Proceedings of Machine Learning Research vol 178:1–110, 2022 35th Annual Conference on Learning Theory

Big-Step-Little-Step: Efficient Gradient Methods for Objectives with Multiple Scales

Jonathan Kelner KELNER@MIT.EDU
MIT[*]

Annie Marsden MARSDEN@STANFORD.EDU
Stanford University

Vatsal Sharan VSHARAN@USC.EDU
USC[†]

Aaron Sidford SIDFORD@STANFORD.EDU
Stanford University

Gregory Valiant VALIANT@STANFORD.EDU
Stanford University

Honglin Yuan HONGL.YUAN@GMAIL.COM
Stanford University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We provide new gradient-based methods for efficiently solving a broad class of ill-conditioned optimization problems. We consider the problem of minimizing a function $f : \mathbb{R}^d \to \mathbb{R}$ which is implicitly decomposable as the sum of m unknown non-interacting smooth, strongly convex functions and provide a method which solves this problem with a number of gradient evaluations that scales (up to logarithmic factors) as the product of the square-root of the condition numbers of the components. This complexity bound (which we prove is nearly optimal) can improve almost exponentially on that of accelerated gradient methods, which grow as the square root of the condition number of f. Additionally, we provide efficient methods for solving stochastic, quadratic variants of this multiscale optimization problem. Rather than learn the decomposition of f (which would be prohibitively expensive), our methods apply a clean recursive "Big-Step-Little-Step" interleaving of standard methods. The resulting algorithms use $\tilde{\mathcal{O}}(dm)$ space, are numerically stable, and open the door to a more fine-grained understanding of the complexity of convex optimization beyond condition number.

Keywords: Convex optimization, first-order methods, condition number, memory constrained optimization.

[*] Authors are listed in alphabetical order.
[†] Part of the work was done while the author was at MIT.

1. Introduction

Smooth, strongly-convex function minimization is a fundamental and canonical problem in optimization theory and machine learning. Given an L-smooth, μ-strongly convex $f : \mathbb{R}^d \to \mathbb{R}$ it is well known that gradient descent and accelerated gradient descent minimize f with $\tilde{\mathcal{O}}(\kappa)$ and $\tilde{\mathcal{O}}(\sqrt{\kappa})$ gradient queries respectively for $\kappa \coloneqq L/\mu$.[1] Further, any first-order method, i.e. one restricted to accessing f through an oracle which returns the value and gradient of f at a queried point, must make $\Omega(\sqrt{\kappa})$ queries (Nemirovski and Yudin, 1983) in the (dimension-independent) worst case (even if randomized (Woodworth and Srebro, 2017)). Consequently κ, the *condition number*, nearly captures the worst-case complexity of the problem.

In this paper, we seek to move beyond this traditional measure of problem complexity and obtain a more fine-grained understanding of the complexity of smooth, strongly convex function minimization. In the special case of quadratic function minimization where $\nabla^2 f$ has a small number of distinct eigenvalue clusters, it has long been known that methods like conjugate gradient can efficiently solve the problem with far fewer gradient queries than would be indicated by the condition number of the problem (Trefethen and Bau, 1997). This fact has been leveraged in a variety of contexts for improved methods (Polyak, 1969; Nocedal, 1996; Saad, 2003; Nocedal and Wright, 2006b).

The central question we ask in this paper is *whether there is an analog of this phenomenon for non-quadratic and stochastic optimization problems*. Although methods like non-linear conjugate gradient (Fletcher and Reeves, 1964; Hager and Zhang, 2006) and limited-memory Quasi-Newton methods (Nocedal, 1980; Liu and Nocedal, 1989) are prevalent and effective in practice, we currently lack a complete theoretical understanding of when they are (provably) effective. In this work, we answer this question in the affirmative and give efficient first-order methods for solving natural classes of non-quadratic and stochastic multi-scale optimization problems.

