \geq 0$ *and set* ξ*,*

$$\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\| \leq C_2\left\{\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\| + \|(c_t,\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t})\|\right\}.$$

(c) There exists $C_3 > 0$ *such that for any* $t \geq 0$ *and set* ξ*, if (12) is solvable, then*

$$\left\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right\| \leq C_3\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t \\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|.$$

Proof See Appendix B.3. □

The results in Lemma 6 hold deterministically conditional on $\mathcal{F}_{t-1}$, because the samples ξ for computing $\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$, $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t$ are supposed to be also given by the statement. The following result suggests that if both the gradient $\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ and the function evaluations $\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$, $\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ are precisely estimated, in the sense that the event $\mathcal{E}_1^t\cap\mathcal{E}_2^t$ happens (cf. (14), (23)), then there is a uniform lower bound on $\bar{\alpha}_t$ to make the Armijo condition hold.

Lemma 7 *For $t\geq\bar{t}+1$, suppose $\mathcal{E}_1^t\cap\mathcal{E}_2^t$ happens. There exists $\Upsilon_1>0$ such that the t-th step satisfies the Armijo condition* (26) *(i.e., is a successful step) if*

$$\bar{\alpha}_t\leq\frac{1-\beta}{\Upsilon_1(\kappa_{grad}+\kappa_f+1)}.$$

Proof See Appendix B.4. □

The next result suggests that, if only the function evaluations $\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$, $\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ are precisely estimated, in the sense that the event $\mathcal{E}_2^t$ happens, then a sufficient decrease of $\bar{\mathcal{L}}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ implies a sufficient decrease of $\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$. The proof directly follows [40, Lemma 6], and thus is omitted.

Lemma 8 *For $t\geq\bar{t}+1$, suppose $\mathcal{E}_2^t$ happens. If the t-th step satisfies the Armijo condition* (26)*, then*

$$\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\leq\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}+\frac{\bar{\alpha}_t\beta}{2}(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t.$$

Based on Lemmas 7 and 8, we now establish an error recursion for the potential function Θ^t_ω in (31). Our analysis is separated into three cases according to the events: $\mathcal{E}_1^t\cap\mathcal{E}_2^t$, $(\mathcal{E}_1^t)^c\cap\mathcal{E}_2^t$ and $(\mathcal{E}_2^t)^c$. We will show that Θ^t_ω decreases in the case of $\mathcal{E}_1^t\cap\mathcal{E}_2^t$, while may increase in the other two cases. Fortunately, by letting p_{grad} and p_f be small, Θ^t_ω always decreases in expectation.

We first show in Lemma 9 that Θ^t_ω decreases when $\mathcal{E}_1^t\cap\mathcal{E}_2^t$ happens. We note that the decrease of Θ^t_ω exceeds $\bar{\alpha}_t\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2$ by $\bar{\delta}_t$ (up to a multiplier).

Lemma 9 *For $t\geq\bar{t}+1$, suppose $\mathcal{E}_1^t\cap\mathcal{E}_2^t$ happens. There exists $\Upsilon_2>0$, such that if ω satisfies*

$$\frac{1-\omega}{\omega}\leq\frac{\beta}{\Upsilon_2(\kappa_{grad}\alpha_{max}+\alpha_{max}+1)^2}\wedge\frac{1}{18(\rho-1)},\tag{32}$$

then

$$\Theta^{t+1}_\omega-\Theta^t_\omega\leq-\frac{1}{2}(1-\omega)\left(1-\frac{1}{\rho}\right)\left(\bar{\alpha}_t\left\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\right\|^2+\bar{\delta}_t\right).$$

Proof See Appendix B.5. □

We then show in Lemma 10 that Θ_ω^t may increase, if $\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ is not precisely estimated (i.e., $(\mathcal{E}_1^t)^c$ happens) but $\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$, $\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ are precisely estimated (i.e., $\mathcal{E}_2^t$ happens). The increase is proportional to $\bar{\alpha}_t\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2$.

Lemma 10 *For $t \geq \bar{t}+1$, suppose $(\mathcal{E}_1^t)^c \cap \mathcal{E}_2^t$ happens. Under (32), we have*

$$\Theta_\omega^{t+1} - \Theta_\omega^t \leq \rho(1-\omega)\bar{\alpha}_t \left\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\right\|^2.$$

Proof See Appendix B.6. □

We finally show in Lemma 11 that Θ_ω^t increases and the increase can exceed $\bar{\alpha}_t\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2$, if $\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$, $\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}$ are not precisely estimated. In this case, the exceeding terms have to be controlled by making use of the condition (25).

Lemma 11 *For $t \geq \bar{t}+1$, suppose $(\mathcal{E}_2^t)^c$ happens. Under (32), we have*

$$\begin{aligned}\Theta_\omega^{t+1} - \Theta_\omega^t \leq\ & \rho(1-\omega)\bar{\alpha}_t \left\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\right\|^2 \\ & + \omega\left\{\left|\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\right|\right\}.\end{aligned}$$

Proof See Appendix B.7. □

Combining Lemmas 9, 10, 11, we derive the one-step error recursion of Θ_ω^t. The proof directly follows that of [40, Theorem 2] and is omitted.

Theorem 2 (One-step error recursion) *For $t \geq \bar{t}+1$, suppose ω satisfies (32) and p_{grad} and p_f satisfy*

$$\frac{p_{grad} + \sqrt{(1 \vee \chi_f)\cdot p_f}}{(1-p_{grad})(1-p_f)} \leq \frac{\rho-1}{8\rho}\left\{\frac{1}{\rho} \wedge \frac{1-\omega}{\omega}\right\}. \tag{33}$$

Then

$$\begin{aligned}&\mathbb{E}\left[\Theta_\omega^{t+1} - \Theta_\omega^t \mid \mathcal{F}_{t-1}\right] \\ &\leq -\frac{1}{4}(1-p_{grad})(1-p_f)(1-\omega)\left(1-\frac{1}{\rho}\right)\left(\bar{\delta}_t + \bar{\alpha}_t\left\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\right\|^2\right).\end{aligned}$$

With Theorem 2, we derive the convergence of $\bar{\alpha}_t R_t^2$ in the next theorem, where $R_t = \|(\nabla_{\boldsymbol{x}}\mathcal{L}_t, c_t, \max\{g_t, -\boldsymbol{\lambda}_t\})\|$ is the KKT residual.

Theorem 3 *Under the conditions of Theorem 2, $\lim_{t\to\infty}\bar{\alpha}_t R_t^2 = 0$ almost surely.*

Proof See Appendix B.8. □

Then, we show that the "liminf" of the KKT residuals converges to zero.

Theorem 4 ("liminf" convergence) *Consider Algorithm 1 under Assumptions 3, 4, 5. Suppose ω satisfies (32) and p_{grad}, p_f satisfy (33). Then, almost surely, we have that* $\liminf_{t\to\infty} R_t = 0$.

Proof See Appendix B.9. □

Finally, we strengthen the statement in Theorem 4 and complete the global convergence analysis of Algorithm 1.

Theorem 5 (Global convergence) *Under the same conditions of Theorem 4, we have that*

$$\lim_{t\to\infty} R_t = 0, \quad \textit{almost surely}.$$

Proof See Appendix B.10. □

Our analysis generalizes the results of [40] to inequality constrained problems. The "lim" convergence guarantee in Theorem 5 strengthens the existing "liminf" convergence guarantee of stochastic line search for both unconstrained problems [44, Theorem 4.10] and equality constrained problems [40, Theorem 4]. Theorem 5 also differs from the results in [3, 4, 18], where the authors showed the (liminf) convergence of the *expected* KKT residual under a fully stochastic setup. Compared to [3, 4, 18], our scheme does not tune a deterministic sequence that controls the stepsizes and determines the convergence behavior (i.e., converging to a KKT point or only its neighborhood). Our scheme tunes two probability parameters p_{grad}, p_f. Seeing from (32) and (33), the upper bound conditions on p_{grad}, p_f depend on the inputs $(\rho, \beta, \kappa_{grad}, \alpha_{max})$ and a universal constant Υ_2. Estimating Υ_2 is often difficult in practice; however, p_{grad}, p_f affect the algorithm's performance only via the generated batch sizes, and the batch sizes depend on p_{grad}, p_f only via the logarithmic factors (see (37) and (41) later). Thus, the algorithm is robust to p_{grad}, p_f. We will also empirically test the robustness to parameters for Algorithm 1 in Sect. 4. In addition, (32) and (33) suggest that the larger the parameters $(\rho, 1/\beta, \kappa_{grad}, \alpha_{max})$ we use, the smaller the probabilities p_{grad}, p_f have to be. Such a dependence is consistent with the general intuition: the algorithm performs more aggressive updates with less restrictive Armijo condition when $(\rho, 1/\beta, \kappa_{grad}, \alpha_{max})$ are large; thus, a more precise model estimation in each iteration is desired in this case.

3.4 Discussion on sample complexities

As introduced in Sect. 1, the stochastic line search is performed by generating a batch of samples in each iteration to have a precise model estimation, which is standard in the literature [11, 13, 14, 20, 22, 29, 44]. The batch sizes are adaptively controlled based on the iteration progress. We now discuss the batch sizes $|\xi_1^t|$ and $|\xi_2^t|$ to ensure the generic conditions (15), (16), (24), (25) of Algorithm 1. We show that, if the KKT residual R_t does not vanish, all the conditions are satisfied by properly choosing $|\xi_1^t|$ and $|\xi_2^t|$.

Sample complexity of ξ_1^t The samples ξ_1^t are used to estimate ∇f_t and $\nabla^2 f_t$ in Step 1 of Algorithm 1. The estimators $\bar{\nabla} f_t$ and $\bar{\nabla}^2 f_t$ can be computed with different amount of samples, and their samples may or may not be independent. Let us suppose $\bar{\nabla} f_t$ is computed by samples ξ_1^t, while $\bar{\nabla}^2 f_t$ is computed by a subset of samples $\tau_1^t \subseteq \xi_1^t$. The case where $\bar{\nabla} f_t$ and $\bar{\nabla}^2 f_t$ are computed by two disjoint subsets of ξ_1^t can be studied following the same analysis. We define

$$\bar{\nabla} f_t = \frac{1}{|\xi_1^t|} \sum_{\xi \in \xi_1^t} \nabla F(\boldsymbol{x}_t; \xi), \qquad \bar{\nabla}^2 f_t = \frac{1}{|\tau_1^t|} \sum_{\xi \in \tau_1^t} \nabla^2 F(\boldsymbol{x}_t; \xi).$$

By Lemma 6(a1), we know that (15) holds if, with probability $1 - p_{grad}$,

$$\|\bar{\nabla} f_t - \nabla f_t\| \le O(\kappa_{grad} \bar{\alpha}_t \bar{R}_t), \quad \|\bar{\nabla}^2 f_t - \nabla^2 f_t\| \le O(\kappa_{grad} \bar{\alpha}_t \bar{R}_t / (\bar{R}_t \wedge 1)), \tag{34}$$

where we suppress universal constants (such as the variance of a single sample) in $O(\cdot)$ notation. By matrix Bernstein inequality [58, Theorem 7.7.1], (34) is satisfied if

$$|\xi_1^t| \ge O\left(\frac{\log(d/p_{grad})}{\kappa_{grad}^2 \bar{\alpha}_t^2 \bar{R}_t^2}\right) \quad \text{and} \quad |\tau_1^t| \ge (\bar{R}_t^2 \wedge 1) \cdot |\xi_1^t|. \tag{35}$$

Furthermore, we use the bound $\mathbb{E}[\|\bar{\nabla}^2 f_t - \nabla^2 f_t\|^2 \mid \mathcal{F}_{t-1}] \le O(\log d / |\tau_1^t|)$ (cf. [58, (6.1.6)]) and know that (16) holds if

$$|\xi_1^t| \ge O\left(\frac{\bar{\alpha}_t \log d}{\chi_{grad}^2 \bar{\delta}_t}\right) \quad \text{and} \quad |\tau_1^t| \ge (\bar{R}_t^2 \wedge 1) \cdot |\xi_1^t|. \tag{36}$$

Combining (35) and (36) together, we know that the conditions (15) and (16) are satisfied if

$$|\xi_1^t| \ge O\left(\frac{\log(d/p_{grad})}{\kappa_{grad}^2 \bar{\alpha}_t^2 \bar{R}_t^2 \wedge \chi_{grad}^2 \bar{\delta}_t / \bar{\alpha}_t}\right), \quad |\tau_1^t| \ge (\bar{R}_t^2 \wedge 1) \cdot |\xi_1^t|. \tag{37}$$

Since (16) is imposed only when $t-1$ is a successful step, the term $\chi_{grad}^2 \bar{\delta}_t / \bar{\alpha}_t$ on the denominator in (37) can be removed when $t-1$ is an unsuccessful step. In contrast to [40], where the gradient ∇f_t and Hessian $\nabla^2 f_t$ are computed based on the same set of samples, we sharpen the calculation and realize that the batch size $|\tau_1^t|$ for $\nabla^2 f_t$ can be significantly less than $|\xi_1^t|$ for ∇f_t. When $\bar{R}_t$ gets close to zero, the ratio $|\tau_1^t|/|\xi_1^t|$ will also decay to zero.

We mention that $\bar{R}_t$ on the right hand side of the condition $|\xi_1^t|$ in (37) has to be computed by samples ξ_1^t. A practical algorithm can first specify ξ_1^t, then compute $\bar{R}_t$, and finally check if (37) holds. For example, a While loop can be designed to gradually increase $|\xi_1^t|$ until (37) holds (cf. [40, Algorithm 4]). Such a While loop always terminates in finite time when $R_t > 0$, because $\bar{R}_t \to R_t$ as $|\xi_1^t|$ increases (by the law of large number) so that the right hand side of (37) does not diverge.

Sample complexity of ξ_2^t The samples ξ_2^t are used to estimate f_t, f_{s_t}, ∇f_t, ∇f_{s_t} in Step 4 of Algorithm 1. Similar to the discussion above, the estimators $\bar{f}_t$, $\bar{f}_{s_t}$ and $\bar{\bar{\nabla}} f_t$, $\bar{\bar{\nabla}} f_{s_t}$ can be computed with different amount of samples, and their samples may or may not be independent. Let us suppose $\bar{f}_t$, $\bar{f}_{s_t}$ are computed by samples ξ_2^t, while $\bar{\bar{\nabla}} f_t$, $\bar{\bar{\nabla}} f_{s_t}$ are computed by a subset of samples $\tau_2^t \subseteq \xi_2^t$. We define (similar for $\bar{f}_{s_t}$, $\bar{\bar{\nabla}} f_{s_t}$)

$$\bar{f}_t = \frac{1}{|\xi_2^t|} \sum_{\xi \in \xi_2^t} F(\boldsymbol{x}_t; \xi), \qquad \bar{\bar{\nabla}} f_t = \frac{1}{|\tau_2^t|} \sum_{\xi \in \tau_2^t} \nabla F(\boldsymbol{x}_t; \xi).$$

By Lemma 6(a2), we know that (24) holds if, with probability $1 - p_f$,

$$\begin{aligned} &|\bar{f}_t - f_t| \vee |\bar{f}_{s_t} - f_{s_t}| \leq O(-\kappa_f \bar{\alpha}_t^2 (\bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t)^T \check{\Delta}_t), \\ &\quad \|\bar{\bar{\nabla}} f_t - \nabla f_t\| \vee \|\bar{\bar{\nabla}} f_{s_t} - \nabla f_{s_t}\| \leq O \left\{ \frac{-\kappa_f \bar{\alpha}_t^2 (\bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t)^T \check{\Delta}_t}{\left\{ \bar{\bar{R}}_t \vee \bar{\bar{R}}_{s_t} \vee \{-\kappa_f \bar{\alpha}_t^2 (\bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t)^T \check{\Delta}_t\}^{1/2} \right\} \wedge 1} \right\}, \end{aligned} \tag{38}$$

where $\bar{\bar{R}}_t$ and $\bar{\bar{R}}_{s_t}$ are computed by τ_2^t and we use the fact that, for scalars a, b, $(a \wedge 1) \vee (b \wedge 1) = (a \vee b) \wedge 1$. By Bernstein inequality, (38) is satisfied if

$$\begin{aligned} |\xi_2^t| &\geq O \left(\frac{\log(d/p_f)}{\kappa_f^2 \bar{\alpha}_t^4 \{(\bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t)^T \check{\Delta}_t\}^2} \right), \\ |\tau_2^t| &\geq \left(\{\bar{\bar{R}}_t^2 \vee \bar{\bar{R}}_{s_t}^2 \vee -\kappa_f \bar{\alpha}_t^2 (\bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t)^T \check{\Delta}_t\} \wedge 1 \right) \cdot |\xi_2^t| \\ &= \left(\{\bar{\bar{R}}_t^2 \vee \bar{\bar{R}}_{s_t}^2\} \cdot |\xi_2^t| \vee \{\log(d/p_f) \cdot |\xi_2^t|\}^{1/2} \right) \wedge |\xi_2^t|. \end{aligned} \tag{39}$$

Moreover, by $\mathbb{E}[|\bar{f}_t - f_t| \mid \mathcal{F}_{t-0.5}] \leq O(1/|\xi_2^t|)$ and $\mathbb{E}[\|\bar{\bar{\nabla}} f_t - \nabla f_t\|^4] \leq O(1/|\tau_2^t|^2)$, we can see that (25) holds if

$$|\xi_2^t| \geq O(1/(\chi_f \bar{\delta}_t^2)), \quad |\tau_2^t| \geq \left(\{\bar{\bar{R}}_t^2 \vee \bar{\bar{R}}_{s_t}^2 \vee \chi_f \bar{\delta}_t^2\} \wedge 1 \right) \cdot |\xi_2^t|. \tag{40}$$

Combining (39) and (40) together, the conditions (24) and (25) are satisfied if

$$\begin{aligned} |\xi_2^t| &\geq O \left(\frac{\log(d/p_f)}{\kappa_f^2 \bar{\alpha}_t^4 \{(\bar{\nabla} \mathcal{L}_{\bar{\epsilon}_t, \bar{\nu}_t, \eta}^t)^T \check{\Delta}_t\}^2 \wedge \chi_f \bar{\delta}_t^2} \right), \\ |\tau_2^t| &\geq \left(\{\bar{\bar{R}}_t^2 \vee \bar{\bar{R}}_{s_t}^2\} \cdot |\xi_2^t| \vee \{\log(d/p_f) \cdot |\xi_2^t|\}^{1/2} \right) \wedge |\xi_2^t|. \end{aligned} \tag{41}$$

Similar to the complexity (37), (41) suggests that the batch size $|\tau_2^t|$ for ∇f_t, ∇f_{s_t} is significantly less than $|\xi_2^t|$ for f_t, f_{s_t}, with the ratio $|\tau_2^t|/|\xi_2^t|$ decaying to zero when t increases. The denominator in (41) is nonzero if $\bar{R}_t \neq 0$ (which is always the case;

otherwise, we should stop the iteration). In particular, if $\check{\Delta}_t = \bar{\Delta}_t$, then

$$\begin{aligned}-\kappa_f\bar{\alpha}_t^2(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\check{\Delta}_t &\overset{\text{(B.24)}}{\ge} \frac{\kappa_f\bar{\alpha}_t^2(\gamma_B\wedge\eta)}{4}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\text{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2\\ &\overset{\text{(B.26)}}{\ge} O(\kappa_f\bar{\alpha}_t^2\bar{R}_t^2)>0;\end{aligned}$$

if $\check{\Delta}_t = \hat{\Delta}_t$, then

$$-\kappa_f\bar{\alpha}_t^2(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\check{\Delta}_t \overset{(20)}{\ge} \kappa_f\bar{\alpha}_t^2/\chi_u\left\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right\|^2 \overset{\text{(B.28)}}{\ge} O(\kappa_f\bar{\alpha}_t^2\bar{R}_t^2)>0.$$

3.5 Discussion on computations and limitations

We now briefly discuss the per-iteration computational cost of Algorithm 1, and present some limitations and extensions of the algorithm.

Objective evaluations By Sect. 3.4 and the complexities in (37) and (41), Algorithm 1 generates $|\xi_1^t| + |\xi_2^t|$ samples in each iteration, and evaluates $2|\xi_2^t|$ function values, $|\xi_1^t| + 2|\tau_2^t|$ gradients, and $|\tau_1^t|$ Hessians for the objective. To see their orders from (37) and (41) clearly, let us suppose $\bar{\alpha}_t$ stabilizes at α_{max} (i.e., the steps are successful) and $\bar{\delta}_t = O((\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta})^T\check{\Delta}_t)$ (see (27) for the reasonability). We also replace the stochastic quantities in (37), (41) by deterministic counterparts and let $R_t \approx R_{s_t}$. Then, we can see that $|\xi_1^t| = |\tau_2^t| = O(1/R_t^2)$, $|\xi_2^t| = O(1/R_t^4)$, and $|\tau_1^t| = O(1)$. Thus, the objective evaluations are

$$\text{function values: } O(1/R_t^4), \quad \text{gradients: } O(1/R_t^2), \quad \text{Hessians: } O(1).$$

We note that the evaluations for the function values and gradients are increasing as the iteration proceeds, and the function evaluations are square of the gradient evaluations. Under the same setup, our evaluation complexities for the functions and gradients are consistent with the unconstrained stochastic line search [44, Section 2.3] with R_t replaced by $\|\nabla f_t\|$. Although the augmented Lagrangian merit function requires the Hessian evaluations, the Hessian complexity is significantly less than that of functions and gradients, and does not have to increase during the iteration. Such an observation is missing in the prior work [40].

Constraint evaluations Since the constraints are deterministic, Algorithm 1 has the same constraint evaluations as deterministic schemes. In particular, the algorithm evaluates four function values (two for equalities and two for inequalities; and for each type of constraint, one for current point and one for trial point), four Jacobians, and two Hessians in each iteration.

Computational cost Same as deterministic SQP schemes, solving Newton system dominates the computational cost. If we do not consider the potential sparse or block-diagonal structures that many problems have, solving the system (12) requires $O((d + m + |\text{active set}|)^3) + O((m + r)^3) = O(d^3 + m^3 + r^3)$ flops. Such computational cost is larger than solving a standard SQP system (see [50, (8.9)]) by the extra term $O((m+r)^3)$. However, as explained in Remark 1, the analysis of standard SQP system relies on the exact Hessian, which is inaccessible in our stochastic setting. When the SQP direction is not employed, the backup direction can be obtained with $O(d+m+r)$ flops for the gradient step, $O((d+m+r)^3)$ flops for the regularized Newton step, and between for the truncated Newton step. Such computational cost is standard in the literature [53, 55], where a safeguarding direction satisfying (20) is required to minimize the augmented Lagrangian. We should mention that, as the EQP scheme, the above computations are not very comparable with the IQP schemes. In that case, the SQP systems include inequality constraints and are more expensive to solve, although less iterations may be performed.

Limitations of the design Algorithm 1 has few limitations. First, it solves the SQP systems exactly. In practice, one may apply conjugate gradient (CG) or minimum residual (MINRES) methods, or apply randomized iterative solvers to solve the systems inexactly. The inexact direction can reduce the computational cost significantly [18]. Second, our backup direction does not fully utilize the computations of the SQP direction. Although our analysis allows any backup direction satisfying (20), and utilizing Newton direction as a backup is standard in the literature [53, 55], a better choice is to directly modify the SQP direction. Then, we may derive a direction that has a faster convergence than the gradient direction, and less computations than the (regularized) Newton direction. We leave the refinements of these two limitations to the future.

4 Numerical experiments

We implement the following two algorithms on 39 nonlinear problems collected in CUTEst test set [27]. We select the problems that have a non-constant objective with less than 1000 free variables. We also require the problems to have at least one inequality constraint, no infeasible constraints, no network constraints; and require the number of constraints to be less than the number of variables. The setup of each algorithm is as follows.

(a) **AdapNewton**: the adaptive scheme in Algorithm 1 with the safeguarding direction given by the regularized Newton step. We set the inputs as $\bar{\alpha}_0 = \alpha_{max} = 1.5$, $\beta = 0.3$, $\kappa_f = \beta/(4\alpha_{max}) = 0.05$, $\kappa_{grad} = \chi_{grad} = \chi_f = \bar{\delta}_0 = 1$, $\bar{\epsilon}_0 = 10^{-2}$, $\eta = 10^{-4}$, $p_{grad} = p_f = 0.1$, $\rho = 2$. Here, we set $\alpha_{max} > 1$ since a stochastic scheme can select a stepsize that is greater than one (cf. Fig. 4). β is close to the middle of the interval (0, 0.5), which is a common range for deterministic schemes. $(\bar{\epsilon}_0, \bar{\delta}_0)$ are adaptively selected during the iteration, while we prefer a small initial $\bar{\epsilon}_0$ to run less adjustments on it. κ_f is set as the allowed largest value $\beta/(4\alpha_{max})$ (cf. Algorithm 1); however, the parameters $(\kappa_{grad}, \kappa_f, \chi_{grad}, \chi_f, p_{grad}, p_f)$ all affect the batch sizes and play the same role as the constant C that we study later. We let η be small so that the last penalty term of (8) is almost negligible, and the merit function (8) is close to a standard augmented Lagrangian function. We also test the robustness of the algorithm to three parameters C, κ, χ_{err}. Here, C is the constant multiplier of the big "O" notation in (37) and (41) (the variance σ^2 of a single sample is also absorbed in "O", which we introduce later). κ is a parameter of the set $\mathcal{T}_\nu$ ($\kappa = 2$ in (5)), and χ_{err} is a parameter of the feasibility error condition (17). Their default values are $C = \kappa = 2$ and $\chi_{err} = 1$, while we allow to vary them in wide ranges: $C, \kappa \in \{2, 2^3, 2^6\}$ and $\chi_{err} \in \{1, 10, 10^2\}$. When we vary one parameter, the other two are set as default.

(b) **AdapGD**: the adaptive scheme in Algorithm 1 with the safeguarding direction given by the steepest descent step. The setup is the same as (b).