We focus on the problem of minimizing a function f which is decomposable as the sum of m non-interacting smooth, strongly convex functions f_i each with condition number κ_i. When each κ_i is small and the smoothness of each component, L_i, is similar, the overall function is well-conditioned and can be solved efficiently. However, when the components are at different scales, i.e. the L_i vary, the overall function can be ill-conditioned. We provide methods that depend only poly-logarithmically on the overall condition number and polynomially on the κ_i, improving almost exponentially on the complexity of AGD in certain cases. We complement this result with novel and nearly matching lower bounds which show that our methods are close to optimal for the class of first-order methods.

The motivation for considering the specific setting of a sum of non-interacting functions that are at different scales is largely theoretical. This model serves as a natural starting place when considering what types of structure can be algorithmically leveraged to go beyond guarantees in terms of the (global) condition number. Still, the setting we focus on is not too removed from the types of structure one might encounter in practice. Indeed, many optimization problems possess structure at unknown and widely varying scales, and optimization approaches designed to gracefully handle such scaling, such as `AdaGrad` (Duchi et al., 2011), have proved to be extremely useful in practice. Our assumption that the components of the objective function are completely non-interacting may be unrealistic, but we are hopeful that our techniques might extend to more general classes of structured, poorly conditioned, optimization settings.

1. $\tilde{\mathcal{O}}(\cdot)$ hides factors poly-logarithmic in the function error of the initial point, desired accuracy, and condition numbers.

1.1. Setup and overview

Definition 1 *The **multiscale (convex) optimization problem** asks to approximately solve the problem that can be implicitly decomposed as follows:*

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \sum_{i\in[m]} f_i(\mathbf{P}_i\mathbf{x}) \quad \textit{where} \quad \mathbf{P}_i \in \mathbb{R}^{d_i\times d} \quad \textit{with} \quad \mathbf{P}_i\mathbf{P}_j^\top = \begin{cases} \mathbf{I}_{d_i\times d_i} & \textit{if } i=j, \\ \mathbf{0}_{d_i\times d_j} & \textit{otherwise,} \end{cases}$$

*the projections $\mathbf{P}_i$ and functions f_i are **unknown**, each $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ has condition number $\kappa_i := L_i/\mu_i$ where f_i is L_i-smooth and μ_i-strongly convex with $L_i < \mu_{i+1}$ for all $i < m$.*[2]

We focus on optimizing such objectives given *only* a gradient oracle for f. We also note that though the above problem (and our yet to be introduced algorithms) are well-defined for any $m \in [d]$, we will treat m as a constant when stating the asymptotic bounds of our algorithms for simplicity.

For intuition, one simple case of Theorem 1 is the quadratic minimization problem $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{b}^\top\mathbf{x}$ where $\mathbf{A}$ is positive-definite and its eigenvalues are all located in $\bigcup_{i\in[m]}[\mu_i, L_i]$. To see this, note that the spectral decomposition of $\mathbf{A}$ can be written as $\mathbf{A} = \sum_{i\in[m]} \mathbf{P}_i^\top\mathbf{\Lambda}_i\mathbf{P}_i$ such that $\mathbf{\Lambda}_i$ is a diagonal matrix with diagonal entries within $[\mu_i, L_i]$, and that the matrices $\{\mathbf{P}_i\}_{i=1}^m$ are pairwise orthogonal. By setting $f_i(\mathbf{y}) := \frac{1}{2}\mathbf{y}^\top\mathbf{\Lambda}_i\mathbf{y} - \mathbf{b}^\top\mathbf{P}_i^\top\mathbf{y}$ one can cast the problem in the form of Theorem 1.

Since any objective f satisfying Theorem 1 must be μ_1-strongly-convex and L_m-smooth, one may directly apply gradient descent or accelerated gradient descent to minimize f within $\tilde{\mathcal{O}}(\kappa_{\text{glob}})$ or $\tilde{\mathcal{O}}(\sqrt{\kappa_{\text{glob}}})$ gradient queries, where κ_{glob} is the global condition number L_m/μ_1. However, as we show in Appendix F.4, gradient descent with constant step-size or line-search does not take full advantage of the additional structure beyond the μ_1-strong-convexity and L_m-smoothness from Theorem 1. In this work, we aim to leverage the structure of Theorem 1 to develop faster algorithms.

Our first contribution is to develop a new algorithm "Big-Step-Little-Step" or `BSLS` (Algorithm 1) which takes advantage of the structure of Theorem 1 and as a result solves the problem within $\tilde{\mathcal{O}}(\prod_{i\in[m]}\kappa_i)$ gradient queries of f (see Theorem 6). Since $\mu_1 \le L_1 < \mu_2 \le L_2 < \cdots < \mu_m \le L_m$ by Theorem 1, it is always the case that $\prod_{i\in[m]}\kappa_i \le \kappa_{\text{glob}}$. Therefore, `BSLS` asymptotically outperforms gradient descent (which has complexity $\tilde{\mathcal{O}}(\kappa_{\text{glob}})$). Moreover, if the clusters $[\mu_i, L_i]$ are well-separated, i.e. $\prod_{i\in[m]}\kappa_i \ll \kappa_{\text{glob}}$, the complexity of `BSLS` can *significantly* outperform accelerated gradient descent. Indeed, in the case where m and each κ_i are constant, the asymptotic performance (with respect to κ_{glob}) of `BSLS` is almost an exponential improvement on the performance of accelerated gradient method; see Fig. 1 for a small experimental comparison in the quadratic minimization setting . We also show in Theorem 14 that `BSLS` is numerically stable under finite-precision arithmetic.

The `BSLS` algorithm consists of a natural interleaving of steps at different sizes. Intuitively, `BSLS` alternates between taking a bigger step-size to make progress on an objective f_i which has a smaller scale (i.e. a small value of L_i), followed by a sequence of smaller steps to fix the errors caused due to this large step in the objectives f_j which have a larger scale (i.e. all $j > i$). The entire framework is recursive – the sequence of smaller steps for $j > i$ are themselves defined recursively, see Section 3 for further intuition.

2. This is without loss of generality (up to logarithmic factors in our claimed bounds) by re-defining any f_i, f_j pair with $[\mu_i, L_i] \cap [\mu_j, L_j] \neq \emptyset$ as a single f_i, sorting the L_i, and noting that $\mu_i \le L_i$ (by known properties of smoothness and convexity).

Next, we develop an accelerated version of BSLS, namely AcBSLS (Algorithm 2), that solves the problem within $\tilde{\mathcal{O}}(\prod_{i\in[m]}\sqrt{\kappa_i})$ gradient queries of f (see Theorem 15). Again, as $\prod_{i\in[m]}\kappa_i \leq \kappa_{\text{glob}}$, AcBSLS complexity is never worse than the $\tilde{O}(\sqrt{\kappa_{\text{glob}}})$ complexity of AGD and can significantly improve when the clusters are well separated. We also show in Theorem 21 that AcBSLS is numerically stable under finite-precision arithmetic.

We conclude the study of the multiscale optimization problem (Theorem 1) by developing a lower bound of $\tilde{\Omega}(\prod_{i\in[m]}\sqrt{\kappa_i})$ across first-order deterministic methods (see Theorem 10). This shows that AcBSLS is asymptotically optimal up to poly-logarithmic factors. Our proof framework consists of 1) a novel reduction of a first-order lower bound to discrete ℓ_2 polynomial approximations on multiple intervals and a further reduction to a uniform approximation, 2) a standard reduction to Green's function based on potential theory Driscoll et al. (1998), and 3) a novel estimate of Green's function associated with multiple intervals.

We summarize the main results in the following theorem.