For both algorithms, the initial iterate $(\boldsymbol{x}_0, \boldsymbol{\mu}_0, \boldsymbol{\lambda}_0)$ is specified by the CUTEst package. The package also provides the deterministic function, gradient and Hessian evaluation, $f_t, \nabla f_t, \nabla^2 f_t$, in each iteration. We generate their stochastic counterparts by adding a Gaussian noise with variance σ^2. In particular, we let $\bar{f}_t \sim \mathcal{N}(f_t, \sigma^2)$, $\bar{\nabla} f_t \sim \mathcal{N}(\nabla f_t, \sigma^2(I + \mathbf{1}\mathbf{1}^T))$, and $(\bar{\nabla}^2 f_t)_{ij} \sim \mathcal{N}((\nabla f_t)_{ij}, \sigma^2)$. We try four levels of variance: $\sigma^2 \in \{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}\}$. Throughout the implementation, we let $B_t = I$ (cf. (12), (21)) and set the iteration budget to be 10^4. The stopping criterion is

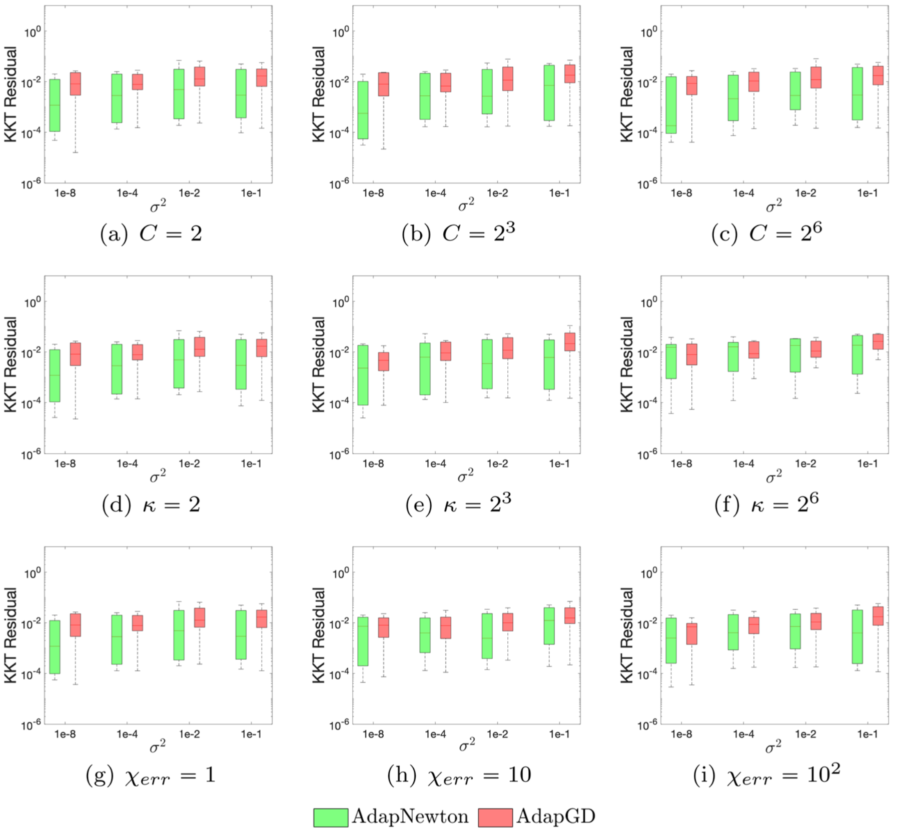
$$\bar{\alpha}_t \|\check{\Delta}_t\| \leq 10^{-7} \quad \text{OR} \quad R_t \leq 10^{-5} \quad \text{OR} \quad t \geq 10^4.$$

The former two cases suggest that the iteration converges within the budget. For each algorithm, each problem, and each setup, we average the results of all convergent runs among 5 runs. Our code is available at https://github.com/senna1128/Constrained-Stochastic-Optimization-Inequality.

KKT residuals We draw the KKT residual boxplots for AdapNewton and AdapGD in Fig. 1. From the figure, we see that both algorithms are robust to tuning parameters (C, κ, χ_{err}). For both algorithms, the median of the KKT residuals gradually increases as σ^2 increases, which is reasonable since the model estimation of each sample is more noisy when σ^2 is larger. However, the increase of the KKT residuals is mild since, regardless of σ^2, both methods generate enough samples in each iteration to enforce the model accuracy conditions (i.e., (15), (16), (24), (25)). Figure 1 also suggests that AdapNewton outperforms AdapGD although the improvement is limited. In fact, the convergence on a few problems may be improved by utilizing the regularized Newton step as the backup of the SQP step; however, the SQP step will be employed eventually.

Sample sizes We draw the sample size boxplots for AdapNewton and AdapGD in Fig. 2. From the figure, we see that both methods generate much less samples for estimating the objective Hessian compared to estimating the objective value and gradient, between which the the objective gradient is estimated with less samples than the objective value. The sample size differences of the three quantities—objective value, gradient, Hessian—are clearer as σ^2 increases. For a fixed σ^2, the sample sizes of different setups of (C, κ, χ_{err}) do not vary much. In fact, the parameters κ, χ_{err} do not directly affect the sample complexities. The parameter C plays a similar role to σ^2 and affects the sample complexities via changing the multipliers in (37) and (41). However, varying C from 2 to 64 is marginal compared to varying σ^2 from 10^{-8} to 10^{-1}. Thus, Fig. 2 again illustrates the robustness of the designed adaptive algorithm.

Moreover, as discussed in Sects. 3.4 and 3.5, the objective value, gradient, and Hessian have different sample complexities in each iteration, which depend on different



(a) $C = 2$ (b) $C = 2^3$ (c) $C = 2^6$

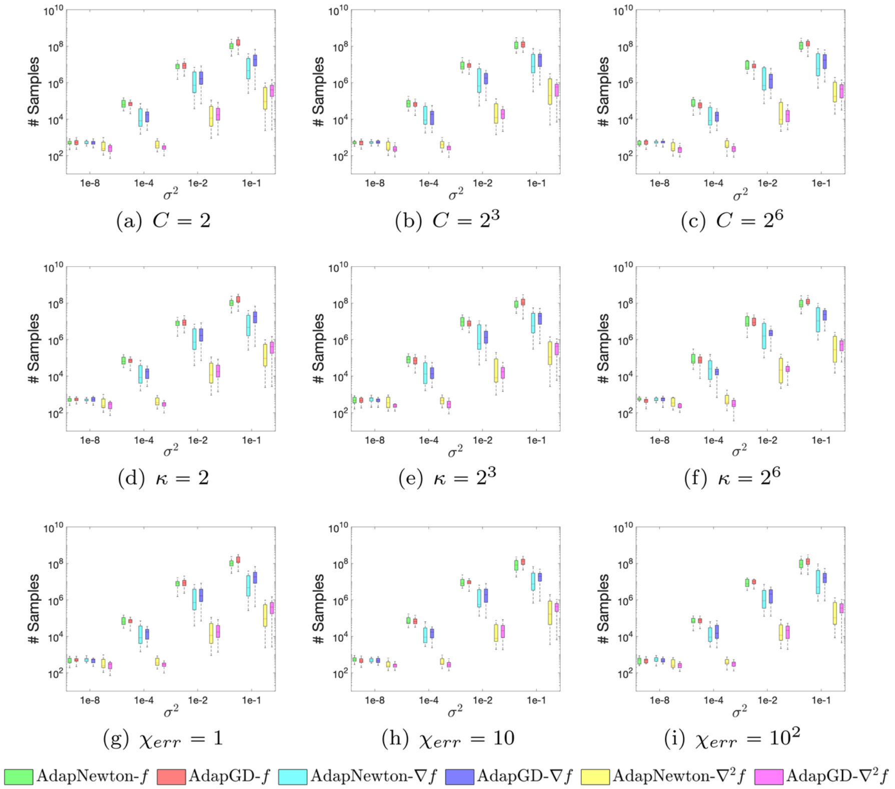
(d) $\kappa = 2$ (e) $\kappa = 2^3$ (f) $\kappa = 2^6$

(g) $\chi_{err} = 1$ (h) $\chi_{err} = 10$ (i) $\chi_{err} = 10^2$

Fig. 1 KKT residual boxplots. Each panel corresponds to a setup of (C, κ, χ_{err}). The default values are $C = \kappa = 2$ and $\chi_{err} = 1$. When we vary one parameter, the other two are set as default. Thus, the three figures on the left column are the same

powers of the reciprocal of the KKT residual $1/R_t$. When $\sigma^2 = 10^{-8}$, the small variance dominates the effect of $1/R_t$ so that all three quantities can be estimated with very few samples. When $\sigma^2 = 0.1$, the different dependencies of the sample sizes on $1/R_t$ are more evident. Overall, Fig. 2 reveals the fact that different objective quantities can be estimated with different amount of samples. Such an aspect improves the prior work [40], where the quantities with different sample complexities are estimated based on the same set of samples, and the effect of the variance σ^2 on the sample complexities is neglected.

In addition, we draw the trajectories of the sample size ratios. In particular, for both algorithms, we randomly pick 5 convergent problems and draw two ratio trajectories for each problem: one is the sample size of the gradient over the sample size of the value, and one is the sample size of the Hessian over the sample size of the gradient. We take $C = 64$ as an example. The plot is shown in Fig. 3. From the figure, we note that the sample size ratios tend to be stabilized at a small level, and the trend is more evident when $\sigma^2 = 0.1$. As we explained for Fig. 2 above, such an observation is consistent with our discussions in Sect. 3.4, and illustrates the improvement of our analysis over



(a) $C = 2$ (b) $C = 2^3$ (c) $C = 2^6$

(d) $\kappa = 2$ (e) $\kappa = 2^3$ (f) $\kappa = 2^6$

(g) $\chi_{err} = 1$ (h) $\chi_{err} = 10$ (i) $\chi_{err} = 10^2$

Fig. 2 Sample size boxplots. Each panel corresponds to a setup of (C, κ, χ_{err}). The default values are $C = \kappa = 2$ and $\chi_{err} = 1$. When we vary one parameter, the other two are set as default. Thus, the three figures on the left column are the same

[40] for performing the stochastic line search on the augmented Lagrangian merit function.

Stepsize trajectories Figure 4 plots the stepsize trajectories that are selected by stochastic line search. We take the default setup as an example, i.e., $C = \kappa = 2$, $\chi_{err} = 1$. Similar to Fig. 3, for each level of σ^2, we randomly pick 5 convergent problems to show the trajectories. Although there is no clear trend for the stepsize trajectories due to stochasticity, we clearly see for both methods that the stepsize can increase significantly from a very small value and even exceed 1. This exclusive property of the line search procedure ensures a fast convergence of the scheme, which is not enjoyed by many non-adaptive schemes where the stepsize often monotonically decays to zero.

We also examine some other aspects of the algorithm, such as the proportion of the iterations with failed SQP steps, with unstabilized penalty parameters, or with a triggered feasibility error condition (17). We also study the effect of a multiplicative noise, and implement the algorithm on an inequality constrained logistic regression problem. Due to the space limit, these auxiliary experiments are provided in Appendix D.

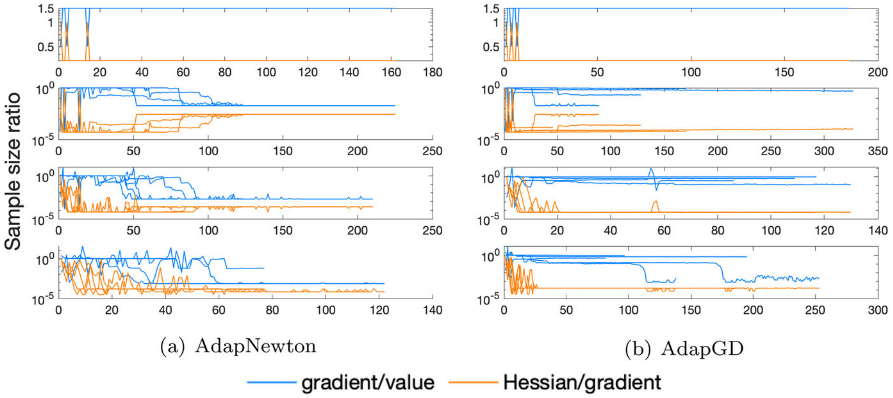


Fig. 3 Sample size ratio trajectories ($C = 64$). Each plot has four rows, from top to bottom, corresponding to $\sigma^2 = 10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}$. Each plot has ten lines with two colors. The five lines with the same color correspond to the five convergent problems

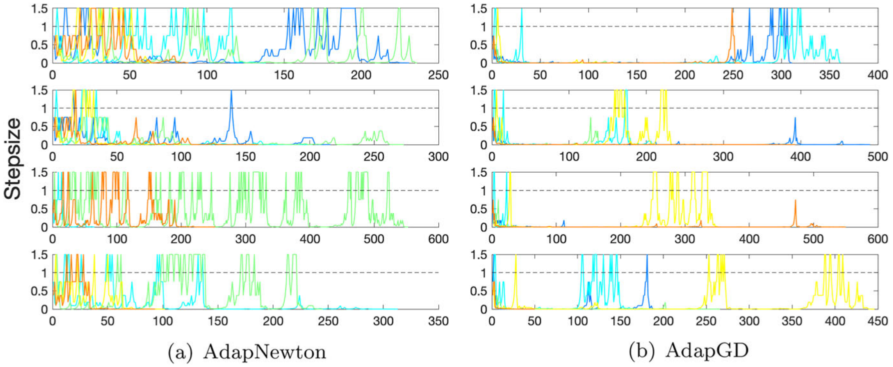


Fig. 4 Stepsize trajectories. Each plot has four rows, from top to bottom, corresponding to $\sigma^2 = 10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}$. Each plot has five lines, corresponding to the five problems

5 Conclusion

This paper studied inequality constrained stochastic nonlinear optimization problems. We designed an active-set StoSQP algorithm that exploits the exact augmented Lagrangian merit function. The algorithm adaptively selects the penalty parameters of the augmented Lagrangian, and selects the stepsize via stochastic line search. We proved that the KKT residuals converge to zero almost surely, which generalizes and strengthens the result for unconstrained and equality constrained problems in [40, 44] to enable wider applications.

The extension of this work includes studying more advanced StoSQP schemes. As mentioned in Sect. 3.5, the proposed StoSQP scheme has to solve the SQP system exactly. We note that, recently, [18] designed a StoSQP scheme where an inexact Newton direction is employed, and [3] designed a StoSQP scheme to relax LICQ condition. It is still open how to design related schemes to achieve relaxations with inequality constraints. In addition, some advanced SQP schemes solve inequality constrained problems by mixing IQP with EQP: one solves a *convex* IQP to obtain an active set, and then solves an EQP to obtain the search direction. See the "SQP+" scheme in [37] for example. Investigating this kind of mixed scheme with a stochastic objective is promising. Besides SQP, there are other classical methods for solving nonlinear problems that can be exploited to deal with stochastic objectives, such as the augmented Lagrangian methods and interior point methods. Different methods have different benefits and all of them deserve studying in the setup where the model can only be accessed with certain noise.

Finally, as mentioned in Sect. 3.2, non-asymptotic analysis and iteration complexity of the proposed scheme are missing in our global analysis. Further, it is known for deterministic setting that differentiable merit functions can overcome the Maratos effect and facilitate a fast local rate, while non-smooth merit functions (without advanced local modifications) cannot. This raises the questions: what is the local rate of the proposed StoSQP, and is the local rate better than the one using non-smooth merit functions? To answer these questions, we need a better understanding on the local behavior of stochastic line search. Such a local study would complement the established global analysis, recognize the benefits of the differentiable merit functions, and bridge the understanding gap between stochastic SQP and deterministic SQP.

Acknowledgements We thank Associate Editor and two anonymous reviewers for instructive comments, which help us further enhance the algorithm design and presentation. This material was completed in part with resources provided by the University of Chicago Research Computing Center. This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) under Contract DE-AC02-06CH11347 and by NSF through award CNS-1545046.

A Proofs of Sect. 2

A.1 Proof of Lemma 2

Throughout the proof, we denote $g^\star = g(\boldsymbol{x}^\star)$, $\boldsymbol{w}^\star_{\epsilon,\nu} = \boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, $\nabla\mathcal{L}^\star = \nabla\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)$ (similar for $c^\star$, $a^\star_\nu$, $q^\star_\nu$ etc.) to be the quantities evaluated at $(\boldsymbol{x}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star) \in \mathcal{T}_\nu \times \mathbb{R}^m \times \mathbb{R}^r$. Since $\boldsymbol{w}^\star_{\epsilon,\nu} = \boldsymbol{0}$, we know from Lemma 1 that $g^\star \leq \boldsymbol{0}$, $\boldsymbol{\lambda}^\star \geq \boldsymbol{0}$, $(\boldsymbol{\lambda}^\star)^T g^\star = 0$. This implies that $\mathrm{diag}^2(g^\star)\boldsymbol{\lambda}^\star = \boldsymbol{0}$. Furthermore, by $c^\star = \boldsymbol{0}$, $\boldsymbol{w}^\star_{\epsilon,\nu} = \boldsymbol{0}$, $a^\star_\nu, \eta, \epsilon > 0$, and $\nabla_{\boldsymbol{\mu},\boldsymbol{\lambda}}\mathcal{L}^\star_{\epsilon,\nu,\eta} = \boldsymbol{0}$, we obtain from (10) that

$$\begin{pmatrix} M^\star_{11} & M^\star_{12} \\ M^\star_{21} & M^\star_{22} \end{pmatrix} \begin{pmatrix} J^\star \\ G^\star \end{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}^\star = \boldsymbol{0}. \tag{A.1}$$

Recalling the definition of $M^\star$ in (9), we multiply the matrix $\nabla_{\boldsymbol{x}}^T\mathcal{L}^\star((J^\star)^T \quad (G^\star)^T)$ from the left and obtain

$$\begin{aligned}\mathbf{0} &\overset{(A.1)}{=} \nabla_{\boldsymbol{x}}^T\mathcal{L}^\star\left((J^\star)^T \;\; (G^\star)^T\right)\begin{pmatrix} J^\star(J^\star)^T & J^\star(G^\star)^T \\ G^\star(J^\star)^T & G^\star(G^\star)^T+\mathrm{diag}^2(g^\star)\end{pmatrix}\begin{pmatrix} J^\star \\ G^\star\end{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}^\star \\ &= \left\|\left((J^\star)^TJ^\star+(G^\star)^TG^\star\right)\nabla_{\boldsymbol{x}}\mathcal{L}^\star\right\|^2+\left\|\mathrm{diag}(g^\star)G^\star\nabla_{\boldsymbol{x}}\mathcal{L}^\star\right\|^2.\end{aligned}$$

This implies $\left((J^\star)^TJ^\star+(G^\star)^TG^\star\right)\nabla_{\boldsymbol{x}}\mathcal{L}^\star=\mathbf{0}$. Multiplying $\nabla_{\boldsymbol{x}}\mathcal{L}^\star$ from the left, we have $J^\star\nabla_{\boldsymbol{x}}\mathcal{L}^\star=\mathbf{0}$ and $G^\star\nabla_{\boldsymbol{x}}\mathcal{L}^\star=\mathbf{0}$. Plugging into (10) and noting that $\nabla_{\boldsymbol{x}}\mathcal{L}^\star_{\epsilon,\nu,\eta}=\mathbf{0}$, $\boldsymbol{w}^\star_{\epsilon,\nu}=\mathbf{0}$, $c^\star=\mathbf{0}$, $\mathrm{diag}^2(g^\star)\boldsymbol{\lambda}^\star=\mathbf{0}$, and $q^\star_\nu, a^\star_\nu, \epsilon>0$, we obtain $\nabla_{\boldsymbol{x}}\mathcal{L}^\star=\mathbf{0}$. This shows $(\boldsymbol{x}^\star,\boldsymbol{\mu}^\star,\boldsymbol{\lambda}^\star)$ satisfies (4), and we complete the proof.

A.2 Proof of Lemma 3

We require the following two preparation lemmas.

Lemma 12 *Let $\mathcal{I}(\boldsymbol{x}^\star)$ be the active set defined in (3), and $\mathcal{I}^+(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)=\left\{i\in\mathcal{I}(\boldsymbol{x}^\star):\boldsymbol{\lambda}_i^\star>0\right\}$. For any $\epsilon,\nu>0$, there exists a compact set $\mathcal{X}_{\epsilon,\nu}\times\Lambda_{\epsilon,\nu}\ni(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)$ depending on (ϵ,ν), such that*

$$\mathcal{I}^+(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)\subseteq\mathcal{A}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})\subseteq\mathcal{I}(\boldsymbol{x}^\star),\qquad\forall(\boldsymbol{x},\boldsymbol{\lambda})\in\mathcal{X}_{\epsilon,\nu}\times\Lambda_{\epsilon,\nu}.$$

Proof See Appendix A.3. □

Lemma 13 *Under Assumption 1, there exist a compact set $X\ni\boldsymbol{x}^\star$ and a constant $\gamma_H\in(0,1]$ such that $M(\boldsymbol{x})\succeq\gamma_HI$ for any $\boldsymbol{x}\in X$, where $M(\boldsymbol{x})$ is defined in (9). Furthermore, for any $\epsilon,\nu>0$, there exists a compact set $\mathcal{X}_{\epsilon,\nu}\times\Lambda_{\epsilon,\nu}\ni(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)$ depending on (ϵ,ν), such that*

$$\begin{pmatrix} J(\boldsymbol{x}) \\ G_{\mathcal{A}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})}(\boldsymbol{x})\end{pmatrix}\left(J(\boldsymbol{x})^T \quad G_{\mathcal{A}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})}(\boldsymbol{x})^T\right)\succeq\gamma_HI,\quad\forall(\boldsymbol{x},\boldsymbol{\lambda})\in\mathcal{X}_{\epsilon,\nu}\times\Lambda_{\epsilon,\nu}.$$

Proof See Appendix A.4. □

We now prove Lemma 3. We suppress the evaluation point and the iteration index t. Let $\mathcal{X}\times\mathcal{M}\times\Lambda\subseteq\mathcal{T}_\nu\times\mathbb{R}^m\times\mathbb{R}^r$ be any compact set around $(\boldsymbol{x}^\star,\boldsymbol{\mu}^\star,\boldsymbol{\lambda}^\star)$ (independent of ϵ,ν,η) and suppose $(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\lambda})\in\mathcal{X}\times\mathcal{M}\times\Lambda$. By Lemma 13, we know there exist a constant $\gamma_H\in(0,1]$ and, for any $\epsilon,\nu>0$, a compact subset $\mathcal{X}_{\epsilon,\nu}\times\Lambda_{\epsilon,\nu}\subseteq\mathcal{X}\times\Lambda$ such that for any point in the subset,

$$M\succeq\gamma_HI\quad\text{and}\quad\begin{pmatrix} J \\ G_a\end{pmatrix}\left(J^T \quad G_a^T\right)\succeq\gamma_HI. \tag{A.2}$$

Thus, by Assumption 2, we know from [41, Lemma 16.1] that K_a is invertible, and thus (12) is solvable. Furthermore, we can also show that (see [40, Lemma 1] for a simple proof)

$$\|K_a^{-1}\| \leq 7\Upsilon_B^2/(\gamma_B\gamma_H). \tag{A.3}$$

With the above two results, we conduct our analysis. Throughout the proof, we use $\Upsilon_1, \Upsilon_2 \ldots$ to denote generic upper bounds of functions evaluated in the set $\mathcal{X}\times\mathcal{M}\times\Lambda$, which are independent of $(\epsilon, \nu, \eta, \gamma_B, \gamma_H)$. As they are upper bounds, without loss of generality, $\Upsilon_i \geq 1, \forall i$.