Theorem 2 (Informal version of Theorems 6, 10, 14, 15 and 21) BSLS *solves the multiscale optimization problem to ϵ-optimality with $\tilde{\mathcal{O}}(\prod_{i\in[m]}\kappa_i)$ gradient evaluations. The accelerated version* AcBSLS *solves to ϵ-optimality with $\tilde{\mathcal{O}}(\prod_{i\in[m]}\sqrt{\kappa_i})$ gradient evaluations. Both* BSLS *and* AcBSLS *only require logarithmic bits of precision and use $\tilde{\mathcal{O}}(d)$ and $\tilde{\mathcal{O}}(md)$ space, respectively. Further,* AcBSLS *is worst-case optimal across first-order deterministic algorithms up to poly-logarithmic factors.*

In the case where the objective in Theorem 1 is quadratic, we show that the conjugate gradient method (CG) (Hestenes and Stiefel, 1952) also solves the problem with $\tilde{\mathcal{O}}(\prod_{i\in[m]}\sqrt{\kappa_i})$ gradient queries and is numerically stable (see Appendix F.6). AcBSLS matches this performance in the quadratic setting and further extends the guarantee to a much broader class of non-quadratic problems. We discuss this implication further in Section 4.

Remark 3 *We provide several remarks on the setup of the multiscale optimization and our results.*

(a) Theorem 1 does not assume knowledge of the decomposition. If in addition one assumes that the decompositions (i.e., f_i and $\mathbf{P}_i$) are individually accessible, the problem can be solved much more efficiently (and more trivially), using $\tilde{\mathcal{O}}(\sum_{i\in[m]}\sqrt{\kappa_i})$ sub-objective gradient queries (see Appendix F.2), in contrast to $\tilde{\mathcal{O}}(\prod_{i\in[m]}\sqrt{\kappa_i})$.

(b) Given the existence of efficient algorithms if the decomposition is known, one may be tempted to first recover the decomposition (f_i and $\mathbf{P}_i$) before solving. However, we show in Appendix F.3 that recovering the $\mathbf{P}_i$ with access to a gradient oracle is costly, in that it takes $\Omega(d)$ queries in the worst case.

(c) Theorem 1 assumes orthogonality conditions on $\{\mathbf{P}_i\}_{i\in[m]}$. We remark that some amount of disjointness is critical to obtaining these upper bounds. In Appendix F.5 we show a $\Omega(\sqrt{\kappa_{\text{glob}}})$ lower bound on the complexity of the problem without such an orthogonality assumption.

(d) Our algorithms do not *necessarily require the knowledge of all the μ_i and L_i's. In fact, due to a simple grid-search, our theorems only require that m, μ_1, L_m and $\prod_{i\in[m]}\kappa_i$ are known to obtain the claimed asymptotic query complexity (see Appendix F.1).*

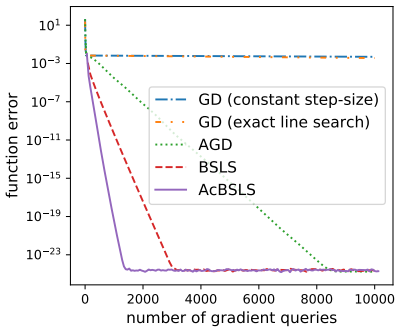


Figure 1: **A numerical example demonstrating the efficiency of `BSLS` and `AcBSLS`.** Our objective is $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A}$ has eigenvalues in $[0.0001, 0.0002] \cup [1, 10]$. This objective satisfies Theorem 1 with $\kappa_1 = 2$, $\kappa_2 = 10$ and $\kappa_{\text{glob}} = 10^5$. We compare `BSLS` and `AcBSLS` with Gradient Descent (GD) with constant step-size, Gradient Descent with exact line search, and Accelerated Gradient Descent (AGD). Observe that `AcBSLS` and `BSLS` clearly outperform the other algorithms.

Next, we consider the following stochastic, quadratic variant of the multiscale optimization.

Definition 4 (Stochastic Quadratic Multiscale Optimization Problem) *The **stochastic quadratic multiscale optimization problem** asks to approximately solve the following problem* $\min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}} \left[\frac{1}{2}(\mathbf{a}^\top \mathbf{x} - b)^2\right]$, *where* $b = \mathbf{a}^\top \mathbf{x}^\star$ *for some fixed, unknown* $\mathbf{x}^\star$ *and the eigenvalues of the covariance matrix* $\mathbb{E}_{\mathcal{D}}[\mathbf{a}\mathbf{a}^\top]$ *can be partitioned into* m *"bands" such that for* $i = 1, \ldots, m$ *and* $j = 1, \ldots, d_i$, *each eigenvalue* λ_{i_j} *satisfies* $\lambda_{i_j} \in [\mu_i, L_i]$ *with* $L_i < \mu_{i+1}\ \forall i < m$.

This objective is a special case of the multiscale optimization problem in Definition 1 where each f_i is a quadratic function of $\mathbf{x}$ (we make the connection explicit in Section E). Therefore, in the non-stochastic case where we have access to noiseless gradients of $f(\mathbf{x})$, our guarantees for `AcBSLS` imply that the problem can be solved with $\tilde{\mathcal{O}}(\prod_{i\in[m]} \sqrt{\kappa_i})$ gradient evaluations. In the next theorem we show that `StochBSLS` provides similar performance guarantees which are robust to stochasticity.

Theorem 5 (Informal version of Theorem 11) *Under certain second-order independence and fourth moment assumptions on* $\mathcal{D}$, `StochBSLS` *solves the stochastic quadratic multiscale optimization problem in expectation with* ϵ*-optimality using* $d \cdot \prod_{i\in[m]} \tilde{\mathcal{O}}(\kappa_i^2)$ *stochastic gradient queries and* $\tilde{\mathcal{O}}(d)$ *space.*

2. Prior work

There is a vast literature on designing and analyzing first-order methods. Here, we survey several lines of research that are most closely related to our contributions.

Complexity measures for first-order methods. There are many results which consider notions other than smoothness and strong convexity for first-order methods. Some examples of this is work on star-convexity (Guminov and Gasnikov, 2017; Nesterov et al., 2018; Hinder et al., 2020), quasi-strong convexity (Necoara et al., 2019), semi-convexity (Van Ngai and Penot, 2007), the quadratic growth condition (Pang, 1997; Anitescu, 2000), the error bound property (Luo and Tseng, 1993; Fabian et al., 2010), restricted strong convexity (Zhang and Yin, 2013; Zhang and Cheng, 2015) and Hölder continuity (Zhang and Yin, 2013; Devolder et al., 2014; Yashtini, 2016; Grimmer, 2019). However, we are unaware of notions of fine-grained condition numbers for non-linear or stochastic problems appearing previously in the literature.

Structured linear systems. As mentioned before, the conjugate gradient method also solves the quadratic, noiseless version of the multiscale optimization problem. We refer the reader to some of the surveys for more discussion, including various preconditioning procedures (Greenbaum, 1997; Saad, 2003; Nocedal and Wright, 2006b). There is also work on improving the condition number dependence of first-order methods to an average condition number (ratio of the average of the eigenvalues of the Hessian and the smallest eigenvalue), which can be smaller than the condition number (Johnson and Zhang, 2013; Shalev-Shwartz and Zhang, 2013; Musco et al., 2018b). There is also work on preconditioning the matrix by deflating large eigenvalues and hence reducing the average condition number in cases with a few very large eigenvalues (Gonen et al., 2016; Musco et al., 2018b).