We start from $(\nabla\mathcal{L}^{(1)}_{\epsilon,\nu,\eta})^T\Delta$ and suppose $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{X}_{\epsilon,\nu}\times\mathcal{M}\times\Lambda_{\epsilon,\nu} \subseteq \mathcal{X}\times\mathcal{M}\times\Lambda$, where $\mathcal{X}_{\epsilon,\nu}$ and $\Lambda_{\epsilon,\nu}$ come from Lemma 13. We have

$$\begin{aligned}
(\nabla\mathcal{L}^{(1)}_{\epsilon,\nu,\eta})^T\Delta \overset{(13)}{=} & \Delta\boldsymbol{x}^T\nabla_{\boldsymbol{x}}\mathcal{L} + \eta\Delta\boldsymbol{x}^T\begin{pmatrix}Q_1 & Q_2\end{pmatrix}\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix} + \frac{1}{\epsilon}\Delta\boldsymbol{x}^T J^T c \\
& + \frac{1}{\epsilon q_\nu}\Delta\boldsymbol{x}^T G^T\boldsymbol{w}_{\epsilon,\nu} + \begin{pmatrix}\Delta\boldsymbol{\mu}\\ \Delta\boldsymbol{\lambda}\end{pmatrix}^T\begin{pmatrix} c \\ \boldsymbol{w}_{\epsilon,\nu}\end{pmatrix} \\
& + \eta\begin{pmatrix}\Delta\boldsymbol{\mu}\\ \Delta\boldsymbol{\lambda}\end{pmatrix}^T\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22}\end{pmatrix}\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix} \\
\overset{(12b)}{=} & \Delta\boldsymbol{x}^T\nabla_{\boldsymbol{x}}\mathcal{L} + \frac{1}{\epsilon}\Delta\boldsymbol{x}^T J^T c + \frac{1}{\epsilon q_\nu}\Delta\boldsymbol{x}^T G^T\boldsymbol{w}_{\epsilon,\nu} + \begin{pmatrix}\Delta\boldsymbol{\mu}\\ \Delta\boldsymbol{\lambda}\end{pmatrix}^T\begin{pmatrix} c \\ \boldsymbol{w}_{\epsilon,\nu}\end{pmatrix} \\
& - \eta\left\|\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2 \\
\overset{\substack{(7)\\(11)}}{=} & \Delta\boldsymbol{x}^T(\nabla_{\boldsymbol{x}}\mathcal{L} - G_c^T\boldsymbol{\lambda}_c) + \frac{1}{\epsilon}\Delta\boldsymbol{x}^T J^T c + \frac{1}{\epsilon q_\nu}\Delta\boldsymbol{x}^T G_a^T g_a + \begin{pmatrix} c \\ g_a\end{pmatrix}^T\begin{pmatrix}\Delta\boldsymbol{\mu}\\ \Delta\boldsymbol{\lambda}_a\end{pmatrix} \\
& - \epsilon q_\nu\Delta\boldsymbol{\lambda}_c^T\boldsymbol{\lambda}_c - \eta\left\|\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2 \\
\overset{(12a)}{=} & -\Delta\boldsymbol{x}^T B\Delta\boldsymbol{x} + \begin{pmatrix} c \\ g_a\end{pmatrix}^T\begin{pmatrix}\tilde{\Delta}\boldsymbol{\mu} + \Delta\boldsymbol{\mu}\\ \tilde{\Delta}\boldsymbol{\lambda}_a + \Delta\boldsymbol{\lambda}_a\end{pmatrix} - \frac{1}{\epsilon}\|c\|^2 - \frac{1}{\epsilon q_\nu}\|g_a\|^2 - \epsilon q_\nu\Delta\boldsymbol{\lambda}_c^T\boldsymbol{\lambda}_c \\
& - \eta\left\|\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2.
\end{aligned} \tag{A.4}$$

Since $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{X}\times\mathcal{M}\times\Lambda$, there exists $\Upsilon_1 \geq 1$ such that $\|(Q_1\ Q_2)\| \leq \Upsilon_1$. Thus, we have

$$\begin{aligned}
\left\|\begin{pmatrix}\Delta\boldsymbol{\mu}\\ \Delta\boldsymbol{\lambda}\end{pmatrix}\right\| \overset{(12b)}{=} & \left\| M^{-1}\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix} + M^{-1}\begin{pmatrix} Q_1^T \\ Q_2^T\end{pmatrix}\Delta\boldsymbol{x}\right\| \\
\overset{(A.2)}{\leq} & \frac{1}{\gamma_H}\left\|\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\| + \frac{\Upsilon_1}{\gamma_H}\|\Delta\boldsymbol{x}\| \\
\leq & \frac{2\Upsilon_1}{\gamma_H}\left\|\begin{pmatrix}\Delta\boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\| \qquad (\text{since } 1 \leq \Upsilon_1).
\end{aligned} \tag{A.5}$$

Moreover, we note that

$$
\begin{aligned}
&\left\{\begin{pmatrix} J\\ G_a\\ G_c\end{pmatrix}\begin{pmatrix} J^T & G_a^T & G_c\end{pmatrix} + \begin{pmatrix}\mathbf{0}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\text{diag}^2(g_a)&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\text{diag}^2(g_c)\end{pmatrix}\right\}\begin{pmatrix}\tilde{\Delta}\boldsymbol{\mu}\\ \tilde{\Delta}\boldsymbol{\lambda}_a\\ -\boldsymbol{\lambda}_c\end{pmatrix}\\
&\overset{(12a)}{=} -\begin{pmatrix} J\\ G_a\\ G_c\end{pmatrix} B\Delta\boldsymbol{x} - \begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G_a\nabla_{\boldsymbol{x}}\mathcal{L}\\ G_c\nabla_{\boldsymbol{x}}\mathcal{L}+\text{diag}^2(g_c)\boldsymbol{\lambda}_c\end{pmatrix} + \begin{pmatrix}\mathbf{0}\\ \text{diag}^2(g_a)\tilde{\Delta}\boldsymbol{\lambda}_a\\ \mathbf{0}\end{pmatrix}\\
&\overset{(12a)}{=} -\begin{pmatrix} J\\ G_a\\ G_c\end{pmatrix} B\Delta\boldsymbol{x} - \begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G_a\nabla_{\boldsymbol{x}}\mathcal{L}\\ G_c\nabla_{\boldsymbol{x}}\mathcal{L}+\text{diag}^2(g_c)\boldsymbol{\lambda}_c\end{pmatrix} - \begin{pmatrix}\mathbf{0}\\ \text{diag}(g_a)\text{diag}(\tilde{\Delta}\boldsymbol{\lambda}_a)G_a\Delta\boldsymbol{x}\\ \mathbf{0}\end{pmatrix}.
\end{aligned}
$$

By $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{X}\times\mathcal{M}\times\Lambda$, there exist $\Upsilon_2, \Upsilon_3, \Upsilon_4 \geq 1$ such that

$$
\left\|\begin{pmatrix} J^T & G^T\end{pmatrix}\right\| \leq \Upsilon_2, \qquad \|\tilde{\Delta}\boldsymbol{\lambda}_a\| \overset{(12a)}{\leq} \left\| K_a^{-1}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L} - G_c^T\boldsymbol{\lambda}_c\\ c\\ g_a\end{pmatrix}\right\| \overset{(A.3)}{\leq} \frac{\Upsilon_3}{\gamma_H\gamma_B},
$$

and

$$
\|\text{diag}(g_a)\text{diag}(\tilde{\Delta}\boldsymbol{\lambda}_a)G_a\| \leq \frac{\Upsilon_4}{\gamma_H\gamma_B}.
$$

Combining the above three displays,

$$
\begin{aligned}
\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{\mu}\\ \tilde{\Delta}\boldsymbol{\lambda}_a\\ -\boldsymbol{\lambda}_c\end{pmatrix}\right\| &\overset{(A.2)}{\leq} \tfrac{1}{\gamma_H}\left(\Upsilon_2\|B\Delta\boldsymbol{x}\| + \tfrac{\Upsilon_4}{\gamma_H\gamma_B}\|\Delta\boldsymbol{x}\|\right) + \tfrac{1}{\gamma_H}\left\|\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|\\
&\leq \tfrac{\Upsilon_2\Upsilon_B+\Upsilon_4+1}{\gamma_H^2\gamma_B}\left\|\begin{pmatrix}\Delta\boldsymbol{x}\\ J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\| \quad (\text{since } \gamma_H\vee\gamma_B\leq 1),
\end{aligned}
\tag{A.6}
$$

where the second inequality also uses $\|B\| \leq \Upsilon_B$ by Assumption 2. Combining (A.4), (A.5), (A.6), and using $0 < q_\nu \leq \nu$ and $\gamma_H \vee \gamma_B \leq 1$,

$$
\begin{aligned}
&(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(1)})^T\Delta\\
&\overset{(A.4)}{\leq} -\Delta\boldsymbol{x}^T B\Delta\boldsymbol{x} + \left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|\left\{\left\|\begin{pmatrix}\tilde{\Delta}\boldsymbol{\mu}\\ \tilde{\Delta}\boldsymbol{\lambda}_a\end{pmatrix}\right\| + \left\|\begin{pmatrix}\Delta\boldsymbol{\mu}\\ \Delta\boldsymbol{\lambda}_a\end{pmatrix}\right\|\right\} - \frac{1}{\epsilon(1\vee\nu)}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|^2\\
&\quad + \epsilon\nu\|\Delta\boldsymbol{\lambda}_c\|\|\boldsymbol{\lambda}_c\| - \eta\left\|\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2\\
&\overset{\substack{(A.5)\\(A.6)}}{\leq} -\Delta\boldsymbol{x}^T B\Delta\boldsymbol{x} + \frac{2\Upsilon_1 + \Upsilon_2\Upsilon_B + \Upsilon_4 + 1}{\gamma_H^2\gamma_B}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|\left\|\begin{pmatrix}\Delta\boldsymbol{x}\\ J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{\epsilon(1\vee\nu)}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|^2\\
&+\epsilon\nu\cdot\frac{2\Upsilon_1(\Upsilon_2\Upsilon_B+\Upsilon_4+1)}{\gamma_H^3\gamma_B}\left\|\begin{pmatrix}\Delta\boldsymbol{x}\\ J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2\\
&-\eta\left\|\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2\\
&\le -\Delta\boldsymbol{x}^T B\Delta\boldsymbol{x}+\frac{\Upsilon_5}{\gamma_H^2\gamma_B}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|\left\|\begin{pmatrix}\Delta\boldsymbol{x}\\ J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|-\frac{1}{\epsilon(1\vee\nu)}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|^2\\
&+\frac{\epsilon\nu\Upsilon_5}{\gamma_H^3\gamma_B}\left\|\begin{pmatrix}\Delta\boldsymbol{x}\\ J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2-\eta\left\|\begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2, \qquad \text{(A.7)}
\end{aligned}
$$

where the last inequality holds by defining

$$\Upsilon_5 = 2\Upsilon_1+\Upsilon_2\Upsilon_B+\Upsilon_4+1\vee 2\Upsilon_1(\Upsilon_2\Upsilon_B+\Upsilon_4+1).$$

To deal with $\Delta\boldsymbol{x}^T B\Delta\boldsymbol{x}$ in (A.7), we decompose $\Delta\boldsymbol{x}$ as $\Delta\boldsymbol{x} = \Delta\boldsymbol{u}+\Delta\boldsymbol{v}$ where $\Delta\boldsymbol{u}\in$ Image $\left\{(J^T \;\; G_a^T)\right\}$ and $\Delta\boldsymbol{v}\in$ Ker $\left\{(J^T \;\; G_a^T)^T\right\}$. Note that

$$
\begin{aligned}
-\begin{pmatrix} c\\ g_a\end{pmatrix}=\begin{pmatrix} J\\ G_a\end{pmatrix}\Delta\boldsymbol{x}=\begin{pmatrix} J\\ G_a\end{pmatrix}\Delta\boldsymbol{u}\Longrightarrow\Delta\boldsymbol{\mu}=-\left(J^T \quad G_a^T\right)\left\{\begin{pmatrix} J\\ G_a\end{pmatrix}(J^T \; G_a^T)\right\}^{-1}\begin{pmatrix} c\\ g_a\end{pmatrix}\\
\overset{\text{(A.2)}}{\Longrightarrow}\|\Delta\boldsymbol{\mu}\|\le\frac{1}{\sqrt{\gamma_H}}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|. \qquad \text{(A.8)}
\end{aligned}
$$

Thus, by Assumption 2,

$$
\begin{aligned}
&-\Delta\boldsymbol{x}^T B\Delta\boldsymbol{x}\\
&\quad=-\Delta\boldsymbol{v}^T B\Delta\boldsymbol{v}-2\Delta\boldsymbol{u}^T B\Delta\boldsymbol{v}-\Delta\boldsymbol{u}^T B\Delta\boldsymbol{u}\le-\gamma_B\|\Delta\boldsymbol{v}\|^2+2\Upsilon_B\|\Delta\boldsymbol{v}\|\|\Delta\boldsymbol{u}\|+\Upsilon_B\|\Delta\boldsymbol{u}\|^2\\
&\quad\le-\frac{3\gamma_B}{4}\|\Delta\boldsymbol{v}\|^2+(\Upsilon_B+\frac{4\Upsilon_B^2}{\gamma_B})\|\Delta\boldsymbol{u}\|^2=-\frac{3\gamma_B}{4}\|\Delta\boldsymbol{x}\|^2+(\Upsilon_B+\frac{4\Upsilon_B^2}{\gamma_B}+\frac{3\gamma_B}{4})\|\Delta\boldsymbol{u}\|^2\\
&\quad\overset{\text{(A.8)}}{\le}-\frac{3\gamma_B}{4}\|\Delta\boldsymbol{x}\|^2+(\Upsilon_B+\frac{4\Upsilon_B^2}{\gamma_B}+\frac{3\gamma_B}{4})\frac{1}{\gamma_H}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|^2\\
&\quad\le-\frac{3\gamma_B}{4}\|\Delta\boldsymbol{x}\|^2+\frac{\Upsilon_6}{\gamma_H\gamma_B}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|^2, \qquad \text{(A.9)}
\end{aligned}
$$

where the last inequality holds with $\Upsilon_6 = \Upsilon_B+4\Upsilon_B^2+1$ by noting that $\gamma_B\le 1$. Combining the above display with (A.7) and using the following Young's inequality,

$$\frac{\Upsilon_5}{\gamma_H^2\gamma_B}\left\|\begin{pmatrix} c\\ g_a\end{pmatrix}\right\|\left\|\begin{pmatrix}\Delta\boldsymbol{x}\\ J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|$$

$$\leq \left(\frac{\gamma_B}{8} \wedge \frac{\eta}{4}\right) \left\| \begin{pmatrix} \Delta \boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\|^2 + \frac{2\Upsilon_5^2}{\gamma_H^4\gamma_B^2(\gamma_B \wedge \eta)} \left\| \begin{pmatrix} c \\ g_a \end{pmatrix} \right\|^2,$$

we have

$$\begin{aligned}(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(1)})^T\Delta &\leq -\frac{3\gamma_B}{4}\|\Delta\boldsymbol{x}\|^2 + \left\{\left(\frac{\gamma_B}{8} \wedge \frac{\eta}{4}\right) + \frac{\epsilon\nu\Upsilon_5}{\gamma_H^3\gamma_B}\right\} \left\| \begin{pmatrix} \Delta \boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\|^2 \\ &\quad + \left\{\frac{\Upsilon_6}{\gamma_H\gamma_B} + \frac{2\Upsilon_5^2}{\gamma_H^4\gamma_B^2(\gamma_B \wedge \eta)} - \frac{1}{\epsilon(1 \vee \nu)}\right\} \left\| \begin{pmatrix} c \\ g_a \end{pmatrix} \right\|^2 - \eta \left\| \begin{pmatrix} J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\|^2 \\ &\leq -\left\{\frac{\gamma_B \wedge \eta}{2} + \left(\frac{\gamma_B}{8} \wedge \frac{\eta}{4}\right) - \frac{\epsilon\nu\Upsilon_5}{\gamma_H^3\gamma_B}\right\} \left\| \begin{pmatrix} \Delta \boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\|^2 \\ &\quad - \left\{\frac{1}{\epsilon(1 \vee \nu)} - \frac{\Upsilon_6}{\gamma_H\gamma_B} - \frac{2\Upsilon_5^2}{\gamma_H^4\gamma_B^2(\gamma_B \wedge \eta)}\right\} \left\| \begin{pmatrix} c \\ g_a \end{pmatrix} \right\|^2.\end{aligned}$$

Therefore, as long as

$$\frac{\gamma_B}{8} \wedge \frac{\eta}{4} \geq \frac{\epsilon\nu\Upsilon_5}{\gamma_H^3\gamma_B} \Longleftarrow \frac{1}{\epsilon} \geq \frac{8\nu\Upsilon_5}{\gamma_H^3\gamma_B(\gamma_B \wedge \eta)}, \tag{A.10a}$$

$$\frac{1}{\epsilon(1 \vee \nu)} - \frac{\Upsilon_6}{\gamma_H\gamma_B} - \frac{2\Upsilon_5^2}{\gamma_H^4\gamma_B^2(\gamma_B \wedge \eta)} \geq 0 \Longleftarrow \frac{1}{\epsilon} \geq \frac{(1 \vee \nu)(2\Upsilon_5^2 + \Upsilon_6)}{\gamma_H^4\gamma_B^2(\gamma_B \wedge \eta)}, \tag{A.10b}$$

we have

$$(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(1)})^T\Delta \leq -\frac{\gamma_B \wedge \eta}{2} \left\| \begin{pmatrix} \Delta \boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\|^2.$$

Thus, letting $\Upsilon = \left\{8\Upsilon_5 \vee (2\Upsilon_5^2 + \Upsilon_6)\right\}/\gamma_H^4$ and noting that (A.10a) is implied by (A.10b), we complete the first part of the statement.

We now prove the second part of the statement. By (13), $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{X} \times \mathcal{M} \times \Lambda$ (and hence (A.12)), and the fact that $a_\nu \geq \nu/2$, there exists $\Upsilon_7 \geq 1$ such that

$$\begin{aligned}(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(2)})^T\Delta &\overset{(13)}{=} \frac{3\|\boldsymbol{w}_{\epsilon,\nu}\|^2}{2\epsilon q_\nu a_\nu}\Delta\boldsymbol{x}^T G^T \boldsymbol{l} + \eta\Delta\boldsymbol{x}^T Q_{2,a}\text{diag}^2(g_a)\boldsymbol{\lambda}_a + \frac{\|\boldsymbol{w}_{\epsilon,\nu}\|^2}{\epsilon a_\nu}\Delta\boldsymbol{\lambda}^T\boldsymbol{\lambda} \\ &\quad + \eta(\Delta\boldsymbol{\mu}^T \;\; \Delta\boldsymbol{\lambda}^T)\begin{pmatrix} M_{12,a} \\ M_{22,a} \end{pmatrix} \text{diag}^2(g_a)\boldsymbol{\lambda}_a \\ &\leq \Upsilon_7\left\{\frac{1}{\epsilon\nu^2}\left(\|g_a\|^2 + \epsilon^2\nu^2\|\boldsymbol{\lambda}_c\|^2\right)\|\Delta\boldsymbol{x}\| + \eta\|g_a\|^2\|\Delta\boldsymbol{x}\|\right. \\ &\quad \left. + \frac{1}{\epsilon\nu}\left(\|g_a\|^2 + \epsilon^2\nu^2\|\boldsymbol{\lambda}_c\|^2\right)\|\Delta\boldsymbol{\lambda}\| + \eta\|g_a\|^2\|(\Delta\boldsymbol{\mu}, \Delta\boldsymbol{\lambda})\|\right\}.\end{aligned}$$

Since $\epsilon \le 1$ by (A.10) (noting that $\Upsilon \ge 1 \ge \gamma_H \vee \gamma_B$), we simplify the above display by

$$\begin{aligned}
&(\nabla \mathcal{L}_{\epsilon,\nu,\eta}^{(2)})^T \Delta \\
&\le \Upsilon_7 \left\{ \frac{1 \vee \nu^2}{\epsilon\nu(1 \wedge \nu)} (\|g_a\|^2 + \|\boldsymbol{\lambda}_c\|^2)(\|\Delta \boldsymbol{x}\| + \|\Delta\boldsymbol{\lambda}\|) + \sqrt{2}\eta\|g_a\|^2\|(\Delta\boldsymbol{x}, \Delta\boldsymbol{\mu}, \Delta\boldsymbol{\lambda})\| \right\} \\
&\le \sqrt{2}\Upsilon_7 \left\{ \frac{1 \vee \nu^2}{\epsilon\nu(1 \wedge \nu)} (\|g_a\|^2 + \|\boldsymbol{\lambda}_c\|^2)\|(\Delta\boldsymbol{x}, \Delta\boldsymbol{\lambda})\| + \eta\|g_a\|^2\|(\Delta\boldsymbol{x}, \Delta\boldsymbol{\mu}, \Delta\boldsymbol{\lambda})\| \right\} \\
&\le 2\sqrt{2}\Upsilon_7 \left(\frac{1 \vee \nu}{\epsilon(1 \wedge \nu^2)} \vee \eta \right) (\|g_a\|^2 + \|\boldsymbol{\lambda}_c\|^2)\|(\Delta\boldsymbol{x}, \Delta\boldsymbol{\mu}, \Delta\boldsymbol{\lambda})\|.
\end{aligned}$$

Noting that

$$\begin{aligned}
\left\| \begin{pmatrix} \Delta\boldsymbol{x} \\ \Delta\boldsymbol{\mu} \\ \Delta\boldsymbol{\lambda} \end{pmatrix} \right\| &\le \|\Delta\boldsymbol{x}\| + \left\| \begin{pmatrix} \Delta\boldsymbol{\mu} \\ \Delta\boldsymbol{\lambda} \end{pmatrix} \right\| \overset{(A.5)}{\le} \|\Delta\boldsymbol{x}\| + \frac{2\Upsilon_1}{\gamma_H} \left\| \begin{pmatrix} \Delta\boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\| \\
&\le \frac{3\Upsilon_1}{\gamma_H} \left\| \begin{pmatrix} \Delta\boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\| \qquad (\text{since } \gamma_H \le 1 \le \Upsilon_1),
\end{aligned}$$

and

$$\begin{aligned}
\left\| \begin{pmatrix} g_a \\ \boldsymbol{\lambda}_c \end{pmatrix} \right\| &\le \|g_a\| + \|\boldsymbol{\lambda}_c\| \overset{\substack{(12a)\\(A.6)}}{\le} \Upsilon_2\|\Delta\boldsymbol{x}\| + \frac{\Upsilon_2\Upsilon_B + \Upsilon_4 + 1}{\gamma_H^2\gamma_B} \left\| \begin{pmatrix} \Delta\boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\| \\
&\le \frac{\Upsilon_2(\Upsilon_B + 1) + \Upsilon_4 + 1}{\gamma_H^2\gamma_B} \left\| \begin{pmatrix} \Delta\boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\| \qquad (\text{since } \gamma_H \vee \gamma_B \le 1),
\end{aligned}$$

we define $\Upsilon_8 = 6\sqrt{2}\Upsilon_7\Upsilon_1(\Upsilon_2(\Upsilon_B + 1) + \Upsilon_4 + 1)$ and have

$$(\nabla \mathcal{L}_{\epsilon,\nu,\eta}^{(2)})^T \Delta \le \frac{\Upsilon_8}{\gamma_H^3\gamma_B} \left(\frac{1 \vee \nu}{\epsilon(1 \wedge \nu^2)} \vee \eta \right) (\|g_a\| + \|\boldsymbol{\lambda}_c\|) \left\| \begin{pmatrix} \Delta\boldsymbol{x} \\ J\nabla_{\boldsymbol{x}}\mathcal{L} \\ G\nabla_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix} \right\|^2. \tag{A.11}$$

By Lemma 12, we can find a compact subset of $\mathcal{X}_{\epsilon,\nu} \times \Lambda_{\epsilon,\nu}$ depending only on (ϵ, ν) such that $\mathcal{A}_{\epsilon,\nu} \subseteq \mathcal{I}(\boldsymbol{x}^\star)$ and $\mathcal{A}_{\epsilon,\nu}^c \subseteq \{\mathcal{I}^+(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)\}^c$; thus

$$\|g_a\| \le \|g_{\mathcal{I}(\boldsymbol{x}^\star)}\| \qquad \text{and} \qquad \|\boldsymbol{\lambda}_c\| \le \|\boldsymbol{\lambda}_{(\mathcal{I}^+(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star))^c}\|.$$

Furthermore, we let $\mathcal{X}_{\epsilon,\nu,\eta} \times \Lambda_{\epsilon,\nu,\eta} \subseteq \mathcal{X}_{\epsilon,\nu} \times \Lambda_{\epsilon,\nu}$ be a compact subset depending additionally on η, such that

$$\|g_{\mathcal{I}(\boldsymbol{x}^\star)}\| \le \frac{\gamma_H^3\gamma_B}{\Upsilon_8} \left(\frac{\epsilon(1 \wedge \nu^2)}{1 \vee \nu} \wedge \frac{1}{\eta} \right) \frac{\gamma_B \wedge \eta}{8},$$

$$\|\boldsymbol{\lambda}_{(\mathcal{I}^+(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star))^c}\| \le \frac{\gamma_H^3\gamma_B}{\Upsilon_8}\left(\frac{\epsilon(1\wedge\nu^2)}{1\vee\nu}\wedge\frac{1}{\eta}\right)\frac{\gamma_B\wedge\eta}{8}.$$

Then, combining (A.11) with the above two displays leads to

$$(\nabla\mathcal{L}_{\epsilon,\nu,\eta}^{(2)})^T\Delta \le \frac{\gamma_B\wedge\eta}{4}\left\|\begin{pmatrix}\Delta\boldsymbol{x}\\ J\nabla_{\boldsymbol{x}}\mathcal{L}\\ G\nabla_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2.$$

This completes the proof.

A.3 Proof of Lemma 12

Let $\mathcal{X}\times\Lambda\subseteq\mathcal{T}_\nu\times\mathbb{R}^r$ be any compact set around $(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)$. For any $(\boldsymbol{x},\boldsymbol{\lambda})\in\mathcal{X}\times\Lambda$, we have

$$q_\nu(\boldsymbol{x},\boldsymbol{\lambda}) \overset{(6)}{\ge} \frac{\nu}{2}\cdot\frac{1}{1+\max_{\boldsymbol{\lambda}\in\Lambda}\|\boldsymbol{\lambda}\|^2} =: \kappa_\nu. \tag{A.12}$$

For any $i\in\mathcal{I}^+(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)$, we know $g_i^\star=0$ and $\boldsymbol{\lambda}_i^\star>0$. Thus, $g_i^\star+\epsilon\kappa_\nu\boldsymbol{\lambda}_i^\star>0$. Consider the ball $\mathcal{B}_i^{\boldsymbol{x}}=\{\boldsymbol{x}:\|\boldsymbol{x}-\boldsymbol{x}^\star\|\le r_i\}\cap\mathcal{X}$ and $\mathcal{B}_i^{\boldsymbol{\lambda}}=\{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^\star\|\le r_i\}\cap\Lambda$. For a sufficiently small r_i (depending on ϵ and ν), we have $(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)\in\mathcal{B}_i^{\boldsymbol{x}}\times\mathcal{B}_i^{\boldsymbol{\lambda}}\subseteq\mathcal{X}\times\Lambda$ and, for any $(\boldsymbol{x},\boldsymbol{\lambda})\in\mathcal{B}_i^{\boldsymbol{x}}\times\mathcal{B}_i^{\boldsymbol{\lambda}}$,

$$g_i(\boldsymbol{x})\ge-\epsilon\kappa_\nu\boldsymbol{\lambda}_i\overset{(A.12)}{\ge}-\epsilon q_\nu(\boldsymbol{x},\boldsymbol{\lambda})\boldsymbol{\lambda}_i.$$

The first inequality is due to the continuity of g_i. This implies $i\in\mathcal{A}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})$. Therefore, for any $(\boldsymbol{x},\boldsymbol{\lambda})$ in the compact set $\cap_{i\in\mathcal{I}^+(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)}\mathcal{B}_i^{\boldsymbol{x}}\times\mathcal{B}_i^{\boldsymbol{\lambda}}$, we have $\mathcal{I}^+(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)\subseteq\mathcal{A}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})$. The argument $\mathcal{A}_{\epsilon,\nu}(\boldsymbol{x},\boldsymbol{\lambda})\subseteq\mathcal{I}(\boldsymbol{x}^\star)$ can be proved in the same way.