Nonlinear CG. Various nonlinear versions of CG have also been proposed such as Fletcher-Reeves (FR) method (Fletcher and Reeves, 1964) and Polak-Ribière (PR) method (Polak and Ribiere, 1969). These methods are effective in practice and have been widely applied by the numerical optimization community (Nocedal and Wright, 2006a; Hager and Zhang, 2006; Dai, 2011). However, for nonlinear CG, there is still a substantial gap between its practical performance and our theoretical understanding. On the negative side, it is known from Chap. 7 of Nemirovski and Yudin (1983) that the FR and PR method do not match the accelerated GD rate $\tilde{\mathcal{O}}(\sqrt{\kappa_{\text{glob}}})$.

Adaptive step sizes for gradient descent. Similar to `BSLS` and its variants, previous works have also explored the use of an adaptive step size sequence. For example, the Barzilai-Borwein method (Barzilai and Borwein, 1988) is a chaotic method which makes progress by making large step sizes and then correcting errors. Malitsky and Mishchenko (2019) provide a gradient descent method with step sizes that adapt to local geometry, without the use of a line search. Agarwal et al. (2021) show that for quadratic objectives, vanilla gradient descent with interlaced small and large step sizes achieves acceleration without the use of momentum. However, these papers do not consider our multiscale optimization problem setup and as far as we are aware these algorithms cannot recover the same complexity results. We also note that cyclical learning rate schedules such as those employed by `BSLS` have also been applied in deep learning (Loshchilov and Hutter, 2016; Smith, 2017; Fu et al., 2019) and it could be interesting to extend the present work to provide a theoretical grounding to these methods.

Leveraging second-order structure via first-order methods. There is a large body of work on methods to approximate second-order information including quasi-Newton methods such as DFP (Davidon, 1991), BFGS (Broyden, 1970), L-BFGS (Nocedal, 1980; Liu and Nocedal, 1989), methods based on subsampling and sketching the Hessian (Pilanci and Wainwright, 2017; Xu et al., 2020; Roosta-Khorasani and Mahoney, 2019), methods which learn diagonal preconditioners such as `AdaGrad` (Duchi et al., 2011) and `Adam` (Kingma and Ba, 2014), stochastic second order methods (Agarwal et al., 2017) and Newton-CG (Royer et al., 2019; Curtis et al., 2021). Carmon and Duchi (2018) also provide accelerated methods that only use gradient and Hessian-vector queries and improve on the complexity of gradient descent for finding stationary points for certain non-convex problems. However, it is not known whether any of these algorithms achieves a worst-case complexity that does not depend polynomially on the overall condition number.

Stochastic methods. Stochastic gradient methods are the workhorse for large scale optimization and machine learning problems (Bottou and Bousquet, 2008) and there is extensive work on stochastic gradient algorithms for solving linear systems, including randomized Kaczmarz (Strohmer and

Vershynin, 2006; Needell et al., 2016), variance reduction techniques (Johnson and Zhang, 2013; Zhang et al., 2013; Schmidt et al., 2017) and accelerated methods (Liu and Wright, 2016; Allen-Zhu, 2017; Jain et al., 2018). However, the complexity of all these methods depends polynomially on some measure of eigenvalue range or conditioning of the underlying matrix.

Lower bounds. Starting with the seminal work of Nemirovski and Yudin (1983), there is a rich body of work on lower bounds for first-order methods. More recently, several works have extended these results to randomized algorithms (Woodworth and Srebro, 2017; Simchowitz, 2018; Braverman et al., 2020; Woodworth, 2021), and we use these results to show necessity of the orthogonality assumption in the multiscale optimization problem. To show query-complexity lower bounds for first-order methods for the multiscale optimization problem, we show a reduction from a first-order lower bound to a polynomial approximation problem on multiple intervals. There is extensive literature on polynomial approximations we leverage here, especially the work of Widom (1969) (more references appear in Section D). We also note that there is a long history of relating the convergence rates of optimization algorithms to polynomial approximation problems, including the work of Greenbaum (1989a); Musco et al. (2018a) on convergence of Lanczos and CG.

3. Approach and results

In this section we give an overview of our algorithms and results.