A.4 Proof of Lemma 13

By Assumption 1, there exists a compact set $X\ni\boldsymbol{x}^\star$ small enough such that $(J^T(\boldsymbol{x})\;\; G_{\mathcal{I}(\boldsymbol{x}^\star)}^T(\boldsymbol{x}))$ has full column rank for all $\boldsymbol{x}\in X$. Furthermore, for any $(\boldsymbol{a},\boldsymbol{b})\in\mathbb{R}^{m+r}$, we note that

$$\begin{aligned}0=(\boldsymbol{a}^T\;\;\boldsymbol{b}^T)M(\boldsymbol{x})\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}&\Longrightarrow\boldsymbol{b}_{\mathcal{I}^c(\boldsymbol{x}^\star)}=\boldsymbol{0}\\ &\Longrightarrow\|J^T(\boldsymbol{x})\boldsymbol{a}+G_{\mathcal{I}(\boldsymbol{x}^\star)}^T(\boldsymbol{x})\boldsymbol{b}_{\mathcal{I}(\boldsymbol{x}^\star)}\|=0\Longrightarrow(\boldsymbol{a},\boldsymbol{b})=\boldsymbol{0},\end{aligned}\tag{A.13}$$

where the first implication is due to $\mathrm{diag}(g(\boldsymbol{x}))\boldsymbol{b}=0$ and $\mathcal{I}^c(\boldsymbol{x}^\star)\subseteq\mathcal{I}^c(\boldsymbol{x})$ (since X is small), and the second implication is due to $\|J^T(\boldsymbol{x})\boldsymbol{a}+G^T(\boldsymbol{x})\boldsymbol{b}\|=0$. Therefore,

$M(\boldsymbol{x})$ is invertible. Moreover, for any $\mathcal{A} \subseteq \mathcal{I}(\boldsymbol{x}^\star)$, we have

$$\sigma_{\min}\left\{\begin{pmatrix} J(\boldsymbol{x}) \\ G_{\mathcal{A}}(\boldsymbol{x}) \end{pmatrix}\left(J^T(\boldsymbol{x})\ G_{\mathcal{A}}^T(\boldsymbol{x})\right)\right\} \geq \sigma_{\min}\left\{\begin{pmatrix} J(\boldsymbol{x}) \\ G_{\mathcal{I}(\boldsymbol{x}^\star)}(\boldsymbol{x}) \end{pmatrix}\left(J^T(\boldsymbol{x})\ G_{\mathcal{I}(\boldsymbol{x}^\star)}^T(\boldsymbol{x})\right)\right\} > 0, \tag{A.14}$$

where $\sigma_{\min}(\cdot)$ denotes the least singular value of a matrix. By (A.13), (A.14), and the compactness of X, we know that there exists $\gamma_H \in (0, 1]$ such that

$$M(\boldsymbol{x}) \succeq \gamma_H I, \quad \begin{pmatrix} J(\boldsymbol{x}) \\ G_{\mathcal{A}}(\boldsymbol{x}) \end{pmatrix}\left(J^T(\boldsymbol{x})\ G_{\mathcal{A}}^T(\boldsymbol{x})\right) \succeq \gamma_H I, \quad \forall \boldsymbol{x} \in X \text{ and } \mathcal{A} \subseteq \mathcal{I}(\boldsymbol{x}^\star). \tag{A.15}$$

To show the second part of the statement, we apply Lemma 12, and know that there exists a compact set $\mathcal{X}_{\epsilon,\nu} \times \Lambda_{\epsilon,\nu} \subseteq X \times \mathbb{R}^r$ such that $\mathcal{A}(\boldsymbol{x}, \boldsymbol{\lambda}) \subseteq \mathcal{I}(\boldsymbol{x}^\star)$, $\forall(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathcal{X}_{\epsilon,\nu} \times \Lambda_{\epsilon,\nu}$. Combining this fact with (A.15), we complete the proof.

B Proofs of Sect. 3

B.1 Proof of Lemma 4

It suffices to show that there exists a threshold $\tilde{\epsilon} > 0$ such that for any samples ξ_1, any parameter $\nu \in [\bar{\nu}_0, \tilde{\nu}]$, where $\bar{\nu}_0$ is the fixed initial input of Algorithm 1 and $\tilde{\nu}$ is defined in (30), and any point $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{X} \times \mathcal{M} \times \Lambda$ with $\boldsymbol{x} \in \mathcal{T}_\nu$, if $\epsilon \leq \tilde{\epsilon}$, then

$$\left\|\left(c(\boldsymbol{x}), \boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x}, \boldsymbol{\lambda})\right)\right\| \leq \chi_{err} \cdot \left\|\bar{\nabla}\mathcal{L}_{\epsilon,\nu,\eta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})\right\|,$$

where $\bar{\nabla}\mathcal{L}_{\epsilon,\nu,\eta}$ is computed using samples in ξ_1 and η, $\chi_{err} > 0$ are any given positive constants. Note that everything above is deterministic; that is, our analysis does not depend on a specific iteration sequence $\{(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)\}_t$. Thus, the threshold $\tilde{\epsilon}$ is deterministic. Let us prove the above statement by contradiction. Without loss of generality, we suppose $\chi_{err} \leq 1$.

Suppose the statement is false, then there exist a sequence $\{\epsilon_j, \xi_1^j, \nu_j\}_j$ and an evaluation point sequence $\{(\boldsymbol{x}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j)\}_j \in \mathcal{X} \times \mathcal{M} \times \Lambda$ such that $\nu_j \in [\bar{\nu}_0, \tilde{\nu}]$, $\boldsymbol{x}_j \in \mathcal{T}_{\nu_j}$, $\epsilon_j \searrow 0$ and

$$\|\bar{\nabla}\mathcal{L}_{\epsilon_j,\nu_j,\eta}^j\| < 1/\chi_{err} \cdot \|(c_j, \boldsymbol{w}_{\epsilon_j,\nu_j}^j)\|, \qquad \forall j \geq 0, \tag{B.1}$$

where $\bar{\nabla}\mathcal{L}_{\epsilon_j,\nu_j,\eta}^j$ is computed using samples ξ_1^j, and η and χ_{err} are fixed constants. By the compactness condition, we suppose $(\boldsymbol{x}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j) \to (\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\lambda}}) \in \mathcal{X} \times \mathcal{M} \times \Lambda$ and $\nu_j \to \nu$ as $j \to \infty$ (otherwise, we can consider a convergent subsequence, which must exist). Noting that $c_j = c(\boldsymbol{x}_j)$ and $\boldsymbol{w}_{\epsilon_j,\nu_j}^j = \max\{g(\boldsymbol{x}_j), -\epsilon_j q_{\nu_j}(\boldsymbol{x}_j, \boldsymbol{\lambda}_j)\boldsymbol{\lambda}_j\}$ are bounded due to the compactness of $(\boldsymbol{x}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j)$ and the boundedness of ν_j and ϵ_j, we have from (B.1) that

$$\epsilon_j \|\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}^j_{\epsilon_j,\nu_j,\eta}\| \to 0 \quad \text{as} \quad j \to \infty. \tag{B.2}$$

Moreover, since $\boldsymbol{x}_j \in \mathcal{T}_{\nu_j}$, we have $\sum_{i=1}^r \max\{(g_j)_i, 0\}^3 \le \nu_j/2$. Taking limit $j \to \infty$ leads to $\tilde{\boldsymbol{x}} \in \mathcal{T}_\nu$. Furthermore, by (10), (B.2), and the convergence of $(\boldsymbol{x}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j)$, we get

$$J^T(\tilde{\boldsymbol{x}})c(\tilde{\boldsymbol{x}}) + \frac{1}{q_\nu(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}})} G^T(\tilde{\boldsymbol{x}}) \max\{g(\tilde{\boldsymbol{x}}), \boldsymbol{0}\} + \frac{3\|\max\{g(\tilde{\boldsymbol{x}}), \boldsymbol{0}\}\|^2}{2q_\nu(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}})a_\nu(\tilde{\boldsymbol{x}})} G(\tilde{\boldsymbol{x}})^T \boldsymbol{l}(\tilde{\boldsymbol{x}}) = \boldsymbol{0},$$

which is further simplified as

$$\sum_{i:c_i(\tilde{\boldsymbol{x}})\neq 0} c_i(\tilde{\boldsymbol{x}})\nabla c_i(\tilde{\boldsymbol{x}}) + \sum_{i:g_i(\tilde{\boldsymbol{x}})>0} \left\{ \frac{1}{q_\nu(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}})} + \frac{3\|\max\{g(\tilde{\boldsymbol{x}}), \boldsymbol{0}\}\|^2 g_i(\tilde{\boldsymbol{x}})}{2q_\nu(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}})a_\nu(\tilde{\boldsymbol{x}})} \right\} g_i(\tilde{\boldsymbol{x}})\nabla g_i(\tilde{\boldsymbol{x}}) = \boldsymbol{0}. \tag{B.3}$$

Suppose $\tilde{\boldsymbol{x}} \in \mathcal{X}\backslash\Omega$ and let $\mathcal{I}_c(\tilde{\boldsymbol{x}}) = \{i : 1 \le i \le m, c_i(\tilde{\boldsymbol{x}}) \neq 0\}$, and $\mathcal{I}_g(\tilde{\boldsymbol{x}}) = \{i : 1 \le i \le r, g_i(\tilde{\boldsymbol{x}}) > 0\}$. By Assumption 4, the set

$$\left\{ \boldsymbol{z} \in \mathbb{R}^d : c_i(\tilde{\boldsymbol{x}})\nabla^T c_i(\tilde{\boldsymbol{x}})\boldsymbol{z} < 0, i \in \mathcal{I}_c(\tilde{\boldsymbol{x}}) \text{ and } \nabla^T g_i(\tilde{\boldsymbol{x}})\boldsymbol{z} < 0, i \in \mathcal{I}_g(\tilde{\boldsymbol{x}}) \right\}$$

is nonempty. By the Gordan's theorem [26], for any $a_i, b_i \ge 0$ such that

$$\sum_{i\in\mathcal{I}_c(\tilde{\boldsymbol{x}})} a_i c_i(\tilde{\boldsymbol{x}})\nabla c_i(\tilde{\boldsymbol{x}}) + \sum_{i\in\mathcal{I}_g(\tilde{\boldsymbol{x}})} b_i \nabla g_i(\tilde{\boldsymbol{x}}) = \boldsymbol{0}, \tag{B.4}$$

we have $a_i = b_i = 0$. Comparing (B.4) with (B.3), and noting that the coefficients of (B.3) are all positive (since $\tilde{\boldsymbol{x}} \in \mathcal{T}_\nu$), we immediately get the contradiction. Thus, $\tilde{\boldsymbol{x}} \in \Omega$.

By Assumption 4 and following the same reasoning as (A.13), $M(\tilde{\boldsymbol{x}})$ is invertible and, particularly, is positive definite. Thus, M_j is invertible for large enough j. Let us suppose $\|M_j^{-1}\| \le \Upsilon_M$ for some $\Upsilon_M > 0$. Further, by direct calculation, we have

$$\operatorname{diag}(g_j)\boldsymbol{\lambda}_j = \operatorname{diag}(\boldsymbol{\lambda}_j)\boldsymbol{w}^j_{\epsilon_j,\nu_j} - \frac{1}{\epsilon_j q^j_{\nu_j}}(\operatorname{diag}(g_j) - \operatorname{diag}(\boldsymbol{w}^j_{\epsilon_j,\nu_j}))\boldsymbol{w}^j_{\epsilon_j,\nu_j}. \tag{B.5}$$

Thus, we can obtain

$$\begin{pmatrix} J_j \\ G_j \end{pmatrix} \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}^j_{\epsilon_j,\nu_j,\eta} \overset{(10)}{=} \begin{pmatrix} J_j \\ G_j \end{pmatrix} \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j + \eta \begin{pmatrix} J_j \\ G_j \end{pmatrix} \begin{pmatrix} Q_{1,j} & Q_{2,j} \end{pmatrix} \begin{pmatrix} J_j \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j \\ G_j \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j + \operatorname{diag}^2(g_j)\boldsymbol{\lambda}_j \end{pmatrix}$$
$$+ \frac{1}{\epsilon_j} \begin{pmatrix} J_j \\ G_j \end{pmatrix} \begin{pmatrix} J_j^T & \frac{G_j^T}{q^j_{\nu_j}} + \frac{3G_j^T \boldsymbol{l}_j(\boldsymbol{w}^j_{\epsilon_j,\nu_j})^T}{2q^j_{\nu_j} a^j_{\nu_j}} \end{pmatrix} \begin{pmatrix} c_j \\ \boldsymbol{w}^j_{\epsilon_j,\nu_j} \end{pmatrix}$$

$$
\begin{aligned}
&\overset{(B.5)}{=} \left\{ I + \eta \begin{pmatrix} J_j \\ G_j \end{pmatrix} \begin{pmatrix} Q_{1,j} & Q_{2,j} \end{pmatrix} \right\} \begin{pmatrix} J_j \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j \\ G_j \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j + \operatorname{diag}^2(g_j) \boldsymbol{\lambda}_j \end{pmatrix} \\
&\quad + \frac{1}{\epsilon_j} \left\{ \begin{pmatrix} J_j \\ G_j \end{pmatrix} \begin{pmatrix} J_j^T & \frac{G_j^T}{q_{\nu_j}^j} + \frac{3 G_j^T \boldsymbol{l}_j (\boldsymbol{w}_{\epsilon_j,\nu_j}^j)^T}{2 q_{\nu_j}^j a_{\nu_j}^j} \end{pmatrix} \right. \\
&\quad \left. + \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{\operatorname{diag}^2(g_j) - \operatorname{diag}(g_j)\operatorname{diag}(\boldsymbol{w}_{\epsilon_j,\nu_j}^j)}{q_{\nu_j}^j} - \epsilon_j \operatorname{diag}(g_j)\operatorname{diag}(\boldsymbol{\lambda}_j) \end{pmatrix} \right\} \begin{pmatrix} c_j \\ \boldsymbol{w}_{\epsilon_j,\nu_j}^j \end{pmatrix} \\
&=: \mathcal{H}_{1,j} \begin{pmatrix} J_j \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j \\ G_j \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j + \operatorname{diag}^2(g_j) \boldsymbol{\lambda}_j \end{pmatrix} + \frac{1}{\epsilon_j} \mathcal{H}_{2,j} \begin{pmatrix} c_j \\ \boldsymbol{w}_{\epsilon_j,\nu_j}^j \end{pmatrix}.
\end{aligned} \tag{B.6}
$$

Let us focus on $\mathcal{H}_{2,j}$. We know that

$$
\begin{aligned}
\mathcal{H}_{2,j} &= \begin{pmatrix} J_j J_j^T & J_j G_j^T / q_{\nu_j}^j \\ G_j J_j^T & \left\{ G_j G_j^T + \operatorname{diag}^2(g_j) \right\} / q_{\nu_j}^j \end{pmatrix} + \underbrace{\begin{pmatrix} \boldsymbol{0} & \frac{3}{2 q_{\nu_j}^j a_{\nu_j}^j} J_j G_j^T \boldsymbol{l}_j (\boldsymbol{w}_{\epsilon_j,\nu_j}^j)^T \\ \boldsymbol{0} & \begin{matrix} \frac{3}{2 q_{\nu_j}^j a_{\nu_j}^j} G_j G_j^T \boldsymbol{l}_j (\boldsymbol{w}_{\epsilon_j,\nu_j}^j)^T \\ - \frac{\operatorname{diag}(g_j)\operatorname{diag}(\boldsymbol{w}_{\epsilon_j,\nu_j}^j)}{q_{\nu_j}^j} - \epsilon_j \operatorname{diag}(g_j)\operatorname{diag}(\boldsymbol{\lambda}_j) \end{matrix} \end{pmatrix}}_{\Delta \mathcal{H}_{2,j}} \\
&= M_j \begin{pmatrix} I & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{q_{\nu_j}^j} I \end{pmatrix} + \Delta \mathcal{H}_{2,j}.
\end{aligned}
$$

Recalling that $\sigma_{\min}(\cdot)$ denotes the least singular value of a matrix, by the Weyl's inequality,

$$
\sigma_{\min}(\mathcal{H}_{2,j}) \geq \sigma_{\min} \left\{ M_j \begin{pmatrix} I & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{q_{\nu_j}^j} I \end{pmatrix} \right\} - \|\Delta \mathcal{H}_{2,j}\| \geq \frac{\sigma_{\min}(M_j)}{1 \vee q_{\nu_j}^j} - \|\Delta \mathcal{H}_{2,j}\|.
$$

Since $\epsilon_j \rightarrow 0$ and $\boldsymbol{w}_{\epsilon_j,\nu_j}^j \rightarrow 0$ as $j \rightarrow \infty$ (because $\tilde{\boldsymbol{x}} \in \Omega$), we know $\Delta \mathcal{H}_{2,j} \rightarrow \boldsymbol{0}$. In addition, since $M_j \rightarrow M(\tilde{\boldsymbol{x}})$ with $M(\tilde{\boldsymbol{x}})$ being positive definite, and $q_{\nu_j}^j \leq \nu_j = \tilde{\nu}$, we know for some constant $\varphi > 0$ and sufficiently large j,

$$
\sigma_{\min}(\mathcal{H}_{2,j}) \geq \varphi. \tag{B.7}
$$

Now we bound the first term in (B.6). By (10) and the invertibility of M_j, we know

$$
\begin{aligned}
\left\| \begin{pmatrix} J_j \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j \\ G_j \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_j + \operatorname{diag}^2(g_j) \boldsymbol{\lambda}_j \end{pmatrix} \right\| &\overset{(10)}{=} \frac{1}{\eta} \left\| M_j^{-1} \left\{ \begin{pmatrix} \bar{\nabla}_{\boldsymbol{\mu}} \mathcal{L}_{\epsilon_j,\nu_j,\eta}^j \\ \bar{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}_{\epsilon_j,\nu_j,\eta}^j \end{pmatrix} - \begin{pmatrix} c_j \\ \boldsymbol{w}_{\epsilon_j,\nu_j}^j + \frac{\|\boldsymbol{w}_{\epsilon_j,\nu_j}^j\|^2}{\epsilon_j a_{\nu_j}^j} \boldsymbol{\lambda}_j \end{pmatrix} \right\} \right\| \\
&\leq \frac{\Upsilon_M}{\eta} \left\{ \begin{pmatrix} \bar{\nabla}_{\boldsymbol{\mu}} \mathcal{L}_{\epsilon_j,\nu_j,\eta}^j \\ \bar{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}_{\epsilon_j,\nu_j,\eta}^j \end{pmatrix} + \left\| \begin{pmatrix} j \\ \boldsymbol{w}_{\epsilon_j,\nu_j}^j \end{pmatrix} \right\| + \frac{\|\boldsymbol{w}_{\epsilon_j,\nu_j}^j\|^2 \|\boldsymbol{\lambda}_j\|}{\epsilon_j a_{\nu_j}^j} \right\} \\
&\overset{(B.1)}{\leq} \frac{\Upsilon_M}{\eta} \left\{ \left(1 + \frac{1}{\chi_{err}} \right) \left\| \begin{pmatrix} c_j \\ \boldsymbol{w}_{\epsilon_j,\nu_j}^j \end{pmatrix} \right\| + \frac{\|\boldsymbol{w}_{\epsilon_j,\nu_j}^j\|^2 \|\boldsymbol{\lambda}_j\|}{\epsilon_j a_{\nu_j}^j} \right\}
\end{aligned}
$$

$$\overset{(6)}{\leq} \frac{2\Upsilon_M}{\chi_{err}\eta}\left\{\left\|\begin{pmatrix} c_j \\ \boldsymbol{w}^j_{\epsilon_j,\nu_j}\end{pmatrix}\right\| + \frac{\|\boldsymbol{w}^j_{\epsilon_j,\nu_j}\|^2\|\boldsymbol{\lambda}_j\|}{\epsilon_j\nu_j}\right\}$$

(also use $\chi_{err} \leq 1$)

$$\leq \frac{2\Upsilon_M}{\chi_{err}\eta\epsilon_j}\left\{\epsilon_j + \frac{\|\boldsymbol{w}^j_{\epsilon_j,\nu_j}\|\|\boldsymbol{\lambda}_j\|}{\nu_j}\right\}\left\|\begin{pmatrix} c_j \\ \boldsymbol{w}^j_{\epsilon_j,\nu_j}\end{pmatrix}\right\|. \tag{B.8}$$

Moreover, by the compactness condition, we have $\|\mathcal{H}_{1,j}\| \leq \Upsilon_1$ and $\|(J_j^T\ G_j^T)\| \leq \Upsilon_2$ for some constants $\Upsilon_1, \Upsilon_2 > 0$. Combining (B.7), (B.8) with (B.6), we have

$$\begin{aligned}
\epsilon_j\Upsilon_2\left\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}^j_{\epsilon_j,\nu_j,\eta}\right\| &\geq \epsilon_j\left\|\begin{pmatrix} J_j \\ G_j\end{pmatrix}\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}^j_{\epsilon_j,\nu_j,\eta}\right\| \\
&\overset{(B.6)}{\geq} \left\|\mathcal{H}_{2,j}\begin{pmatrix} c_j \\ \boldsymbol{w}^j_{\epsilon_j,\nu_j}\end{pmatrix}\right\| - \epsilon_j\left\|\mathcal{H}_{1,j}\begin{pmatrix} J_j\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_j \\ G_j\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_j + \text{diag}^2(g_j)\boldsymbol{\lambda}_j\end{pmatrix}\right\| \\
&\overset{(B.7)}{\geq} \varphi\cdot\left\|\begin{pmatrix} c_j \\ \boldsymbol{w}^j_{\epsilon_j,\nu_j}\end{pmatrix}\right\| - \epsilon_j\Upsilon_1\left\|\begin{pmatrix} J_j\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_j \\ G_j\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_j + \text{diag}^2(g_j)\boldsymbol{\lambda}_j\end{pmatrix}\right\| \\
&\overset{(B.8)}{\geq} \left\{\varphi - \frac{2\Upsilon_1\Upsilon_M}{\chi_{err}\eta}\left(\epsilon_j + \frac{\|\boldsymbol{w}^j_{\epsilon_j,\nu_j}\|\|\boldsymbol{\lambda}_j\|}{\nu_j}\right)\right\}\left\|\begin{pmatrix} c_j \\ \boldsymbol{w}^j_{\epsilon_j,\nu_j}\end{pmatrix}\right\| \\
&=: (\varphi - \varphi_j)\|(c_j, \boldsymbol{w}^j_{\epsilon_j,\nu_j})\|.
\end{aligned}$$

Noting that $\varphi_j \to 0$ as $j \to \infty$ (since $\boldsymbol{w}^j_{\epsilon_j,\nu_j} \to 0$ and $\epsilon_j \to 0$), we obtain for large j that

$$\epsilon_j\Upsilon_2/\chi_{err}\cdot\|(c_j, \boldsymbol{w}^j_{\epsilon_j,\nu_j})\| \overset{(B.1)}{\geq} \epsilon_j\Upsilon_2\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}^j_{\epsilon_j,\nu_j,\eta}\| \geq \varphi/2\cdot\|(c_j, \boldsymbol{w}^j_{\epsilon_j,\nu_j})\|,$$

which cannot hold because $\epsilon_j \searrow 0$. This is a contradiction, and thus we complete the proof.

B.2 Proof of Lemma 5

The proof closely follows the proof of Lemma 3 in Appendix A.2 We suppress the iteration t and assume ξ_1^t is any sample set. Our analysis is independent of the sample set ξ_1^t for computing $\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}$, and we will see that the threshold is independent of t. Like Lemma 3, we use $\Upsilon_1, \Upsilon_2, \ldots$ to denote generic constants that are independent of $(\bar{\epsilon}_t, \bar{\nu}_t, \eta, \gamma_B, \gamma_H)$, whose existence is ensured by the compactness of the iterates.