3.1. Big-Step-Little-Step Algorithm (BSLS)

We begin by introducing our main algorithm "Big-Step-Little-Step" (BSLS) for the multiscale optimization problem (Theorem 1). As the name suggests, BSLS adopts the idea of running a series of gradient descent steps with alternating step-sizes ranging from L_m^{-1} to L_1^{-1}. To see the rationale behind alternating step-sizes consider the simple case of $m = 2$ sub-objectives. If we were to run one step of GD on f, the smoother sub-objective f_1 would favor a "big step" of size L_1^{-1}, while the less smooth f_2 would favor a "little step" of size L_2^{-1}. In fact, the "big-step" will decrease f_1 considerably (by a factor of $1 - \kappa_1^{-1}$), but it may also increase f_2 (by no more than a factor of κ_{glob}^2). On the other hand, the "little-step" will decrease f_2 considerably (by a factor of $1 - \kappa_2^{-1}$), but will not decrease f_1 substantially (though it will not increase f_1 either). Thus, in order to make progress on both f_1 and f_2 efficiently, one could run one big step, followed by multiple little steps to fix the increase in f_2 from the previous large step-size. The BSLS algorithm is a careful interleaving of these big and little steps. This intuition extends readily to the case of $m > 2$ sub-objectives with a recursive framework (see Algorithm 1). We begin by executing $\texttt{BSLS}_1(\mathbf{x}^{(0)})$ on the initialization $\mathbf{x}^{(0)}$. We explain the recursive procedure via an illustrative example in Fig. 2.

The following theorem characterizes the convergence rate of BSLS. The proof appears in Appendix B.1.

Theorem 6 *In the multiscale optimization (Def. 1), for any $\mathbf{x}^{(0)}$ and $\epsilon > 0$, $\texttt{BSLS}_1(\mathbf{x}^{(0)})$ returns an ϵ-optimal solution with $\mathcal{O}\left(\left(\prod_{i\in[m]} \kappa_i\right) \cdot \left(\log^{m-1} \kappa_{\text{glob}}\right) \cdot \log\left(\frac{f(\mathbf{x}^{(0)})-f^\star}{\epsilon}\right)\right)$ gradient queries when $\{(\mu_i, L_i), i \in [m]\}$ are known. Moreover in the case where $\{(\mu_i, L_i), i \in [m]\}$ are unknown and only m, μ_1, L_m and $\pi_\kappa = \prod_{i\in[m]} \kappa_i$ are known, we can achieve the same asymptotic sample complexity (up to constant factors suppressed in the $\mathcal{O}(\cdot)$).*

Algorithm 1 Big-Step-Little-Step Algorithm

Procedure $\texttt{GD}(\mathbf{x}; L)$

1: **return** $\mathbf{x} - \frac{1}{L} \cdot \nabla f(\mathbf{x})$

Procedure $\texttt{BSLS}_i\ (\mathbf{x}^{(0)})$

1: $T_i \leftarrow \begin{cases} \left\lceil \kappa_1 \log\left(\frac{f(\mathbf{x}^{(0)}) - f^\star}{\epsilon}\right)\right\rceil & \text{if } i = 1, \\ \lceil \kappa_i \cdot (2 \log \kappa_{\text{glob}} + 1)\rceil & \text{otherwise.} \end{cases}$

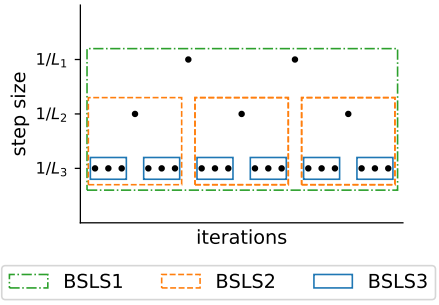
2: **for** $t = 0, 1, \ldots, T_i - 1$ **do**

3: $\quad \tilde{\mathbf{x}}^{(t)} \leftarrow \texttt{BSLS}_{i+1}(\mathbf{x}^{(t)})$ if $i < m$, or $\mathbf{x}^{(t)}$ otherwise.