Following the derivation of (A.4), we have

$$\begin{aligned}
(\bar{\nabla}\mathcal{L}^{(1)}_{\bar{\epsilon},\bar{\nu},\eta})^T\bar{\Delta} = &-\bar{\Delta}\boldsymbol{x}^T B\bar{\Delta}\boldsymbol{x} + \begin{pmatrix} c \\ g_a\end{pmatrix}^T\begin{pmatrix} \bar{\tilde{\Delta}}\boldsymbol{\mu} + \bar{\Delta}\boldsymbol{\mu} \\ \bar{\tilde{\Delta}}\boldsymbol{\lambda}_a + \bar{\Delta}\boldsymbol{\lambda}_a\end{pmatrix} - \frac{1}{\bar{\epsilon}}\|c\|^2 - \frac{1}{\bar{\epsilon}q_{\bar{\nu}}}\|g_a\|^2 \\
&-\bar{\epsilon}q_{\bar{\nu}}\bar{\Delta}\boldsymbol{\lambda}_c^T\boldsymbol{\lambda}_c - \eta\left\|\begin{pmatrix} J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} \\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} + \Pi_c(\text{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2,
\end{aligned} \tag{B.9}$$

where $(\bar{\bar{\varDelta}}\boldsymbol{\mu}, \bar{\bar{\varDelta}}\boldsymbol{\lambda}_a)$ is the dual solution of (12a) with $\nabla_{\boldsymbol{x}}\mathcal{L}$ being replaced by $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}$. Following the derivation of (A.5), there exists $\varUpsilon_1 > 0$ such that

$$\left\|\begin{pmatrix} \bar{\varDelta}\boldsymbol{\mu} \\ \bar{\varDelta}\boldsymbol{\lambda} \end{pmatrix}\right\| \leq \frac{\varUpsilon_1}{\gamma_H}\left\|\begin{pmatrix} \bar{\varDelta}\boldsymbol{x} \\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} \\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} + \varPi_c(\operatorname{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix}\right\|. \tag{B.10}$$

Following the derivation of (A.6), there exists $\varUpsilon_2 > 0$ such that

$$\left\|\begin{pmatrix} \bar{\bar{\varDelta}}\boldsymbol{\mu} \\ \bar{\bar{\varDelta}}\boldsymbol{\lambda}_a \\ -\boldsymbol{\lambda}_c \end{pmatrix}\right\| \leq \frac{\varUpsilon_2}{\gamma_H^2\gamma_B}\left\|\begin{pmatrix} \bar{\varDelta}\boldsymbol{x} \\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} \\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} + \varPi_c(\operatorname{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix}\right\|. \tag{B.11}$$

Following the derivation of (A.7) by combining (B.9), (B.10), and (B.11), and noting that $0 < q_{\bar{\nu}} \leq \bar{\nu} \leq \tilde{\nu}$ where $\tilde{\nu}$ is defined in (30), there exists $\varUpsilon_3 > 0$ such that

$$\begin{aligned}
(\bar{\nabla}\mathcal{L}^{(1)}_{\bar{\epsilon},\bar{\nu},\eta})^T\bar{\varDelta} \leq & -\bar{\varDelta}\boldsymbol{x}^T B\bar{\varDelta}\boldsymbol{x} + \frac{\varUpsilon_3}{\gamma_H^2\gamma_B}\left\|\begin{pmatrix} c \\ g_a \end{pmatrix}\right\|\left\|\begin{pmatrix} \bar{\varDelta}\boldsymbol{x} \\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} \\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} + \varPi_c(\operatorname{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix}\right\| \\
& - \frac{1}{\bar{\epsilon}(1 \vee \tilde{\nu})}\left\|\begin{pmatrix} c \\ g_a \end{pmatrix}\right\|^2 \\
& + \frac{\bar{\epsilon}\tilde{\nu}\varUpsilon_3}{\gamma_H^3\gamma_B}\left\|\begin{pmatrix} \bar{\varDelta}\boldsymbol{x} \\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} \\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} + \varPi_c(\operatorname{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix}\right\|^2 \\
& - \eta\left\|\begin{pmatrix} J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} \\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} + \varPi_c(\operatorname{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix}\right\|^2.
\end{aligned} \tag{B.12}$$

Following the derivation of (A.9), there exists $\varUpsilon_4 > 0$ such that

$$-\bar{\varDelta}\boldsymbol{x}^T B\bar{\varDelta}\boldsymbol{x} \leq -\frac{3\gamma_B}{4}\|\bar{\varDelta}\boldsymbol{x}\|^2 + \frac{\varUpsilon_4}{\gamma_H\gamma_B}\left\|\begin{pmatrix} c \\ g_a \end{pmatrix}\right\|^2.$$

Combining the above display with (B.12) and using the following Young's inequality

$$\begin{aligned}
& \frac{\varUpsilon_3}{\gamma_H^2\gamma_B}\left\|\begin{pmatrix} c \\ g_a \end{pmatrix}\right\|\left\|\begin{pmatrix} \bar{\varDelta}\boldsymbol{x} \\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} \\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} + \varPi_c(\operatorname{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix}\right\| \\
& \leq \left(\frac{\gamma_B}{8} \wedge \frac{\eta}{4}\right)\left\|\begin{pmatrix} \bar{\varDelta}\boldsymbol{x} \\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} \\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L} + \varPi_c(\operatorname{diag}^2(g)\boldsymbol{\lambda}) \end{pmatrix}\right\|^2 + \frac{2\varUpsilon_3^2}{\gamma_H^4\gamma_B^2(\gamma_B \wedge \eta)}\left\|\begin{pmatrix} c \\ g_a \end{pmatrix}\right\|^2,
\end{aligned}$$

we have

$$
\begin{aligned}
(\bar{\nabla}\mathcal{L}_{\epsilon,\nu,\eta}^{(1)})^T\bar{\Delta} &\leq -\frac{3\gamma_B}{4}\|\bar{\Delta}\boldsymbol{x}\|^2 + \left\{\left(\frac{\gamma_B}{8}\wedge\frac{\eta}{4}\right) + \frac{\bar{\epsilon}\tilde{\nu}\Upsilon_3}{\gamma_H^3\gamma_B}\right\}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}\\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}\\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2\\
&\quad + \left\{\frac{\Upsilon_4}{\gamma_H\gamma_B} + \frac{2\Upsilon_3^2}{\gamma_H^4\gamma_B^2(\gamma_B\wedge\eta)} - \frac{1}{\bar{\epsilon}(1\vee\tilde{\nu})}\right\}\left\|\begin{pmatrix}c\\ g_a\end{pmatrix}\right\|^2 - \eta\left\|\begin{pmatrix}J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}\\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2\\
&\leq -\left\{\frac{\gamma_B\wedge\eta}{2} + \left(\frac{\gamma_B}{8}\wedge\frac{\eta}{4}\right) - \frac{\bar{\epsilon}\tilde{\nu}\Upsilon_3}{\gamma_H^3\gamma_B}\right\}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}\\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}\\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2\\
&\quad - \left\{\frac{1}{\bar{\epsilon}(1\vee\tilde{\nu})} - \frac{\Upsilon_4}{\gamma_H\gamma_B} - \frac{2\Upsilon_3^2}{\gamma_H^4\gamma_B^2(\gamma_B\wedge\eta)}\right\}\left\|\begin{pmatrix}c\\ g_a\end{pmatrix}\right\|^2.
\end{aligned}
$$

Therefore, as long as

$$
\begin{aligned}
&\frac{\gamma_B}{8}\wedge\frac{\eta}{4} \geq \frac{\bar{\epsilon}\tilde{\nu}\Upsilon_3}{\gamma_H^3\gamma_B} \Longleftarrow \frac{1}{\bar{\epsilon}} \geq \frac{8\tilde{\nu}\Upsilon_3}{\gamma_H^3\gamma_B(\gamma_B\wedge\eta)},\\
&\frac{1}{\bar{\epsilon}(1\vee\tilde{\nu})} - \frac{\Upsilon_4}{\gamma_H\gamma_B} - \frac{2\Upsilon_3^2}{\gamma_H^4\gamma_B^2(\gamma_B\wedge\eta)} \geq 0 \Longleftarrow \frac{1}{\bar{\epsilon}} \geq \frac{(1\vee\tilde{\nu})(2\Upsilon_3^2+\Upsilon_4)}{\gamma_H^4\gamma_B^2(\gamma_B\wedge\eta)},
\end{aligned}
\tag{B.13}
$$

we have

$$
(\bar{\nabla}\mathcal{L}_{\epsilon,\nu,\eta}^{(1)})^T\bar{\Delta} \leq -\frac{\gamma_B\wedge\eta}{2}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}\\ J\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}\\ G\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}+\Pi_c(\mathrm{diag}^2(g)\boldsymbol{\lambda})\end{pmatrix}\right\|^2.
$$

Thus, we can define

$$
\tilde{\epsilon}_2 := \frac{\gamma_H^4\gamma_B^2(\gamma_B\wedge\eta)}{(2\Upsilon_3^2+8\Upsilon_3+\Upsilon_4)(\tilde{\nu}\vee 1)},
$$

which implies (B.13) and completes the proof.

B.3 Proof of Lemma 6

We let $C_1, C_2, \ldots$ be generic constants that are independent of $(\beta, \alpha_{max}, \kappa_{grad}, \kappa_f, p_{grad}, p_f, \chi_{grad}, \chi_f)$. These constants may not be consistent with the constants C_1, C_2, C_3 in the statement. However, the existence of C_1, C_2, C_3 in the statement follows directly from our proof.

(a1) By the definition of $\nabla\mathcal{L}_{\epsilon,\nu,\eta}$ in (10), all quantities depending on ϵ, ν do not depend on the batch samples. We have

$$\begin{aligned}\bar{\nabla}\mathcal{L}_{\epsilon,\nu,\eta}^t-\nabla\mathcal{L}_{\epsilon,\nu,\eta}^t &\overset{(10)}{=}\begin{pmatrix}\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t-\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}+\eta\begin{pmatrix}\bar{Q}_{1,t} & \bar{Q}_{2,t}\\ M_{11,t} & M_{12,t}\\ M_{21,t} & M_{22,t}\end{pmatrix}\begin{pmatrix}J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\operatorname{diag}^2(g_t)\boldsymbol{\lambda}_t\end{pmatrix}\\ &\quad-\eta\begin{pmatrix}Q_{1,t} & Q_{2,t}\\ M_{11,t} & M_{12,t}\\ M_{21,t} & M_{22,t}\end{pmatrix}\begin{pmatrix}J_t\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\nabla_{\boldsymbol{x}}\mathcal{L}_t+\operatorname{diag}^2(g_t)\boldsymbol{\lambda}_t\end{pmatrix}\\ &=\begin{pmatrix}\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t-\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}+\eta\begin{pmatrix}Q_{1,t} & Q_{2,t}\\ M_{11,t} & M_{12,t}\\ M_{21,t} & M_{22,t}\end{pmatrix}\begin{pmatrix}J_t(\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t-\nabla_{\boldsymbol{x}}\mathcal{L}_t)\\ G_t(\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t-\nabla_{\boldsymbol{x}}\mathcal{L}_t)\end{pmatrix}\\ &\quad+\eta\begin{pmatrix}\bar{Q}_{1,t}-Q_{1,t} & \bar{Q}_{2,t}-Q_{2,t}\\ \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\operatorname{diag}^2(g_t)\boldsymbol{\lambda}_t\end{pmatrix}.\end{aligned}$$

By Assumption 3, the definition (9), and the facts that $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t-\nabla_{\boldsymbol{x}}\mathcal{L}_t=\bar{\nabla}f_t-\nabla f_t$ and $\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_t-\nabla_{\boldsymbol{x}}^2\mathcal{L}_t=\bar{\nabla}^2f_t-\nabla^2f_t$, there exists $C_1>0$ (depending on η) such that

$$\begin{aligned}&\|\bar{\nabla}\mathcal{L}_{\epsilon,\nu,\eta}^t-\nabla\mathcal{L}_{\epsilon,\nu,\eta}^t\|\\ &\quad\leq C_1\|\bar{\nabla}f_t-\nabla f_t\|+C_1\|\bar{\nabla}^2f_t-\nabla^2f_t\|(\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|+\|\operatorname{diag}^2(g_t)\boldsymbol{\lambda}_t\|).\end{aligned}$$

Since $\|\operatorname{diag}^2(g_t)\boldsymbol{\lambda}_t\|\leq C_2\|\max\{g_t,-\boldsymbol{\lambda}_t\}\|$ for some constant $C_2>0$, we apply the definition of $\bar{R}_t$ in (14) and the uniform boundedness of $\bar{R}_t$, and know that the above inequality leads to the statement.

(a2) By the definition of $\mathcal{L}_{\epsilon,\nu,\eta}$ in (8), all quantities depending on ϵ, ν do not depend on the batch samples. We have

$$\begin{aligned}&\bar{\mathcal{L}}_{\epsilon,\nu,\eta}^t-\mathcal{L}_{\epsilon,\nu,\eta}^t\\ &\overset{(8)}{=}\bar{\mathcal{L}}_t-\mathcal{L}_t+\frac{\eta}{2}\begin{pmatrix}J_t(\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t-\nabla_{\boldsymbol{x}}\mathcal{L}_t)\\ G_t(\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t-\nabla_{\boldsymbol{x}}\mathcal{L}_t)\end{pmatrix}^T\begin{pmatrix}J_t(\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\nabla_{\boldsymbol{x}}\mathcal{L}_t)\\ G_t(\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\nabla_{\boldsymbol{x}}\mathcal{L}_t)+2\operatorname{diag}^2(g_t)\boldsymbol{\lambda}_t\end{pmatrix}.\end{aligned}$$

By Assumption 3 and the facts that $\bar{\mathcal{L}}_t-\mathcal{L}_t=\bar{f}_t-f_t$ and $\|\operatorname{diag}^2(g_t)\boldsymbol{\lambda}_t\|\leq C_2\|\max\{g_t,-\boldsymbol{\lambda}_t\}\|$, there exists $C_3>0$ (depending on η) such that

$$|\bar{\mathcal{L}}_{\epsilon,\nu,\eta}^t-\mathcal{L}_{\epsilon,\nu,\eta}^t|\leq C_3|\bar{f}_t-f_t|+C_3\|\bar{\nabla}f_t-\nabla f_t\|(R_t+\bar{R}_t).$$

Using $R_t\leq\bar{R}_t+\|\bar{\nabla}f_t-\nabla f_t\|\leq 2(\bar{R}_t\vee\|\bar{\nabla}f_t-\nabla f_t\|)$, we prove the statement.

(b) By (10) and Assumption 3, there exists $C_4>0$ such that

$$\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|\leq\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}^t\|+C_4\left\{\left\|\begin{pmatrix}J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\end{pmatrix}\right\|+\|\operatorname{diag}^2(g_t)\boldsymbol{\lambda}_t\|\right\}$$

$$+\frac{C_4}{\bar{\epsilon}_t(1\wedge q^t_{\bar{\nu}_t})}\left\|\begin{pmatrix} c_t \\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\| + \frac{C_4}{\bar{\epsilon}_t q^t_{\bar{\nu}_t} a^t_{\bar{\nu}_t}}\|\boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\|^2.$$

By Theorem 1, we have

$$\bar{\epsilon}_0 \geq \bar{\epsilon}_t \geq \tilde{\epsilon}, \quad \tilde{\nu} \geq \bar{\nu}_t \geq q^t_{\bar{\nu}_t} \overset{\text{(A.12)}}{\geq} \kappa_{\bar{\nu}_t} \geq \kappa_{\bar{\nu}_0}, \quad \tilde{\nu} \geq \bar{\nu}_t \geq a^t_{\bar{\nu}_t} \geq \frac{\bar{\nu}_t}{2} \geq \frac{\bar{\nu}_0}{2}. \tag{B.14}$$

Thus, there exists $C_5 > 0$ such that

$$\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\| \leq \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\| + C_5\left\{\left\|\begin{pmatrix} J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\end{pmatrix}\right\| + \|\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t\| + \left\|\begin{pmatrix} c_t \\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\| + \|\boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\|^2\right\}.$$

Moreover, there exists $C_6 > 0$ such that

$$\|\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t\| \leq C_6\left\|\begin{pmatrix} g_{t_a} \\ \boldsymbol{\lambda}_{t_c}\end{pmatrix}\right\| \leq \frac{C_6}{\bar{\epsilon}_t q^t_{\bar{\nu}_t}\wedge 1}\left\|\begin{pmatrix} g_{t_a} \\ -\bar{\epsilon}_t q^t_{\bar{\nu}_t}\boldsymbol{\lambda}_{t_c}\end{pmatrix}\right\| \overset{\text{(B.14)}}{\leq} \frac{C_6}{\tilde{\epsilon}\kappa_{\bar{\nu}_0}\wedge 1}\|\boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\|, \tag{B.15}$$

and

$$\|\boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\| \overset{\text{Lem. 14}}{\leq} C_6(\bar{\epsilon}_t q^t_{\bar{\nu}_t}\vee 1) \leq C_6(\bar{\epsilon}_0\tilde{\nu}\vee 1). \tag{B.16}$$

Combining the above three displays, there exists $C_7 > 0$ such that

$$\begin{aligned}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\| \leq{}& \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\| \\ &+C_7\left\{\left\|\begin{pmatrix} J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\end{pmatrix}\right\| + \left\|\begin{pmatrix} c_t \\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\|\right\}.\end{aligned} \tag{B.17}$$

We deal with the middle term. We know that

$$\begin{aligned}&\begin{pmatrix} M_{11,t} & M_{12,t} \\ M_{21,t} & M_{22,t}\end{pmatrix}\begin{pmatrix} J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\end{pmatrix} \\ &\overset{(10)}{=} \frac{1}{\eta}\begin{pmatrix} \bar{\nabla}_{\boldsymbol{\mu}}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta} \\ \bar{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\end{pmatrix} - \frac{1}{\eta}\begin{pmatrix} c_t \\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t} + \frac{\|\boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\|^2}{\bar{\epsilon}_t a^t_{\bar{\nu}_t}}\boldsymbol{\lambda}_t\end{pmatrix} \\ &\quad -\begin{pmatrix} M_{12,t} \\ M_{22,t}\end{pmatrix}\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t.\end{aligned} \tag{B.18}$$

Multiplying $((J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t)^T\ \ (G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t)^T)$ on both sides, there exists $C_8 > 0$ such that

$$\|J_t^T J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + G_t^T G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|^2 \leq \begin{pmatrix} J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\end{pmatrix}^T\begin{pmatrix} M_{11,t} & M_{12,t} \\ M_{21,t} & M_{22,t}\end{pmatrix}\begin{pmatrix} J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\end{pmatrix}$$

$$\overset{(B.18),(B.14)-(B.16)}{\leq} C_8\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|\left\{\left\|\begin{pmatrix}\bar{\nabla}_{\boldsymbol{\mu}}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\\ \bar{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\end{pmatrix}\right\|+\left\|\begin{pmatrix}c_t\\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\|\right\}. \tag{B.19}$$

Furthermore,

$$\begin{aligned}\left\|\begin{pmatrix}J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\end{pmatrix}\right\|^2 &\leq \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|\cdot\|J_t^TJ_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+G_t^TG_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|\\ &\overset{(B.19)}{\leq}\sqrt{C_8}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|^{\frac{3}{2}}\left\{\left\|\begin{pmatrix}\bar{\nabla}_{\boldsymbol{\mu}}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\\ \bar{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\end{pmatrix}\right\|+\left\|\begin{pmatrix}c_t\\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\|\right\}^{\frac{1}{2}}.\end{aligned}$$

Combining the above display with (B.17), there exists $C_9>0$ such that

$$\begin{aligned}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|\leq C_9\left\{\left\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right\|+\left\|\begin{pmatrix}c_t\\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\|\right\}+C_9^{1/4}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|^{\frac{3}{4}}\left\{\left\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right\|+\left\|\begin{pmatrix}c_t\\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\|\right\}^{\frac{1}{4}}\\ \leq\frac{5C_9}{4}\left\{\left\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right\|+\left\|\begin{pmatrix}c_t\\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\|\right\}+\frac{3}{4}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|,\end{aligned}$$

where the second inequality is due to Young's inequality $a^{3/4}b^{1/4}\leq 3a/4+b/4$. Thus,

$$\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|\leq 5C_9\left\{\left\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right\|+\left\|\begin{pmatrix}c_t\\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\|\right\}.$$

(c) By (10) and using (B.14), (B.15) and (B.16), there exists $C_{10}>0$ such that

$$\begin{aligned}\left\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_t,\bar{\nu}_t,\eta}\right\|&\leq\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|+C_{10}\left\|\begin{pmatrix}J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|+C_{10}\left\|\begin{pmatrix}c_t\\ \boldsymbol{w}^t_{\bar{\epsilon}_t,\bar{\nu}_t}\end{pmatrix}\right\|\\ &\leq\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|+C_{10}\left\|\begin{pmatrix}J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|+C_{10}(\bar{\epsilon}_tq^t_{\bar{\nu}_t}\vee 1)\left\|\begin{pmatrix}c_t\\ g_{t_a}\\ \boldsymbol{\lambda}_{t_c}\end{pmatrix}\right\|\\ &\overset{(B.14)}{\leq}\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\|+C_{10}\left\|\begin{pmatrix}J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|+C_{10}(\bar{\epsilon}_0\tilde{\nu}\vee 1)\left\|\begin{pmatrix}c_t\\ g_{t_a}\\ \boldsymbol{\lambda}_{t_c}\end{pmatrix}\right\|.\end{aligned}\tag{B.20}$$

For $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t$, we have the following decomposition

$$\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t=\underbrace{\left\{I-(J_t^T\;G_{t_a}^T)\left\{\begin{pmatrix}J_t\\ G_{t_a}\end{pmatrix}(J_t^T\;G_{t_a}^T)\right\}^{-1}\begin{pmatrix}J_t\\ G_{t_a}\end{pmatrix}\right\}}_{\mathcal{P}^t_{JG}}\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t+(I-\mathcal{P}^t_{JG})\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t.$$

By Assumptions 3 and 5, we know $\|(I-\mathcal{P}^t_{JG})\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\| \le C_{11}\|(J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t, G_{t_a}\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t)\|$ for some constant $C_{11}>0$. Furthermore, for some constant $C_{12}>0$, we also have

$$\|\mathcal{P}^t_{JG}\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\| \overset{(12a)}{=} \left\|\mathcal{P}^t_{JG}\left\{B_t\bar{\Delta}\boldsymbol{x}_t + J_t^T\bar{\tilde{\Delta}}\boldsymbol{\mu}_t + G_{t_a}^T\bar{\tilde{\Delta}}\boldsymbol{\lambda}_{t_a} - G_{t_c}^T\boldsymbol{\lambda}_{t_c}\right\}\right\|$$
$$\le \|\mathcal{P}^t_{JG}B_t\bar{\Delta}\boldsymbol{x}_t\| + \|\mathcal{P}^t_{JG}G_{t_c}^T\boldsymbol{\lambda}_{t_c}\| \overset{(B.11)}{\le} C_{12}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|.$$

Combining the last two displays, we have

$$\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\| \le (C_{11}+C_{12})\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|. \tag{B.21}$$

Moreover, there exists $C_{13}>0$ such that

$$\left\|\begin{pmatrix}c_t\\ g_{t_a}\end{pmatrix}\right\| \overset{(12a)}{\le} C_{13}\|\bar{\Delta}\boldsymbol{x}_t\|, \qquad \|\boldsymbol{\lambda}_{t_c}\| \overset{(B.11)}{\le} C_{13}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|. \tag{B.22}$$

Combining (B.20), (B.21), and (B.22) together, we complete the proof.

B.4 Proof of Lemma 7

Analogous to the proof of Lemma 6, we only track the constants $(\beta, \alpha_{max}, \kappa_{grad}, \kappa_f, p_{grad}, p_f, \chi_{grad}, \chi_f)$. We use $\Upsilon_1, \Upsilon_2, \ldots$ to denote generic constants that are independent from $(\beta, \alpha_{max}, \kappa_{grad}, \kappa_f, p_{grad}, p_f, \chi_{grad}, \chi_f)$. Note that Υ_1 in the proof may not be consistent with Υ_1 in the statement, while the existence of Υ_1 in the statement follows directly from our proof.

Let $\Upsilon_{\epsilon,\nu,\eta}$ be the upper bound of the generalized Hessian of $\mathcal{L}_{\epsilon,\nu,\eta}$ in the compact set $(\mathcal{X}\cap\mathcal{T}_{\theta\nu})\times\mathcal{M}\times\Lambda$ (see [50] for the definition of the generalized Hessian). In particular, $\Upsilon_{\epsilon,\nu,\eta} = \sup_{(\mathcal{X}\cap\mathcal{T}_{\theta\nu})\times\mathcal{M}\times\Lambda}\|\partial^2\mathcal{L}_{\epsilon,\nu,\eta}\|$. Without loss of generality, we suppose $\tilde{\epsilon}$ in Theorem 1 satisfies $\tilde{\epsilon} = \bar{\epsilon}_0/\rho^{\tilde{i}}$ for some integer $\tilde{i}$. Then, with definition $\tilde{j}$ in (30), we let

$$\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta} = \max\{\Upsilon_{\epsilon,\nu,\eta} : \epsilon = \bar{\epsilon}_0/\rho^i, \nu = \rho^j\bar{\nu}_0, 1\le i\le\tilde{i}, 1\le j\le\tilde{j}\}$$

and have $\Upsilon_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} \le \Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}$. Noting that $\boldsymbol{x}_{s_t}, \boldsymbol{x}_t \in \mathcal{T}_{\bar{\nu}_{\bar{t}}}$, we apply the Taylor expansion and have

$$\begin{aligned}\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} &\le \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t(\nabla\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t + \frac{\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\bar{\alpha}_t^2}{2}\|\check{\Delta}_t\|^2\\ &= \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t + \bar{\alpha}_t(\nabla\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t + \frac{\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\bar{\alpha}_t^2}{2}\|\check{\Delta}_t\|^2\end{aligned}$$

$$\begin{aligned}
&\leq \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t (\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \check{\Delta}_t + \bar{\alpha}_t \|\check{\Delta}_t\| \cdot \bar{\Delta}(\nabla \mathcal{L}^t_\eta) + \frac{\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\bar{\alpha}_t^2}{2}\|\check{\Delta}_t\|^2 \\
&\overset{(14)}{\leq} \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t (\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \check{\Delta}_t + \kappa_{grad}\bar{\alpha}_t^2 \cdot \bar{R}_t \|\check{\Delta}_t\| + \frac{\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\bar{\alpha}_t^2}{2}\|\check{\Delta}_t\|^2. \qquad \text{(B.23)}
\end{aligned}$$

We consider the following two cases.
Case 1, $\check{\Delta}_t = \bar{\Delta}_t$ Combining (18) with (19), we have

$$(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \bar{\Delta}_t \leq -\frac{\gamma_B \wedge \eta}{4} \left\| \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\text{diag}^2(g_t)\boldsymbol{\lambda}_t) \end{pmatrix} \right\|^2. \qquad \text{(B.24)}$$

By (B.10), there exists $\Upsilon_1 > 0$ such that

$$\|\bar{\Delta}_t\| \leq \Upsilon_1 \left\| \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\text{diag}^2(g_t)\boldsymbol{\lambda}_t) \end{pmatrix} \right\|. \qquad \text{(B.25)}$$

Furthermore, we have

$$\bar{R}_t \overset{\text{Lem. }14}{\leq} \frac{1}{\bar{\epsilon}_{\bar{t}} q^t_{\bar{\nu}_{\bar{t}}} \wedge 1} \left\| \begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \\ \boldsymbol{w}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}}} \end{pmatrix} \right\| \overset{(B.14)}{\leq} \frac{1}{\tilde{\epsilon}\kappa_{\bar{\nu}_0} \wedge 1} \left\| \begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \\ g_{t_a} \\ -\bar{\epsilon}_{\bar{t}} q^t_{\bar{\nu}_{\bar{t}}} \boldsymbol{\lambda}_{t_c} \end{pmatrix} \right\| \leq \frac{\bar{\epsilon}_0 \tilde{\nu} \vee 1}{\tilde{\epsilon}\kappa_{\bar{\nu}_0} \wedge 1} \left\| \begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \\ g_{t_a} \\ \boldsymbol{\lambda}_{t_c} \end{pmatrix} \right\|,$$

and thus, by (B.21), (B.22), there exists $\Upsilon_2 > 0$ such that

$$\bar{R}_t \leq \Upsilon_2 \left\| \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\text{diag}^2(g_t)\boldsymbol{\lambda}_t) \end{pmatrix} \right\|. \qquad \text{(B.26)}$$

Plugging (B.25) and (B.26) into (B.23), we have

$$\begin{aligned}
\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} &\leq \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t (\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \bar{\Delta}_t \\
&\quad + \left\{ \Upsilon_1 \Upsilon_2 \kappa_{grad} + \frac{\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\Upsilon_1^2}{2} \right\} \bar{\alpha}_t^2 \left\| \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\text{diag}^2(g_t)\boldsymbol{\lambda}_t) \end{pmatrix} \right\|^2 \\
&\overset{(B.24)}{\leq} \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t (\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \bar{\Delta}_t - \left\{ \Upsilon_1 \Upsilon_2 \kappa_{grad} + \frac{\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\Upsilon_1^2}{2} \right\} \frac{4\bar{\alpha}_t^2}{\gamma_B \wedge \eta} (\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \bar{\Delta}_t \\
&\leq \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t \left\{ 1 - \Upsilon_3 \left(\kappa_{grad} + 1 \right) \bar{\alpha}_t \right\} (\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \bar{\Delta}_t, \qquad \text{(B.27)}
\end{aligned}$$

where $\Upsilon_3 = 4\Upsilon_1\Upsilon_2/(\gamma_B \wedge \eta) \vee 2\Upsilon_1^2 \Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}/(\gamma_B \wedge \eta)$.
Case 2, $\check{\Delta}_t = \hat{\Delta}_t$ By Lemma 6(b), Lemma 14, (17), and (B.14), there exists $\Upsilon_4 > 0$ such that

$$\bar{R}_t \leq \Upsilon_4 \|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|. \qquad \text{(B.28)}$$

Plugging (20) and (B.28) into (B.23), we have

$$\begin{aligned}\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} &\leq \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\hat{\Delta}_t + \Upsilon_4\chi_u\kappa_{grad}\bar{\alpha}_t^2\cdot\|\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 + \frac{\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\chi_u^2\bar{\alpha}_t^2}{2}\|\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2\\ &\overset{(20)}{\leq} \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\hat{\Delta}_t - \left(\Upsilon_4\chi_u^2\kappa_{grad} + \frac{\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\chi_u^3}{2}\right)\bar{\alpha}_t^2\cdot(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\hat{\Delta}_t\\ &\leq \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t\left\{1-\Upsilon_5\left(\kappa_{grad}+1\right)\bar{\alpha}_t\right\}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\hat{\Delta}_t,\end{aligned}\tag{B.29}$$

where $\Upsilon_5 = \Upsilon_4\chi_u^2 \vee \Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}\chi_u^3/2$.

Combining (B.27) and (B.29), and letting $\Upsilon_6 = \Upsilon_3 \vee \Upsilon_5 \vee 2$, we obtain

$$\mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} \leq \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t\left\{1-\Upsilon_6(\kappa_{grad}+1)\bar{\alpha}_t\right\}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t.\tag{B.30}$$

By the event $\mathcal{E}_2^t$, we have

$$\begin{aligned}\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} &\overset{(23)}{\leq} \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \kappa_f\bar{\alpha}_t^2(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t\\ &\overset{(B.30)}{\leq} \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t\left\{1-\Upsilon_6\left(\kappa_{grad}+1\right)\bar{\alpha}_t - \kappa_f\bar{\alpha}_t\right\}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t\\ &\overset{(23)}{\leq} \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t\left\{1-\Upsilon_6\left(\kappa_{grad}+1\right)\bar{\alpha}_t - 2\kappa_f\bar{\alpha}_t\right\}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t\\ &\leq \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t\left\{1-\Upsilon_6\left(\kappa_{grad}+\kappa_f+1\right)\bar{\alpha}_t\right\}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t \quad \text{(since } \Upsilon_6 \geq 2).\end{aligned}$$

Therefore, as long as

$$1-\Upsilon_6\left(\kappa_{grad}+\kappa_f+1\right)\bar{\alpha}_t \geq \beta \Longleftrightarrow \bar{\alpha}_t \leq \frac{1-\beta}{\Upsilon_6(\kappa_{grad}+\kappa_f+1)},$$

we have

$$\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} \leq \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} + \bar{\alpha}_t\beta(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\check{\Delta}_t.$$

This completes the proof.

B.5 Proof of Lemma 9

Algorithm 1 has three types of steps: a reliable step (Line 19), an unreliable step (Line 21), and an unsuccessful step (Line 24). For each type of step, $\check{\Delta}_t = \bar{\Delta}_t$ or $\check{\Delta}_t = \hat{\Delta}_t$. Thus, we analyze in the following six cases.

Case 1a, reliable step, $\check{\Delta}_t = \bar{\Delta}_t$ By Lemma 8, we have

$$\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} \leq \frac{\bar{\alpha}_t\beta}{2}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\bar{\Delta}_t \overset{(27)}{\leq} \frac{4\bar{\alpha}_t\beta}{9}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\bar{\Delta}_t - \frac{\bar{\delta}_t}{18}$$

$$\overset{\text{(B.24)}}{\leq} -\frac{\bar{\alpha}_t\beta(\gamma_B \wedge \eta)}{9}\left\|\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2 - \frac{\bar{\delta}_t}{18}. \tag{B.31}$$

Note that

$$\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\| \leq \bar{\Delta}(\nabla\mathcal{L}^t_\eta) + \|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\| \overset{(14)}{\leq} \kappa_{grad}\bar{\alpha}_t\bar{R}_t + \|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|.$$

Combining the above display with (B.26), Lemma 6(c), and using $\bar{\alpha}_t \leq \alpha_{max}$, there exists $\Upsilon_1 > 0$ such that

$$\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\| \leq \Upsilon_1(\kappa_{grad}\alpha_{max} + 1)\left\|\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|. \tag{B.32}$$

Combining the above inequality with (B.31), we have

$$\begin{aligned}\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} \leq &-\frac{\bar{\alpha}_t\beta(\gamma_B \wedge \eta)}{18}\left\|\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2 \\ &-\frac{\bar{\alpha}_t\beta(\gamma_B \wedge \eta)}{18\Upsilon_1^2(\kappa_{grad}\alpha_{max} + 1)^2}\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \frac{\bar{\delta}_t}{18}.\end{aligned}$$

By Line 20 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = (\rho - 1)\bar{\delta}_t$. By the Taylor expansion and $\bar{\alpha}_{t+1} \leq \rho\bar{\alpha}_t$ (Line 18), there exists $\Upsilon_2 > 0$ such that

$$\begin{aligned}\bar{\alpha}_{t+1}\|\nabla\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \bar{\alpha}_t\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 &\leq 2\rho\bar{\alpha}_t\left\{\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 + \Upsilon^2_{\tilde{\epsilon},\tilde{\nu},\eta}\bar{\alpha}_t^2\|\bar{\Delta}_t\|^2\right\} \\ &\overset{\text{(B.25)}}{\leq} 2\rho\bar{\alpha}_t\left\{\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 + \Upsilon^2_{\tilde{\epsilon},\tilde{\nu},\eta}\alpha^2_{max}\Upsilon_2\left\|\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2\right\}.\end{aligned} \tag{B.33}$$

Combining the above two displays with (31), we obtain

$$\begin{aligned}\Theta^{t+1}_\omega - \Theta^t_\omega \leq &-\left(\frac{\omega\beta(\gamma_B \wedge \eta)}{18} - (1-\omega)\rho\Upsilon^2_{\tilde{\epsilon},\tilde{\nu},\eta}\alpha^2_{max}\Upsilon_2\right)\bar{\alpha}_t\left\|\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2 \\ &-\left(\frac{\omega\beta(\gamma_B \wedge \eta)}{18\Upsilon_1^2(\kappa_{grad}\alpha_{max} + 1)^2} - (1-\omega)\rho\right)\bar{\alpha}_t\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 \\ &-\left(\frac{\omega}{18} - \frac{(1-\omega)(\rho-1)}{2}\right)\bar{\delta}_t.\end{aligned}$$

Let

$$\frac{\omega\beta(\gamma_B \wedge \eta)}{36} \geq (1-\omega)\rho\Upsilon^2_{\tilde{\epsilon},\tilde{\nu},\eta}\alpha^2_{max}\Upsilon_2 \Longleftrightarrow \frac{\omega}{1-\omega} \geq \frac{36\rho\Upsilon^2_{\tilde{\epsilon},\tilde{\nu},\eta}\alpha^2_{max}\Upsilon_2}{\beta(\gamma_B \wedge \eta)},$$

$$\frac{\omega\beta(\gamma_B\wedge\eta)}{36\Upsilon_1^2(\kappa_{grad}\alpha_{max}+1)^2}\geq(1-\omega)\rho\Longleftrightarrow\frac{\omega}{1-\omega}\geq\frac{36\rho\Upsilon_1^2(\kappa_{grad}\alpha_{max}+1)^2}{\beta(\gamma_B\wedge\eta)},\tag{B.34}$$

$$\frac{\omega}{36}\geq\frac{(1-\omega)(\rho-1)}{2}\Longleftrightarrow\frac{\omega}{1-\omega}\geq 18(\rho-1),$$

which is further implied by

$$\frac{\omega}{1-\omega}\geq\frac{\Upsilon_3(\kappa_{grad}\alpha_{max}+\alpha_{max}+1)^2}{\beta}\vee 18(\rho-1)\tag{B.35}$$

if we define $\Upsilon_3=(36\rho\Upsilon^2_{\bar\epsilon,\tilde\nu,\eta}\Upsilon_2\vee 36\rho\Upsilon_1^2)/(\gamma_B\wedge\eta)$. Then, we obtain

$$\begin{aligned}\Theta_\omega^{t+1}-\Theta_\omega^t\leq&-\frac{\omega\beta(\gamma_B\wedge\eta)}{36}\cdot\bar\alpha_t\left\|\begin{pmatrix}\bar\Delta\boldsymbol{x}_t\\ J_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2\\&-\frac{\omega\beta(\gamma_B\wedge\eta)}{36\Upsilon_1^2(\kappa_{grad}\alpha_{max}+1)^2}\cdot\bar\alpha_t\|\nabla\mathcal{L}^t_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\|^2-\frac{\omega}{36}\bar\delta_t.\end{aligned}\tag{B.36}$$

Case 2a, unreliable step, $\check{\bar\Delta}_t=\bar\Delta_t$ By Lemma 8, we have

$$\begin{aligned}\mathcal{L}^{t+1}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}-\mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}&\leq\frac{\bar\alpha_t\beta}{2}(\bar\nabla\mathcal{L}^t_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta})^T\bar\Delta_t\\&\overset{(B.24)}{\leq}-\frac{\bar\alpha_t\beta(\gamma_B\wedge\eta)}{8}\left\|\begin{pmatrix}\bar\Delta\boldsymbol{x}_t\\ J_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2\\&\overset{(B.32)}{\leq}-\frac{\bar\alpha_t\beta(\gamma_B\wedge\eta)}{16}\left\|\begin{pmatrix}\bar\Delta\boldsymbol{x}_t\\ J_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2\\&\quad-\frac{\bar\alpha_t\beta(\gamma_B\wedge\eta)}{16\Upsilon_1^2(\kappa_{grad}\alpha_{max}+1)^2}\|\nabla\mathcal{L}^t_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\|^2.\end{aligned}$$

By Line 22 of Algorithm 1, $\bar\delta_{t+1}-\bar\delta_t=-(1-1/\rho)\bar\delta_t$, while (B.33) still holds. Thus, under (B.35), we have

$$\begin{aligned}\Theta_\omega^{t+1}-\Theta_\omega^t\leq&-\frac{\omega\beta(\gamma_B\wedge\eta)}{36}\cdot\bar\alpha_t\left\|\begin{pmatrix}\bar\Delta\boldsymbol{x}_t\\ J_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t+\Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2\\&-\frac{\omega\beta(\gamma_B\wedge\eta)}{36\Upsilon_1^2(\kappa_{grad}\alpha_{max}+1)^2}\cdot\bar\alpha_t\|\nabla\mathcal{L}^t_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\|^2-\frac{1}{2}(1-\omega)\left(1-\frac{1}{\rho}\right)\bar\delta_t.\end{aligned}\tag{B.37}$$

Case 3a, unsuccessful step, $\check{\Delta}_t = \bar{\Delta}_t$ In this case, $(\boldsymbol{x}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$, $\bar{\alpha}_{t+1} = \bar{\alpha}_t/\rho$ and $\bar{\delta}_{t+1} = \bar{\delta}_t/\rho$. Thus, we immediately have

$$\Theta_\omega^{t+1} - \Theta_\omega^t \le -\frac{1}{2}(1-\omega)\left(1-\frac{1}{\rho}\right)\left(\bar{\alpha}_t\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 + \bar{\delta}_t\right). \tag{B.38}$$

Combining (B.36), (B.37), (B.38), and noting that

$$\frac{\omega\beta(\gamma_B\wedge\eta)}{36\Upsilon_1^2(\kappa_{grad}\alpha_{max}+1)^2} \ge \frac{1-\omega}{2}\left(1-\frac{1}{\rho}\right) \Longleftarrow \frac{\omega}{1-\omega} \ge \frac{18\Upsilon_1^2(\kappa_{grad}\alpha_{max}+1)^2}{\beta(\gamma_B\wedge\eta)},$$
$$\frac{\omega}{36} \ge \frac{1-\omega}{2}\left(1-\frac{1}{\rho}\right) \Longleftarrow \frac{\omega}{1-\omega} \ge 18(\rho-1),$$

with the right hand side being implied by (B.34) and further by (B.35), we know (B.38) holds for all three cases with $\check{\Delta}_t = \bar{\Delta}_t$.

Case 1b, reliable step, $\check{\Delta}_t = \hat{\Delta}_t$ By Lemma 8, we have

$$\begin{aligned}\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} &\le \frac{\bar{\alpha}_t\beta}{2}(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\hat{\Delta}_t \\ &\overset{(27)}{\le} \frac{\bar{\alpha}_t\beta}{3}(\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T\hat{\Delta}_t - \frac{\bar{\delta}_t}{6} \overset{(20)}{\le} -\frac{\bar{\alpha}_t\beta}{3\chi_u}\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \frac{\bar{\delta}_t}{6}.\end{aligned}$$

Note that

$$\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\| \le \bar{\Delta}(\nabla\mathcal{L}^t_\eta) + \|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\| \overset{(14)}{\le} \kappa_{grad}\bar{\alpha}_t\bar{R}_t + \|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|.$$

Combining the above display with (B.28) and using $\bar{\alpha}_t \le \alpha_{max}$, there exists $\Upsilon_4 > 0$ such that

$$\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\| \le \Upsilon_4(\kappa_{grad}\alpha_{max}+1)\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|. \tag{B.39}$$

Combining the above three displays,

$$\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} \le -\frac{\bar{\alpha}_t\beta}{6\chi_u}\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \frac{\bar{\alpha}_t\beta}{6\Upsilon_4^2\chi_u(\kappa_{grad}\alpha_{max}+1)^2}\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \frac{\bar{\delta}_t}{6}.$$

By Line 20 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = (\rho-1)\bar{\delta}_t$. By the Taylor expansion and $\bar{\alpha}_{t+1} \le \rho\bar{\alpha}_t$ (Line 18),

$$\begin{aligned}\bar{\alpha}_{t+1}\|\nabla\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \bar{\alpha}_t\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 &\le 2\rho\bar{\alpha}_t\left\{\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 + \Upsilon^2_{\tilde{\epsilon},\tilde{\nu},\eta}\bar{\alpha}_t^2\|\hat{\Delta}_t\|^2\right\} \\ &\overset{(20)}{\le} 2\rho\bar{\alpha}_t\left\{\|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 + \Upsilon^2_{\tilde{\epsilon},\tilde{\nu},\eta}\chi_u^2\alpha_{max}^2\|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2\right\}.\end{aligned} \tag{B.40}$$

Combining the above two displays,

$$\begin{aligned}\Theta_\omega^{t+1}-\Theta_\omega^t \le &-\left(\frac{\omega\beta}{6\chi_u}-(1-\omega)\rho\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}^2\chi_u^2\alpha_{max}^2\right)\bar{\alpha}_t\|\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2\\ &-\left(\frac{\omega\beta}{6\Upsilon_4^2\chi_u(\kappa_{grad}\alpha_{max}+1)^2}-(1-\omega)\rho\right)\bar{\alpha}_t\|\nabla\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2\\ &-\left(\frac{\omega}{6}-\frac{(1-\omega)(\rho-1)}{2}\right)\bar{\delta}_t.\end{aligned}$$

Let

$$\begin{aligned}\frac{\omega\beta}{12\chi_u}\ge(1-\omega)\rho\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}^2\chi_u^2\alpha_{max}^2 &\iff \frac{\omega}{1-\omega}\ge\frac{12\rho\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}^2\chi_u^3\alpha_{max}^2}{\beta},\\ \frac{\omega\beta}{12\Upsilon_4^2\chi_u(\kappa_{grad}\alpha_{max}+1)^2}\ge(1-\omega)\rho &\iff \frac{\omega}{1-\omega}\ge\frac{12\rho\Upsilon_4^2\chi_u(\kappa_{grad}\alpha_{max}+1)^2}{\beta},\\ \frac{\omega}{12}\ge\frac{1-\omega}{2}(\rho-1) &\iff \frac{\omega}{1-\omega}\ge 6(\rho-1),\end{aligned}\tag{B.41}$$

which is implied by (B.35) if we re-define $\Upsilon_3 \leftarrow \Upsilon_3 \vee 12\rho\Upsilon_{\tilde{\epsilon},\tilde{\nu},\eta}^2\chi_u^3 \vee 12\rho\Upsilon_4^2\chi_u$. Then,

$$\begin{aligned}\Theta_\omega^{t+1}-\Theta_\omega^t\le &-\frac{\omega\beta}{12\chi_u}\cdot\bar{\alpha}_t\|\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2\\ &-\frac{\omega\beta}{12\Upsilon_4^2\chi_u(\kappa_{grad}\alpha_{max}+1)^2}\cdot\bar{\alpha}_t\|\nabla\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2-\frac{\omega}{12}\bar{\delta}_t.\end{aligned}\tag{B.42}$$

Case 2b, unreliable step, $\check{\Delta}_t=\hat{\Delta}_t$ By Lemma 8, we have

$$\begin{aligned}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t+1}-\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t}\le\frac{\bar{\alpha}_t\beta}{2}(\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t)^T\hat{\Delta}_t &\overset{(20)}{\le}-\frac{\bar{\alpha}_t\beta}{2\chi_u}\|\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2\\ &\overset{(B.39)}{\le}-\frac{\bar{\alpha}_t\beta}{4\chi_u}\|\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2-\frac{\bar{\alpha}_t\beta}{4\Upsilon_4^2\chi_u\left(\kappa_{grad}\alpha_{max}+1\right)^2}\|\nabla\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2.\end{aligned}$$

By Line 22 of Algorithm 1, $\bar{\delta}_{t+1}-\bar{\delta}_t=-(1-1/\rho)\bar{\delta}_t$, while (B.40) still holds. Thus, under (B.35), we have

$$\begin{aligned}\Theta_\omega^{t+1}-\Theta_\omega^t\le &-\frac{\omega\beta}{12\chi_u}\cdot\bar{\alpha}_t\|\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2\\ &-\frac{\omega\beta}{12\Upsilon_4^2\chi_u(\kappa_{grad}\alpha_{max}+1)^2}\cdot\bar{\alpha}_t\|\nabla\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^t\|^2-\frac{1}{2}(1-\omega)\left(1-\frac{1}{\rho}\right)\bar{\delta}_t.\end{aligned}\tag{B.43}$$

Case 3b, unsuccessful step, $\check{\Delta}_t = \hat{\Delta}_t$ In this case, (B.38) holds. Combining (B.42), (B.43), (B.38), and noting that

$$\frac{\omega\beta}{12\Upsilon_4^2\chi_u(\kappa_{grad}\alpha_{max}+1)^2} \geq \frac{1-\omega}{2}\left(1-\frac{1}{\rho}\right) \Longleftarrow \frac{\omega}{1-\omega} \geq \frac{6\Upsilon_4^2\chi_u(\kappa_{grad}\alpha_{max}+1)^2}{\beta},$$
$$\frac{\omega}{12} \geq \frac{1-\omega}{2}\left(1-\frac{1}{\rho}\right) \Longleftarrow \frac{\omega}{1-\omega} \geq 6(\rho-1),$$

as implied by (B.41) and further by (B.35), we know (B.38) holds for all three cases with $\check{\Delta}_t = \hat{\Delta}_t$. In summary, under (B.35), (B.38) holds for all cases. This completes the proof.

B.6 Proof of Lemma 10

The proof follows the proof of Lemma 9, except that (B.32) and (B.39) do not hold due to $(\mathcal{E}_1^t)^c$. We consider the following six cases.
Case 1a, reliable step, $\check{\Delta}_t = \bar{\Delta}_t$ By Lemma 8, we have

$$\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t+1} - \mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t} \leq \frac{\bar{\alpha}_t\beta}{2}(\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t})^T\bar{\Delta}_t \overset{(27)}{\leq} \frac{4\bar{\alpha}_t\beta}{9}(\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t})^T\bar{\Delta}_t - \frac{\bar{\delta}_t}{18}$$
$$\overset{(B.24)}{\leq} -\frac{\bar{\alpha}_t\beta(\gamma_B\wedge\eta)}{9}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2 - \frac{\bar{\delta}_t}{18}.$$

By Line 20 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = (\rho-1)\bar{\delta}_t$, while (B.33) still holds. By the condition of ω in (B.34) and (B.35), we know that under (32) (which implies (B.35)),

$$\Theta_\omega^{t+1} - \Theta_\omega^t \leq -\frac{\omega\beta(\gamma_B\wedge\eta)}{36}\cdot\bar{\alpha}_t\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t\\ G_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2$$
$$+\rho(1-\omega)\bar{\alpha}_t\|\nabla\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t}\|^2 - \frac{\omega}{36}\bar{\delta}_t. \tag{B.44}$$

Case 2a, unreliable step, $\check{\Delta}_t = \bar{\Delta}_t$ By Lemma 8, we have

$$\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t+1} - \mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t} \leq \frac{\bar{\alpha}_t\beta}{2}(\bar{\nabla}\mathcal{L}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}^{t})^T\bar{\Delta}_t$$
$$\overset{(B.24)}{\leq} -\frac{\bar{\alpha}_t\beta(\gamma_B\wedge\eta)}{8}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ J_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_tG_t\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2.$$

By Line 22 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = -(1-1/\rho)\bar{\delta}_t$, while (B.33) still holds. Thus, under (32),

$$\Theta_\omega^{t+1} - \Theta_\omega^t \le -\frac{\omega\beta(\gamma_B \wedge \eta)}{36} \cdot \bar{\alpha}_t \left\| \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t \\ G_t \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t) \end{pmatrix} \right\|^2$$
$$+\rho(1-\omega)\bar{\alpha}_t \|\nabla \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \frac{1}{2}(1-\omega)\left(1 - \frac{1}{\rho}\right)\bar{\delta}_t. \quad \text{(B.45)}$$

Case 3a, unsuccessful step, $\check{\Delta}_t = \bar{\Delta}_t$ In this case, (B.38) holds. Combining (B.44), (B.45), and (B.38), we have

$$\Theta_\omega^{t+1} - \Theta_\omega^t \le \rho(1-\omega)\bar{\alpha}_t \|\nabla \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2. \quad \text{(B.46)}$$

Case 1b, reliable step, $\check{\Delta}_t = \hat{\Delta}_t$ By Lemma 8, we have

$$\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} \le \frac{\bar{\alpha}_t\beta}{2}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \hat{\Delta}_t$$
$$\overset{(27)}{\le} \frac{\bar{\alpha}_t\beta}{3}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \hat{\Delta}_t - \frac{\bar{\delta}_t}{6} \overset{(20)}{\le} -\frac{\bar{\alpha}_t\beta}{3\chi_u}\|\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \frac{\bar{\delta}_t}{6}.$$

By Line 20 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = (\rho - 1)\bar{\delta}_t$, while (B.40) still holds. By the condition of ω in (B.41), we know that under (32) (which implies (B.41)),

$$\Theta_\omega^{t+1} - \Theta_\omega^t \le -\frac{\omega\beta}{12\chi_u} \cdot \bar{\alpha}_t \|\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 + \rho(1-\omega)\bar{\alpha}_t\|\nabla\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \frac{\omega}{12}\bar{\delta}_t. \quad \text{(B.47)}$$

Case 2b, unreliable step, $\check{\Delta}_t = \hat{\Delta}_t$ By Lemma 8, we have

$$\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta} \le \frac{\bar{\alpha}_t\beta}{2}(\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta})^T \hat{\Delta}_t \overset{(20)}{\le} -\frac{\bar{\alpha}_t\beta}{2\chi_u}\|\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2.$$

By Line 22 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = -(1-1/\rho)\bar{\delta}_t$, while (B.40) still holds. Thus, under (32),

$$\Theta_\omega^{t+1} - \Theta_\omega^t \le -\frac{\omega\beta}{12\chi_u} \cdot \bar{\alpha}_t \|\bar{\nabla}\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2$$
$$+\rho(1-\omega)\bar{\alpha}_t\|\nabla\mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}},\bar{\nu}_{\bar{t}},\eta}\|^2 - \frac{1-\omega}{2}\left(1-\frac{1}{\rho}\right)\bar{\delta}_t. \quad \text{(B.48)}$$

Case 3b, unsuccessful step, $\check{\Delta}_t = \hat{\Delta}_t$ In this case, (B.38) holds. Combining (B.47), (B.48), and (B.38), we note that (B.46) holds as well. Thus, (B.46) holds for all six cases. This completes the proof.

B.7 Proof of Lemma 11

The proof follows the proof of Lemma 10, except that Lemma 8 is not applicable. We consider the following six cases.
Case 1a, reliable step, $\check{\Delta}_t = \bar{\Delta}_t$ We have

$$\begin{aligned}
\mathcal{L}^{t+1}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} &\leq \bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} + \left|\bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| \\
&\leq \bar\alpha_t\beta(\bar\nabla\mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta})^T\bar{\Delta}_t + \left|\bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| \\
&\overset{(27)}{\leq} \frac{4\bar\alpha_t\beta}{5}(\bar\nabla\mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta})^T\bar{\Delta}_t - \frac{\bar\delta_t}{5} + \left|\bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| \\
&\overset{(B.24)}{\leq} -\frac{\bar\alpha_t\beta(\gamma_B\wedge\eta)}{5}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t \\ J_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2 - \frac{\bar\delta_t}{5} \\
&\quad + \left|\bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right|.
\end{aligned}$$

By Line 20 of Algorithm 1, $\bar\delta_{t+1} - \bar\delta_t = (\rho-1)\bar\delta_t$, while (B.33) still holds. By the condition of ω in (B.34) and (B.35), we know that under (32) (which implies (B.35)),

$$\begin{aligned}
\Theta^{t+1}_\omega - \Theta^t_\omega \leq & -\frac{\omega\beta(\gamma_B\wedge\eta)}{36}\cdot\bar\alpha_t\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t \\ J_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2 \\
& +\rho(1-\omega)\bar\alpha_t\|\nabla\mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\|^2 \\
& +\omega\left\{\left|\bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right|\right\} - \frac{\omega}{36}\bar\delta_t. \quad \text{(B.49)}
\end{aligned}$$

Case 2a, unreliable step, $\check{\Delta}_t = \bar{\Delta}_t$ We have

$$\begin{aligned}
\mathcal{L}^{t+1}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} &\leq \bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} + \left|\bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| \\
&\overset{(26)}{\leq} \bar\alpha_t\beta(\bar\nabla\mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta})^T\bar{\Delta}_t + \left|\bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| \\
&\overset{(B.24)}{\leq} -\frac{\bar\alpha_t\beta(\gamma_B\wedge\eta)}{4}\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t \\ J_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t \\ G_t\bar\nabla_{\boldsymbol{x}}\mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t)\end{pmatrix}\right\|^2 \\
&\quad + \left|\bar{\mathcal{L}}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{s_t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right| \\
&\quad + \left|\bar{\mathcal{L}}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta} - \mathcal{L}^{t}_{\bar\epsilon_{\bar t},\bar\nu_{\bar t},\eta}\right|.
\end{aligned}$$

By Line 22 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = -(1 - 1/\rho)\bar{\delta}_t$, while (B.33) still holds. Thus, under (32),

$$
\begin{aligned}
\Theta_\omega^{t+1} - \Theta_\omega^t \le & -\frac{\omega\beta(\gamma_B \wedge \eta)}{36} \cdot \bar{\alpha}_t \left\| \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ J_t \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t \\ G_t \bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t + \Pi_c(\mathrm{diag}^2(g_t)\boldsymbol{\lambda}_t) \end{pmatrix} \right\|^2 \\
& -\frac{1}{2}(1-\omega)\left(1 - \frac{1}{\rho}\right)\bar{\delta}_t \\
& +\omega \left\{ \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| \right\} \\
& +\rho(1-\omega)\bar{\alpha}_t \|\nabla \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|^2. \qquad \text{(B.50)}
\end{aligned}
$$

Case 3a, unsuccessful step, $\check{\Delta}_t = \bar{\Delta}_t$ In this case, (B.38) holds. Combining (B.49), (B.50), and (B.38), we obtain

$$
\begin{aligned}
\Theta_\omega^{t+1} - \Theta_\omega^t \le \omega & \left\{ \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| \right\} \\
& +\rho(1-\omega)\bar{\alpha}_t \|\nabla \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|^2. \qquad \text{(B.51)}
\end{aligned}
$$

Case 1b, reliable step, $\check{\Delta}_t = \hat{\Delta}_t$ We have

$$
\begin{aligned}
\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} & \le \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} + \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| \\
& \le \bar{\alpha}_t \beta (\bar{\nabla} \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta})^T \hat{\Delta}_t + \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| \\
& \overset{(27)}{\le} \frac{\bar{\alpha}_t \beta}{2} (\bar{\nabla} \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta})^T \hat{\Delta}_t - \frac{\bar{\delta}_t}{2} + \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| \\
& \overset{(20)}{\le} -\frac{\bar{\alpha}_t \beta}{2\chi_u} \|\bar{\nabla} \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|^2 - \frac{\bar{\delta}_t}{2} + \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right|.
\end{aligned}
$$

By Line 20 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = (\rho - 1)\bar{\delta}_t$, while (B.40) still holds. By the condition of ω in (B.41), we know that under (32) (which implies (B.41)),

$$
\begin{aligned}
\Theta_\omega^{t+1} - \Theta_\omega^t \le & -\frac{\omega\beta}{12\chi_u} \cdot \bar{\alpha}_t \|\bar{\nabla} \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|^2 + \omega \left\{ \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| \right\} \\
& +\rho(1-\omega)\bar{\alpha}_t \|\nabla \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|^2 - \frac{\omega}{12}\bar{\delta}_t. \qquad \text{(B.52)}
\end{aligned}
$$

Case 2b, unreliable step, $\check{\Delta}_t = \hat{\Delta}_t$ We have

$$
\begin{aligned}
\mathcal{L}^{t+1}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} & \le \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} + \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| \\
& \le \bar{\alpha}_t \beta (\bar{\nabla} \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta})^T \hat{\Delta}_t + \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| \\
& \overset{(20)}{\le} -\frac{\bar{\alpha}_t \beta}{\chi_u} \|\bar{\nabla} \mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|^2 + \left| \bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right| + \left| \bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} \right|.
\end{aligned}
$$

By Line 22 of Algorithm 1, $\bar{\delta}_{t+1} - \bar{\delta}_t = -(1-1/\rho)\bar{\delta}_t$, while (B.40) still holds. Thus, under (32),

$$\begin{aligned}\Theta_\omega^{t+1} - \Theta_\omega^t \leq & -\frac{\omega\beta}{12\chi_u} \cdot \bar{\alpha}_t \|\bar{\nabla}\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|^2 + \omega\left\{\left|\bar{\mathcal{L}}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{s_t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\right| + \left|\bar{\mathcal{L}}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta} - \mathcal{L}^{t}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\right|\right\} \\ & + \rho(1-\omega)\bar{\alpha}_t \|\nabla\mathcal{L}^t_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|^2 - \frac{1}{2}(1-\omega)\left(1 - \frac{1}{\rho}\right)\bar{\delta}_t. \end{aligned} \tag{B.53}$$

Case 3b, unsuccessful step, $\check{\Delta}_t = \hat{\Delta}_t$ In this case, (B.38) holds. Combining (B.52), (B.53), and (B.38), we note that (B.51) holds as well. Thus, (B.51) holds for all six cases. This completes the proof.

B.8 Proof of Theorem 3

We suppose there are infinite many successful steps. Otherwise, $\bar{\alpha}_t$ decreases to zero (cf. Line 25 of Algorithm 1) and the argument holds trivially. We use $\bar{t} < t_1 < t_2 < \ldots$ to denote the subsequence with $t_i - 1, \forall i \geq 1$, being a successful step. By Lemma 14, Lemma 6(b), and (B.14), there exist $\Upsilon_1, \Upsilon_2 > 0$ such that for any $i \geq 1$,

$$R_{t_i} \overset{\text{Lem. } 14}{\leq} \Upsilon_1 \left\|\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_{t_i} \\ c_{t_i} \\ \boldsymbol{w}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}}\end{pmatrix}\right\| \overset{\text{Lem. } 6(b)}{\leq} \Upsilon_2 \left\{\|\nabla\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\| + \left\|\begin{pmatrix} c_{t_i} \\ \boldsymbol{w}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}}\end{pmatrix}\right\|\right\}.$$

Since $t_i \geq \bar{t} + 1$, two parameters $\bar{\epsilon}_{\bar{t}}$, $\bar{\nu}_{\bar{t}}$ are fixed conditional on any σ-algebra $\mathcal{F} \supseteq \mathcal{F}_{\bar{t}}$. Thus, for any $i \geq 1$,

$$\begin{aligned} & \left\|\begin{pmatrix} c_{t_i} \\ \boldsymbol{w}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}}\end{pmatrix}\right\| = \mathbb{E}\left[\left\|\begin{pmatrix} c_{t_i} \\ \boldsymbol{w}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}}\end{pmatrix}\right\| \mid \mathcal{F}_{t_i - 1}\right] \\ & = \mathbb{E}\left[\left\|\begin{pmatrix} c_{t_i} \\ \boldsymbol{w}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}}\end{pmatrix}\right\| \mathbf{1}_{\chi_{err}\|\bar{\nabla}\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\| \leq \bar{R}_{t_i}} \mid \mathcal{F}_{t_i - 1}\right] + \mathbb{E}\left[\left\|\begin{pmatrix} c_{t_i} \\ \boldsymbol{w}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}}\end{pmatrix}\right\| \mathbf{1}_{\bar{R}_{t_i} < \chi_{err}\|\bar{\nabla}\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|} \mid \mathcal{F}_{t_i - 1}\right] \\ & \overset{(17)}{\leq} \mathbb{E}\left[\chi_{err}\|\bar{\nabla}\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\| \cdot \mathbf{1}_{\chi_{err}\|\bar{\nabla}\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\| \leq \bar{R}_{t_i}} \mid \mathcal{F}_{t_i - 1}\right] \\ & \quad + (\bar{\epsilon}_{\bar{t}} q^{t_i}_{\bar{\nu}_{\bar{t}}} \vee 1)\mathbb{E}\left[\left\|\begin{pmatrix} c_{t_i} \\ \max\{g_{t_i}, -\boldsymbol{\lambda}_{t_i}\}\end{pmatrix}\right\| \cdot \mathbf{1}_{\bar{R}_{t_i} < \chi_{err}\|\bar{\nabla}\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|} \mid \mathcal{F}_{t_i - 1}\right] \quad \text{(also use Lemma 14)} \\ & \overset{(B.14)}{\leq} \chi_{err}\mathbb{E}\left[\|\bar{\nabla}\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\| \mid \mathcal{F}_{t_i - 1}\right] + (\bar{\epsilon}_0\tilde{\nu} \vee 1)\mathbb{E}\left[\bar{R}_{t_i} \cdot \mathbf{1}_{\bar{R}_{t_i} < \chi_{err}\|\bar{\nabla}\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\|} \mid \mathcal{F}_{t_i - 1}\right] \\ & \leq \{1 + (\bar{\epsilon}_0\tilde{\nu} \vee 1)\}\chi_{err}\mathbb{E}\left[\|\bar{\nabla}\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\| \mid \mathcal{F}_{t_i - 1}\right] \\ & \leq \{1 + (\bar{\epsilon}_0\tilde{\nu} \vee 1)\}\chi_{err}\left\{\|\nabla\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\| + \mathbb{E}\left[\|\bar{\Delta}(\nabla\mathcal{L}^{t_i}_{\eta})\| \mid \mathcal{F}_{t_i - 1}\right]\right\} \\ & \overset{(16)}{\leq} \{1 + (\bar{\epsilon}_0\tilde{\nu} \vee 1)\}\chi_{err}\left\{\|\nabla\mathcal{L}^{t_i}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\| + \chi_{grad}\sqrt{\bar{\delta}_{t_i}/\bar{\alpha}_{t_i}}\right\}. \end{aligned}$$

Combining the above two displays, we know there exists $\Upsilon_3 > 0$ such that

$$R_{t_i} \leq \Upsilon_3(\chi_{grad} + 1)\left\{\|\nabla \mathcal{L}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}^{t_i}\| + \sqrt{\bar{\delta}_{t_i}/\bar{\alpha}_{t_i}}\right\},$$

which implies

$$\bar{\alpha}_{t_i} R_{t_i}^2 \leq 2\Upsilon_3^2(\chi_{grad} + 1)^2\left\{\bar{\alpha}_{t_i}\|\nabla \mathcal{L}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}^{t_i}\|^2 + \bar{\delta}_{t_i}\right\}. \tag{B.54}$$

On the other hand, by Theorem 2, we sum up the error recursion for $t \geq \bar{t} + 1$, take conditional expectation on $\mathcal{F}_{\bar{t}}$, and have

$$\begin{aligned}
&\sum_{t=\bar{t}+1}^{\infty} \mathbb{E}[\bar{\alpha}_t\|\nabla \mathcal{L}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}^{t}\|^2 + \bar{\delta}_t \mid \mathcal{F}_{\bar{t}}] \\
&\quad \leq \frac{4\rho}{(1 - p_{grad})(1 - p_f)(1 - \omega)(\rho - 1)} \sum_{t=\bar{t}+1}^{\infty} \mathbb{E}[\Theta_{\omega}^{t} \mid \mathcal{F}_{\bar{t}}] - \mathbb{E}\left[\Theta_{\omega}^{t+1} \mid \mathcal{F}_{\bar{t}}\right] \\
&\quad \leq \frac{4\rho}{(1 - p_{grad})(1 - p_f)(1 - \omega)(\rho - 1)}\left(\Theta_{\omega}^{\bar{t}+1} - \min_{\mathcal{X}\times\mathcal{M}\times\Lambda} \omega\mathcal{L}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}\right) < \infty.
\end{aligned} \tag{B.55}$$

Thus, applying the Fubini's theorem to exchange the summation and expectation, we know that $\mathbb{E}[\limsup_{t\to\infty} \bar{\alpha}_t\|\nabla \mathcal{L}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}^{t}\|^2 + \bar{\delta}_t \mid \mathcal{F}_{\bar{t}}] = 0$. Since $\bar{\alpha}_t\|\nabla \mathcal{L}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}^{t}\|^2 + \bar{\delta}_t$ is non-negative, we further obtain $\bar{\alpha}_t\|\nabla \mathcal{L}_{\bar{\epsilon}_{\bar{t}}, \bar{\nu}_{\bar{t}}, \eta}^{t}\|^2 + \bar{\delta}_t \to 0$ as $t \to \infty$ almost surely. By (B.54), we have $\bar{\alpha}_{t_i} R_{t_i}^2 \to 0$ as $i \to \infty$. Noting that $\bar{\alpha}_t R_t^2 \leq \bar{\alpha}_{t_i} R_{t_i}^2$ for any $t_i \leq t < t_{i+1}$, we complete the proof.

B.9 Proof of Theorem 4

We adapt the proof of [40, Theorem 4]. By Theorem 3, it suffices to show that the "limsup" of the random stepsize sequence $\{\bar{\alpha}_t\}_t$ is lower bounded away from zero. To show this, we define two stepsize sequences as follows. For any $t > \bar{t} + 1$, we let

$$\begin{aligned}
\phi_t &= \log(\bar{\alpha}_t), \\
\varphi_t &= \min\{\log(c), \mathbf{1}_{\mathcal{E}_1^{t-1}\cap\mathcal{E}_2^{t-1}}(\log(\rho) + \varphi_{t-1}) + (1 - \mathbf{1}_{\mathcal{E}_1^{t-1}\cap\mathcal{E}_2^{t-1}})(\varphi_{t-1} - \log(\rho))\},
\end{aligned}$$

and let $\phi_{\bar{t}+1} = \varphi_{\bar{t}+1} = \log(\bar{\alpha}_{\bar{t}+1})$. Here, c is a deterministic constant such that

$$c \leq \frac{1 - \beta}{\Upsilon_1(\kappa_{grad} + \kappa_f + 1)} \wedge \alpha_{max}$$

and $c = \rho^{-i}\alpha_{max}$ for some $i > 0$. The first constant comes from Lemma 7. We aim to show $\phi_t \geq \varphi_t, \forall t \geq \bar{t} + 1$.

First, we note that by the stepsize specification in Lines 18 and 25 of Algorithm 1 (Line 13 is not performed since $t \geq \bar{t}+1$), $\bar{\alpha}_t = \rho^{j_t}c$ for some integer j_t. Second, we note that ϕ_t and φ_t are both $\mathcal{F}_{t-1}$-measurable, that is, they are fixed conditional on $\mathcal{F}_{t-1}$. Third, we show that $\phi_t \geq \varphi_t$ by induction. Note that $\phi_{\bar{t}+1} = \varphi_{\bar{t}+1}$. Suppose $\phi_t \geq \varphi_t$, we consider the following three cases.

(a) If $\phi_t > \log(c)$, then $\phi_t \geq \log(c) + \log(\rho)$. Thus, $\phi_{t+1} \geq \phi_t - \log(\rho) \geq \log(c) \geq \varphi_{t+1}$.

(b) If $\phi_t \leq \log(c)$ and $\mathbf{1}_{\mathcal{E}_1^t \cap \mathcal{E}_2^t} = 1$, then Lemma 7 leads to

$$\phi_{t+1} = \min\{\log(\alpha_{max}), \phi_t + \log(\rho)\} \geq \min\{\log(c), \varphi_t + \log(\rho)\} = \varphi_{t+1}.$$

(c) If $\phi_t \leq \log(c)$ and $\mathbf{1}_{\mathcal{E}_1^t \cap \mathcal{E}_2^t} = 0$, then

$$\phi_{t+1} \geq \phi_t - \log(\rho) \geq \varphi_t - \log(\rho) \geq \varphi_{t+1}.$$

Combining the above three cases, we have $\phi_t \geq \varphi_t, \forall t \geq \bar{t}+1$. Note that, conditional on $\mathcal{F}_{\bar{t}}$, $\{\varphi_t\}_{t \geq \bar{t}+1}$ is a random walk with a maximum and a drift upward (cf. [24, Example 6.1.2]). Thus, $\limsup_{t\to\infty} \varphi_t \geq \log(c)$ almost surely. In particular, we have

$$\begin{aligned}
&P\left(\limsup_{t\to\infty} \phi_t \geq \log(c)\right)\\
&= \sum_{i=0}^{\infty} \int_{\mathcal{F}_i} P\left(\limsup_{t\to\infty} \phi_t \geq \log(c) \mid \mathcal{F}_i, \bar{t}=i\right) P\left(\mathcal{F}_i, \bar{t}=i\right)\\
&\overset{\phi_t \geq \varphi_t}{\geq} \sum_{i=0}^{\infty} \int_{\mathcal{F}_i} P\left(\limsup_{t\to\infty} \varphi_t \geq \log(c) \mid \mathcal{F}_i, \bar{t}=i\right) P\left(\mathcal{F}_i, \bar{t}=i\right)\\
&= \sum_{i=0}^{\infty} \int_{\mathcal{F}_i} P\left(\mathcal{F}_i, \bar{t}=i\right)\\
&= 1,
\end{aligned}$$

which means that the "limsup" of $\bar{\alpha}_t$ is lower bounded almost surely. Using Theorem 3, we complete the proof.

B.10 Proof of Theorem 5

Suppose $\limsup_{t\to\infty} R_t = \epsilon > 0$. By Theorem 4, we know there exist two sequences $\{n_i\}_i$ and $\{m_i\}_i$ with $n_i < m_i < n_{i+1}$ for all i, such that

$$R_{n_i} \geq \frac{2\epsilon}{3}, \quad R_t \geq \frac{\epsilon}{3}, \ t = n_i + 1, \ldots, m_i - 1, \quad R_{m_i} < \frac{\epsilon}{3}.$$

For each interval $[n_i, m_i]$, we use $\{t_{i,j}\}_{j=1}^{J_i}$ to denote a subsequence within the interval such that $n_i = t_{i,1} < \ldots < t_{i,j} < \ldots < t_{i,J_i} = m_i$ and $t_{i,j} - 1$ is a successful step.

In other words, $t_{i,j}$ is the first index that we arrive at the new point. Here, we suppose $n_i - 1$ is a successful step; that is, the index n_i is the first time we arrive at the point $(\boldsymbol{x}_{n_i}, \boldsymbol{\mu}_{n_i}, \boldsymbol{\lambda}_{n_i})$ (one can always choose n_i to satisfy this condition). We also note that $t_{i,J_i} = m_i$ because $R_{m_i-1} \geq \epsilon/3$ while $R_{m_i} < \epsilon/3$. With these notation, there exist $\Upsilon_1, \Upsilon_2 > 0$ such that

$$\begin{aligned}
\frac{\epsilon}{3} &\leq R_{n_i} - R_{m_i} \leq \sum_{t=n_i}^{m_i-1} |R_{t+1} - R_t| \\
&\leq \sum_{t=n_i}^{m_i-1} \left\| \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{t+1} - \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_{t+1} - c_t \\ \max\{g_{t+1}, -\boldsymbol{\lambda}_{t+1}\} - \max\{g_t, -\boldsymbol{\lambda}_t\} \end{pmatrix} \right\| \\
&\leq \Upsilon_1 \sum_{t=n_i}^{m_i-1} \|(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t, \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)\| \quad \text{(due to the Lip-continuity)} \\
&= \Upsilon_1 \sum_{j=2}^{J_i} \|(\boldsymbol{x}_{t_{i,j}} - \boldsymbol{x}_{t_{i,j}-1}, \boldsymbol{\mu}_{t_{i,j}} - \boldsymbol{\mu}_{t_{i,j}-1}, \boldsymbol{\lambda}_{t_{i,j}} - \boldsymbol{\lambda}_{t_{i,j}-1})\| \\
&= \Upsilon_1 \sum_{j=2}^{J_i} \bar{\alpha}_{t_{i,j}-1} \|\check{\Delta}_{t_{i,j}-1}\| \leq \Upsilon_2 \sum_{j=2}^{J_i} \bar{\alpha}_{t_{i,j}-1} \quad \text{(due to Assumption 3)} \\
&\leq \Upsilon_2 \sum_{j=1}^{J_i-1} \bar{\alpha}_{t_{i,j}} \quad \text{(due to Line 25 of Algorithm 1).}
\end{aligned} \tag{B.56}$$

Let us define the set $\mathcal{T} = \{t : t - 1 \text{ is successful and } R_t \geq \epsilon/3\}$. We can see from (B.54) and (B.55) that $\sum_{t\in\mathcal{T}} \bar{\alpha}_t < \infty$. This contradicts (B.56) since $\sum_{t\in\mathcal{T}} \bar{\alpha}_t \geq \sum_i \sum_{j=1}^{J_i-1} \bar{\alpha}_{t_{i,j}} \overset{\text{(B.56)}}{=} \infty$. Thus, we know $\limsup_{t\to\infty} R_t = 0$; and thus, we complete the proof.

C Auxiliary lemmas

Lemma 14 *Let* $\epsilon, \nu > 0$ *and* $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathcal{T}_\nu \times \mathbb{R}^r$. *Then*

$$\frac{\|\boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x}, \boldsymbol{\lambda})\|}{\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \vee 1} \leq \|\max\{g(\boldsymbol{x}), -\boldsymbol{\lambda}\}\| \leq \frac{\|\boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x}, \boldsymbol{\lambda})\|}{\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \wedge 1}.$$

Proof To prove Lemma 14, we require the following lemma. □

Lemma 15 *For any two scalars* a, b *and a scalar* $c > 0$, $|\max\{a, b\}| \leq \frac{1}{c\wedge 1} |\max\{a, cb\}|$.

Proof Without loss of generality, we assume $b \neq 0$ and $c \neq 1$. We consider four cases.

Case 1: $b > 0, c < 1$ If $a \le cb < b$, then $|\max\{a,b\}| = b = \frac{1}{c}|\max\{a,cb\}|$. If $cb < a \le b$, then $|\max\{a,b\}| = b \le \frac{1}{c}a = \frac{1}{c}|\max\{a,cb\}|$. If $cb < b < a$, then $|\max\{a,b\}| = a \le \frac{1}{c}|\max\{a,cb\}|$. Thus, the result holds.
Case 2: $b > 0, c > 1$ If $a \le b < cb$, then $|\max\{a,b\}| = b \le cb = |\max\{a,cb\}|$. If $b < a \le cb$, then $|\max\{a,b\}| = a \le cb = |\max\{a,cb\}|$. If $b < cb < a$, then $|\max\{a,b\}| = a = |\max\{a,cb\}|$. Thus, the result holds.
Case 3: $b < 0, c < 1$ If $a \le b < cb$, then $|\max\{a,b\}| = |b| = \frac{1}{c}|\max\{a,cb\}|$. If $b < a \le cb$, then $|\max\{a,b\}| = |a| \le |b| = \frac{1}{c}|\max\{a,cb\}|$. If $b < cb < a$, then $|\max\{a,b\}| = |a| \le \frac{|a|}{c} = \frac{1}{c}|\max\{a,cb\}|$. Thus, the result holds.
Case 4: $b < 0, c > 1$ If $a \le cb < b$, then $|\max\{a,b\}| = |b| \le c|b| = |\max\{a,cb\}|$. If $cb < a \le b$, then $|\max\{a,b\}| = |b| \le |a| = |\max\{a,cb\}|$. If $cb < b < a$, then $|\max\{a,b\}| = |a| = |\max\{a,cb\}|$. Thus, the result holds.
Combining the above four cases, we complete the proof. □

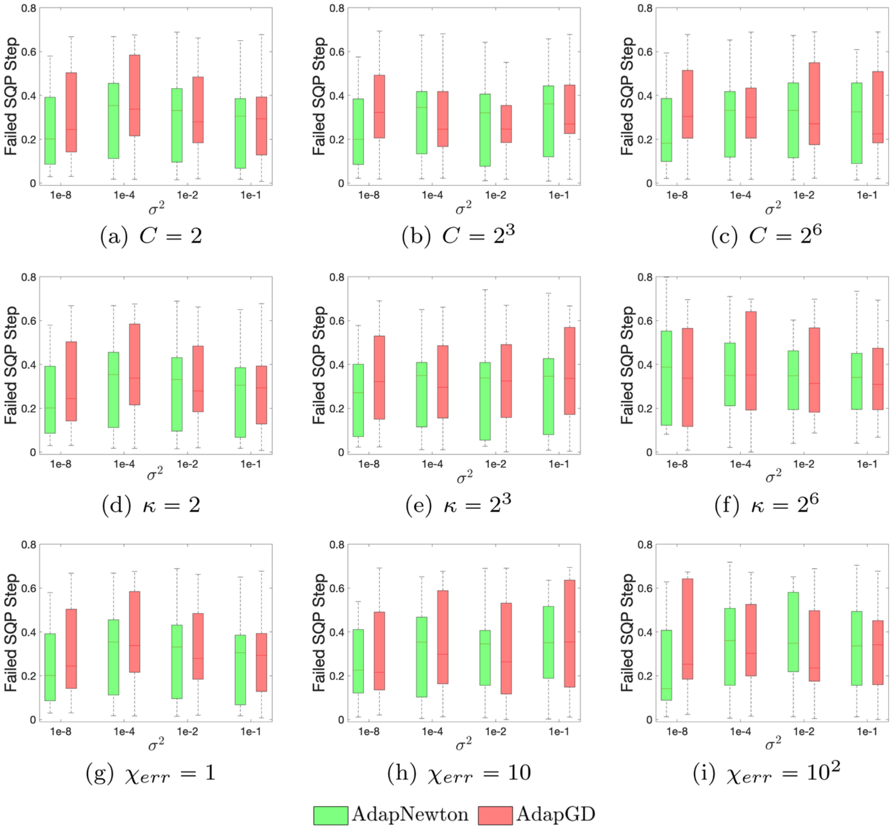
Since $\epsilon, \nu > 0$, $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathcal{T}_\nu \times \mathbb{R}^r$, and $q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) > 0$, we have for any $i \in \{1, 2, \dots, r\}$,

$$\begin{aligned}
&|(\boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x}, \boldsymbol{\lambda}))_i| \\
&\quad = |\max\{g_i(\boldsymbol{x}), -\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda})\boldsymbol{\lambda}_i\}| \le \frac{1}{\frac{1}{\epsilon q_\nu(\boldsymbol{x},\boldsymbol{\lambda})} \wedge 1}|\max\{g_i(\boldsymbol{x}), -\boldsymbol{\lambda}_i\}| \\
&\quad = (\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \vee 1) \cdot |\max\{g_i(\boldsymbol{x}), -\boldsymbol{\lambda}_i\}| \\
&\quad \le \frac{\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \vee 1}{\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \wedge 1}|\max\{g_i(\boldsymbol{x}), -\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda})\boldsymbol{\lambda}_i\}| \\
&\quad = \frac{\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \vee 1}{\epsilon q_\nu(\boldsymbol{x}, \boldsymbol{\lambda}) \wedge 1}\left|(\boldsymbol{w}_{\epsilon,\nu}(\boldsymbol{x}, \boldsymbol{\lambda}))_i\right|,
\end{aligned}$$

where both inequalities are from Lemma 15. Taking ℓ_2 norm on both sides, we finish the proof.

D Auxiliary experiments

We follow the experiments in Sect. 4 and provide additional results. We first examine three proportions: (1) the proportion of the iterations with failed SQP steps, (2) the proportion of the iterations with unstabilized penalty parameters, (3) the proportion of the iterations with a triggered feasibility error condition. We then investigate a multiplicative noise, and apply the method on an inequality constrained logistic regression problem.
Failed SQP steps Figure 5 plots the proportion of the iterations with failed SQP steps. From the figure, we see that the proportion varies from 10% to 60% across the problems, and AdapNewton tends to have a smaller proportion than AdapGD. Although the proportion does not have a clear dependency on the variance σ^2, the noticeable proportion of failed SQP steps illustrates the differences between equality and inequality constrained problems. As analyzed in Sect. 2, the active-set SQP steps may not be



(a) $C = 2$

(b) $C = 2^3$

(c) $C = 2^6$

(d) $\kappa = 2$

(e) $\kappa = 2^3$

(f) $\kappa = 2^6$

(g) $\chi_{err} = 1$

(h) $\chi_{err} = 10$

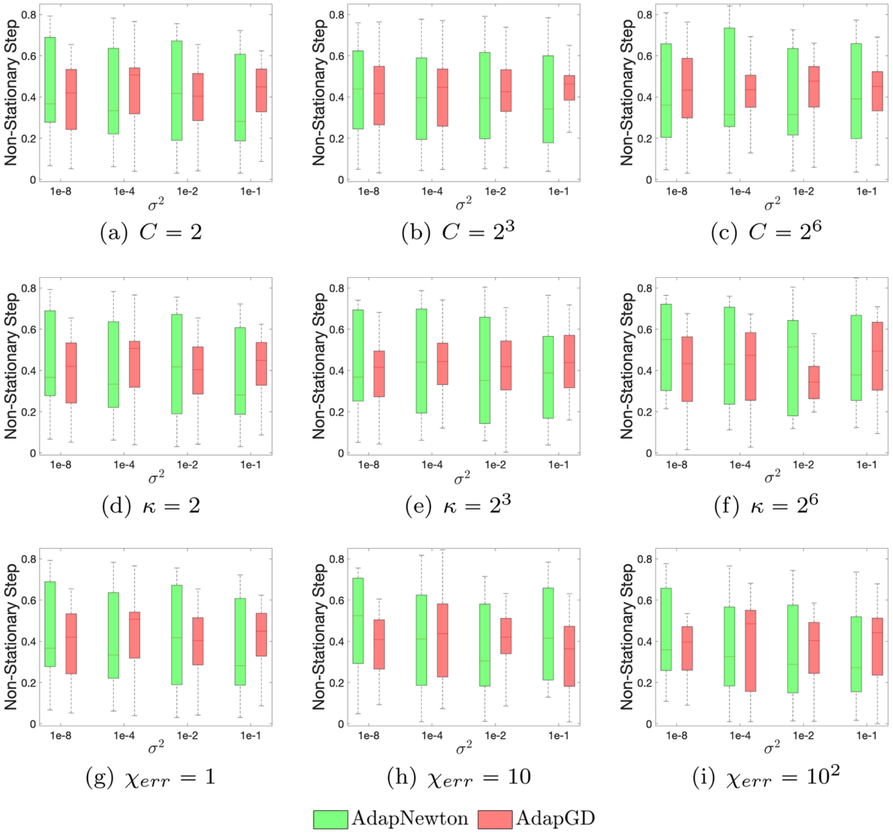
(i) $\chi_{err} = 10^2$

🟩 AdapNewton 🟥 AdapGD

Fig. 5 Failed SQP step boxplots. Each panel corresponds to a setup of (C, κ, χ_{err}). The default values are $C = \kappa = 2$ and $\chi_{err} = 1$. When we vary one parameter, the other two are set as default. Thus, the three figures on the left column are the same

informative if the identified active set is very distinct from the true active set. Due to the potential failure of the SQP steps, utilizing a safeguarding direction is critical in achieving the global convergence for the algorithm.

Non-stationary penalty parameters Figure 6 plots the proportion of the iterations with unstabilized penalty parameters; i.e., the last iteration that we update $\bar{\epsilon}_0$ over the total number of the iterations. From the figure, we observe that the proportion varies from 20% to 70%, and AdapNewton and AdapGD have comparable results. In fact, the proportion highly depends on the adopted initial $\bar{\epsilon}_0$ and the updating rule of $\bar{\epsilon}_0$. For example, a large ρ and a small $\bar{\epsilon}_0$ will reduce the proportion significantly; and the updating rules $\bar{\epsilon}_0 \leftarrow \bar{\epsilon}_0/\rho$ and $\bar{\epsilon}_0 \leftarrow \exp(-1/\bar{\epsilon}_0)$ will also lead to different proportions. The large variation in Fig. 6 suggests that different problems stabilize $\bar{\epsilon}_0$ to different levels; thus, a problem-dependent tuning of $\bar{\epsilon}_0$ is desired in practice. We note in the experiments that the results on some problems can be improved if $\bar{\epsilon}_0 = 10^{-4}$, while such a setup may not be suitable for other problems. Thus, designing a robust scheme to select the penalty parameters deserves further studying.



(a) $C = 2$ (b) $C = 2^3$ (c) $C = 2^6$

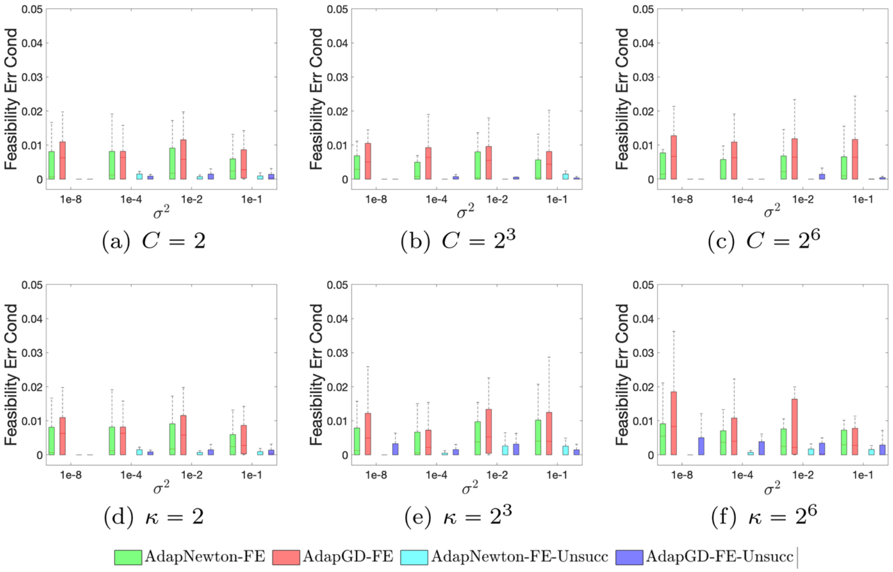
(d) $\kappa = 2$ (e) $\kappa = 2^3$ (f) $\kappa = 2^6$

(g) $\chi_{err} = 1$ (h) $\chi_{err} = 10$ (i) $\chi_{err} = 10^2$

Fig. 6 Unstabilized penalty parameter boxplots. Each panel corresponds to a setup of (C, κ, χ_{err}). The default values are $C = \kappa = 2$ and $\chi_{err} = 1$. When we vary one parameter, the other two are set as default. Thus, the three figures on the left column are the same

Feasibility error condition Figure 7 plots the proportion of the iterations with a triggered feasibility error condition. We do not show the results for the different setups of χ_{err}. In fact, when $\chi_{err} = 1$, the results are identical to $C = 2$ and $\kappa = 2$ (see the left column of Fig. 7). However, when $\chi_{err} = 10$ or 100, the feasibility error condition is *never* triggered. From Fig. 7, we see that the proportion is extremely small (e.g., as small as 1%). This suggests that the condition (17) is hardly triggered in practice. Figure 7 also plots the iteration proportion that (17) is triggered for an unsuccessful step. We see that such an proportion is even smaller (e.g., less than 0.5%). Given these negligible proportions, we can conclude that the condition (17) does not negatively affect the performance of the designed StoSQP scheme.

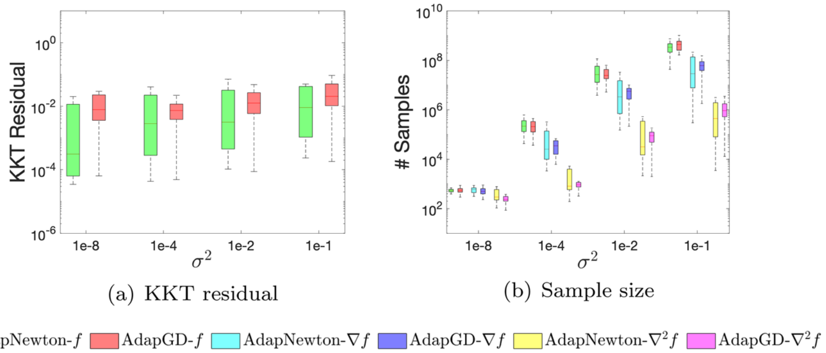
Multiplicative noise We also investigate a multiplicative noise in the experiments. In particular, we employ the default setup $(C, \kappa, \chi_{err}) = (2, 2, 1)$ but replace the noise variance σ^2 by $(1 + \|\boldsymbol{x}_t\|^2)\sigma^2$. Thus, the variance scales linearly with respect to the magnitude of the (primal) iterate. The KKT residual and sample size boxplots are



(a) $C = 2$ (b) $C = 2^3$ (c) $C = 2^6$

(d) $\kappa = 2$ (e) $\kappa = 2^3$ (f) $\kappa = 2^6$

Fig. 7 Feasibility error condition boxplots. Each panel corresponds to a setup of (C, κ). The default values are $C = \kappa = 2$. When we vary one parameter, the other parameter is set as default. Thus, the two figures on the left column are the same



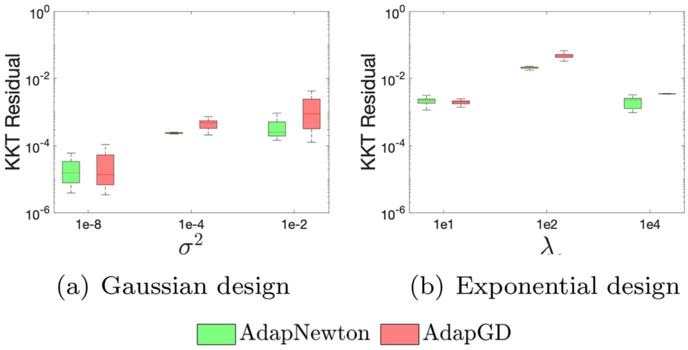
(a) KKT residual (b) Sample size

Fig. 8 Multiplicative noise boxplots. The left figure shows the KKT residual boxplot and the right figure shows the sample size boxplot

shown in Fig. 8. Compared to Figs. 1 and 2, we see that the algorithm achieves comparable results to additive noise. This observation is as expected because, regardless of the noise type, the algorithm enforces the same stochastic conditions on the model estimation accuracy in each iteration, and adaptively selects the batch sizes that are mainly characterized by the current KKT residual.

Logistic regression problem We study an inequality constrained logistic regression problem, where we let

$$F(\boldsymbol{x}; (\xi_{\boldsymbol{a}}, \xi_b)) = \log\{1 + \exp(-\xi_b \cdot \xi_{\boldsymbol{a}}^T \boldsymbol{x})\}, \quad g(\boldsymbol{x}) = C\boldsymbol{x} + \boldsymbol{q}.$$



(a) Gaussian design (b) Exponential design

Fig. 9 KKT residual boxplots. The left figure shows the residual boxplot for the Gaussian design, and the right figure shows the residual boxplot for the exponential design

We set $d = 10$, $r = 5$, and generate each entry of the matrix $C \in \mathbb{R}^{5\times 10}$ and vector $q \in \mathbb{R}^5$ from the standard Gaussian distribution. We let ξ_b be a Rademacher variable (i.e., taking $\{-1, 1\}$ with equal probability), and consider different design distributions for $\xi_{\boldsymbol{a}}$. In particular, we consider both a light tail design $(\xi_{\boldsymbol{a}})_i \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{a}}^2)$ and vary $\sigma_{\boldsymbol{a}}^2 \in \{10^{-8}, 10^{-4}, 10^{-2}\}$, and a heavy tail design $(\xi_{\boldsymbol{a}})_i \sim \text{Exp}(\lambda_{\boldsymbol{a}})$ and vary $\lambda_{\boldsymbol{a}} \in \{10, 10^2, 10^4\}$. Note that $\text{Exp}(\lambda_{\boldsymbol{a}})$ has the variance $1/\lambda_{\boldsymbol{a}}^2$. For each design, we run AdapNewton and AdapGD for 20 times. The default algorithm setup is the same as in Sect. 4.

Figure 9 shows the KKT residual boxplots. From the figure, we observe that AdapNewton performs slightly better than AdapGD. Both methods achieve reasonable performance on all setups of the two designs, although the two methods perform better on the Gaussian design that has a lighter tail than the Exponential design. Overall, the experiments demonstrate the effectiveness of the proposed algorithm.

References

1. Bandeira, A.S., Scheinberg, K., Vicente, L.N.: Convergence of trust-region methods based on probabilistic models. SIAM J. Optim. **24**(3), 1238–1264 (2014). https://doi.org/10.1137/130915984
2. Berahas, A.S., Cao, L., Scheinberg, K.: Global convergence rate analysis of a generic line search algorithm with noise. SIAM J. Optim. **31**(2), 1489–1518 (2021). https://doi.org/10.1137/19m1291832
3. Berahas, A.S., Curtis, F.E., O'Neill, M.J., Robinson, D.P.: A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient Jacobians. arXiv preprint (2021). arXiv:2106.13015
4. Berahas, A.S., Curtis, F.E., Robinson, D., Zhou, B.: Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. SIAM J. Optim. **31**(2), 1352–1379 (2021). https://doi.org/10.1137/20m1354556
5. Berahas, A.S., Bollapragada, R., Zhou, B.: An adaptive sampling sequential quadratic programming method for equality constrained stochastic optimization. arXiv preprint (2022). arXiv:2206.00712
6. Berahas, A.S., Shi, J., Yi, Z., Zhou, B.: Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction. arXiv preprint (2022). arXiv:2204.04161
7. Bertsekas, D.: Constrained Optimization and Lagrange Multiplier Methods. Elsevier, Belmont (1982). https://doi.org/10.1016/c2013-0-10366-2

8. Birge, J.R.: State-of-the-art-survey—stochastic programming: computation and applications. INFORMS J. Comput. **9**(2), 111–133 (1997). https://doi.org/10.1287/ijoc.9.2.111
9. Blanchet, J., Cartis, C., Menickelly, M., Scheinberg, K.: Convergence rate analysis of a stochastic trust-region method via supermartingales. INFORMS J. Optim. **1**(2), 92–119 (2019). https://doi.org/10.1287/ijoo.2019.0016
10. Boggs, P.T., Tolle, J.W.: Sequential quadratic programming. Acta Numer. **4**, 1–51 (1995). https://doi.org/10.1017/s0962492900002518
11. Bollapragada, R., Byrd, R., Nocedal, J.: Adaptive sampling strategies for stochastic optimization. SIAM J. Optim. **28**(4), 3312–3343 (2018). https://doi.org/10.1137/17m1154679
12. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018). https://doi.org/10.1137/16m1080173
13. Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Math. Program. **134**(1), 127–155 (2012). https://doi.org/10.1007/s10107-012-0572-5
14. Cartis, C., Scheinberg, K.: Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. Math. Program. **169**(2), 337–375 (2017). https://doi.org/10.1007/s10107-017-1137-4
15. Chen, C., Tung, F., Vedula, N., Mori, G.: Constraint-aware deep neural network compression. In: Computer Vision—ECCV 2018. Springer, pp. 409–424 (2018). https://doi.org/10.1007/978-3-030-01237-3_25
16. Chen, R., Menickelly, M., Scheinberg, K.: Stochastic optimization using a trust-region method and random models. Math. Program. **169**(2), 447–487 (2017). https://doi.org/10.1007/s10107-017-1141-8
17. Curtis, F.E., O'Neill, M.J., Robinson, D.P.: Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. arXiv preprint (2021). arXiv:2112.14799
18. Curtis, F.E., Robinson, D.P., Zhou, B.: Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. arXiv preprint (2021). arXiv:2107.03512
19. di Serafino, D., Krejić, N., Jerinkić, N.K., Viola, M.: Lsos: Line-search second-order stochastic optimization methods. arXiv preprint (2020). arXiv:2007.15966
20. De, S., Yadav, A., Jacobs, D., Goldstein, T.: Automated inference with adaptive batches. In: Proceedings of Machine Learning Research, PMLR, Fort Lauderdale, FL, USA, vol. 54, pp. 1504–1513 (2017). http://proceedings.mlr.press/v54/de17a.html
21. Fasano, G., Lucidi, S.: A nonmonotone truncated Newton–Krylov method exploiting negative curvature directions, for large scale unconstrained optimization. Optim. Lett. **3**(4), 521–535 (2009). https://doi.org/10.1007/s11590-009-0132-y
22. Friedlander, M.P., Schmidt, M.: Hybrid deterministic-stochastic methods for data fitting. SIAM J. Sci. Comput. **34**(3), A1380–A1405 (2012). https://doi.org/10.1137/110830629
23. Fukuda, E.H., Fukushima, M.: A note on the squared slack variables technique for nonlinear optimization. J. Oper. Res. Soc. Jpn. **60**(3), 262–270 (2017). https://doi.org/10.15807/jorsj.60.262
24. Gallager, R.G.: Stochastic Processes. Cambridge University Press, Cambridge (2013). https://doi.org/10.1017/cbo9781139626514
25. Goh, C.K., Liu, Y., Kong, A.W.K.: A constrained deep neural network for ordinal regression. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2018). https://doi.org/10.1109/cvpr.2018.00093
26. Goldman, A.J., Tucker, A.W.: 4. Theory of linear programming. In: Linear Inequalities and Related Systems. (AM-38). Princeton University Press, pp. 53–98 (1957). https://doi.org/10.1515/9781400881987-005
27. Gould, N.I.M., Orban, D., Toint, P.L.: CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. Comput. Optim. Appl. **60**(3), 545–557 (2014). https://doi.org/10.1007/s10589-014-9687-3
28. Gratton, S., Royer, C.W., Vicente, L.N., Zhang, Z.: Complexity and global rates of trust-region methods based on probabilistic models. IMA J. Numer. Anal. **38**(3), 1579–1597 (2017). https://doi.org/10.1093/imanum/drx043
29. Krejić, N., Krklec, N.: Line search methods with variable sample size for unconstrained optimization. J. Comput. Appl. Math. **245**, 213–231 (2013). https://doi.org/10.1016/j.cam.2012.12.020
30. Liew, C.K.: Inequality constrained least-squares estimation. J. Am. Stat. Assoc. **71**(355), 746–751 (1976). https://doi.org/10.1080/01621459.1976.10481560

31. Liew, C.K.: A two-stage least-squares estimation with inequality restrictions on parameters. Rev. Econ. Stat. **58**(2), 234 (1976). https://doi.org/10.2307/1924031
32. Livieris, I.E., Pintelas, P.: An adaptive nonmonotone active set—weight constrained—neural network training algorithm. Neurocomputing **360**, 294–303 (2019). https://doi.org/10.1016/j.neucom.2019.06.033
33. Livieris, I.E., Pintelas, P.: An improved weight-constrained neural network training algorithm. Neural Comput. Appl. **32**(9), 4177–4185 (2019). https://doi.org/10.1007/s00521-019-04342-2
34. Lucidi, S.: New results on a class of exact augmented Lagrangians. J. Optim. Theory Appl. **58**(2), 259–282 (1988). https://doi.org/10.1007/bf00939685
35. Lucidi, S.: Recursive quadratic programming algorithm that uses an exact augmented Lagrangian function. J. Optim. Theory Appl. **67**(2), 227–245 (1990). https://doi.org/10.1007/bf00940474
36. Lucidi, S.: New results on a continuously differentiable exact penalty function. SIAM J. Optim. **2**(4), 558–574 (1992). https://doi.org/10.1137/0802027
37. Morales, J.L., Nocedal, J., Wu, Y.: A sequential quadratic programming algorithm with an additional equality constrained phase. IMA J. Numer. Anal. **32**(2), 553–579 (2011). https://doi.org/10.1093/imanum/drq037
38. Na, S.: Global convergence of online optimization for nonlinear model predictive control. Adv. Neural Inf. Process. Syst. **34**, 12441–12453 (2021)
39. Na, S., Mahoney, M.W.: Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. arXiv preprint (2022). arXiv:2205.13687
40. Na, S., Anitescu, M., Kolar, M.: An adaptive stochastic sequential quadratic programming with differentiable exact augmented Lagrangians. Math. Program. (2022). https://doi.org/10.1007/s10107-022-01846-z
41. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2006). https://doi.org/10.1007/978-0-387-40065-5
42. Onuk, A.E., Akcakaya, M., Bardhan, J.P., Erdogmus, D., Brooks, D.H., Makowski, L.: Constrained maximum likelihood estimation of relative abundances of protein conformation in a heterogeneous mixture from small angle x-ray scattering intensity measurements. IEEE Trans. Signal Process. **63**(20), 5383–5394 (2015). https://doi.org/10.1109/tsp.2015.2455515
43. Oztoprak, F., Byrd, R., Nocedal, J.: Constrained optimization in the presence of noise. arXiv preprint (2021). arXiv:2110.04355
44. Paquette, C., Scheinberg, K.: A stochastic line search method with expected complexity analysis. SIAM J. Optim. **30**(1), 349–376 (2020). https://doi.org/10.1137/18m1216250
45. Phillips, R.F.: A constrained maximum-likelihood approach to estimating switching regressions. J. Econom. **48**(1–2), 241–262 (1991). https://doi.org/10.1016/0304-4076(91)90040-k
46. Pillo, G.D., Grippo, L.: A new class of augmented Lagrangians in nonlinear programming. SIAM J. Control. Optim. **17**(5), 618–628 (1979). https://doi.org/10.1137/0317044
47. Pillo, G.D., Grippo, L.: A new augmented Lagrangian function for inequality constraints in nonlinear programming problems. J. Optim. Theory Appl. **36**(4), 495–519 (1982). https://doi.org/10.1007/bf00940544
48. Pillo, G.D., Grippo, L.: A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints. SIAM J. Control. Optim. **23**(1), 72–84 (1985). https://doi.org/10.1137/0323007
49. Pillo, G.D., Grippo, L.: An exact penalty function method with global convergence properties for nonlinear programming problems. Math. Program. **36**(1), 1–18 (1986). https://doi.org/10.1007/bf02591986
50. Pillo, G.D., Lucidi, S.: An augmented Lagrangian function with improved exactness properties. SIAM J. Optim. **12**(2), 376–406 (2002). https://doi.org/10.1137/s1052623497321894
51. Pillo, G.D., Grippo, L., Lampariello, F.: A method for solving equality constrained optimization problems by unconstrained minimization. In: Optimization Techniques, Springer-Verlag, Lecture Notes in Control and Information Science, vol. 23, pp. 96–105 (1980). https://doi.org/10.1007/bfb0006592
52. Pillo, G.D., Lucidi, S., Palagi, L.: Convergence to second-order stationary points of a primal-dual algorithm model for nonlinear programming. Math. Oper. Res. **30**(4), 897–915 (2005). https://doi.org/10.1287/moor.1050.0150

53. Pillo, G.D., Liuzzi, G., Lucidi, S., Palagi, L.: A truncated Newton method in an augmented Lagrangian framework for nonlinear programming. Comput. Optim. Appl. **45**(2), 311–352 (2008). https://doi.org/10.1007/s10589-008-9216-3
54. Pillo, G.D., Liuzzi, G.S.L.: A primal-dual algorithm for nonlinear programming exploiting negative curvature directions. Numer. Algebra Control Optim. **1**(3), 509–528 (2011). https://doi.org/10.3934/naco.2011.1.509
55. Pillo, G.D., Liuzzi, G., Lucidi, S.: An exact penalty-Lagrangian approach for large-scale nonlinear programming. Optimization **60**(1–2), 223–252 (2011). https://doi.org/10.1080/02331934.2010.505964
56. Silvapulle, S.: Constrained Statistical Inference, vol. 912. Wiley, New York (2004)
57. Sun, S., Nocedal, J.: A trust region method for the optimization of noisy functions. arXiv preprint (2022). arXiv:2201.00973
58. Tropp, J.A.: An introduction to matrix concentration inequalities. Found. Trends® Mach. Learn. **8**(1–2), 1–230 (2015). https://doi.org/10.1561/2200000048
59. Xu, M., Ye, J.J., Zhang, L.: Smoothing augmented Lagrangian method for nonsmooth constrained optimization problems. J. Glob. Optim. **62**(4), 675–694 (2014). https://doi.org/10.1007/s10898-014-0242-7
60. Zavala, V.M., Anitescu, M.: Scalable nonlinear programming via exact differentiable penalty functions and trust-region Newton methods. SIAM J. Optim. **24**(1), 528–558 (2014). https://doi.org/10.1137/120888181