4: $\quad \mathbf{x}^{(t+1)} \leftarrow \texttt{GD}(\tilde{\mathbf{x}}^{(t)}; L_i)$

5: **return** $\texttt{BSLS}_{i+1}(\mathbf{x}^{(T_i)})$ if $i < m$, or $\mathbf{x}^{(T_i)}$ otherwise.

Figure 2: **An illustrative example of** BSLS **(Algorithm 1) with** $m = 3$**,** $T_1 = 2$**,** $T_2 = 1$**,** $T_3 = 3$**.** The algorithm starts by executing $\texttt{BSLS}_1(\mathbf{x}^{(0)})$ at initialization $\mathbf{x}^{(0)}$. $\texttt{BSLS}_1$ will invoke $\texttt{BSLS}_2$ for $T_1 + 1 = 3$ times, with two "big" steps (GD of step-size $1/L_1$) in between. Within each invocation of $\texttt{BSLS}_2$, it will invoke $\texttt{BSLS}_3$ for $T_2 + 1 = 2$ times, with one "medium" step (GD of step-size $1/L_2$) in between. $\texttt{BSLS}_3$ only consists of $T_3 = 3$ little steps (GD of step-size $1/L_3$) since $m = 3$ is the final layer. Therefore, $\texttt{BSLS}_1$ effectively consists of 3 types of gradient steps structured in an interlacing order.



Given the rationale behind BSLS, it is natural to ask whether such a careful step-size sequence is necessary to obtain fast convergence. Perhaps by simply performing line-searches in the direction of the gradient we can obtain a method which automatically finds the appropriate scale to make progress? As we also mentioned earlier in Section 1.1, Appendix F.4 shows that this is not the case; indeed, we show instances where gradient descent with exact line search or constant step-sizes require $\Omega(\kappa_{\text{glob}})$ gradient evaluations to solve the problem while BSLS only requires $O(\log(\kappa_{\text{glob}}))$ gradient evaluations. This illustrates that it can be difficult to directly guess the right step-size and suggests the need for a step-size schedule such as that employed by BSLS.

3.2. Stability of BSLS and why interlacing order matters

For methods like conjugate gradient, there are known gaps between the best-known theoretical performance with infinite precision and finite-precision arithmetic (Paige, 1971; Greenbaum, 1989a), and there are related robustness issues in the face of statistical errors (Polyak, 1987). Consequently, when designing methods for the multiscale problems we consider, care is needed to ensure methods perform efficiently even without infinite precision arithmetic. Here, we discuss the stability properties of BSLS. We first note that *under exact arithmetic* any reordering of the GD steps in $\texttt{BSLS}_1$ attains the same convergence rate:

Proposition 7 *In the multiscale optimization (Theorem 1), assume all operations are performed under exact arithmetic. Then any reshuffling of* GD *steps in* $\texttt{BSLS}_1$ *(Algorithm 1) attains* ϵ*-optimality.*

In contrast, we show that *under finite-precision*, the interlacing order defined by recursive BSLS (Algorithm 1) is essential to guarantee the stability. Specifically, we show that our recursive $\texttt{BSLS}_1$ only requires roughly *logarithmic bits of precision* (per floating-point number) to match the rate of convergence achieved under exact arithmetic, in contrast to potentially (at least) polynomial bits of precision for problematic orderings.

To understand why order matters in finite-precision, let us again consider simply $m = 2$ sub-objectives. Theorem 7 suggests a total of $\tilde{\Theta}(\kappa_1)$ big steps and $\tilde{\Theta}(\kappa_1\kappa_2)$ little steps are needed to attain ϵ-optimality under exact arithmetic. Consider a problematic ordering: begin with $\tilde{\Theta}(\kappa_1)$ big steps altogether and end with little steps altogether. With this ordering, the initial $\tilde{\Theta}(\kappa_1)$ big steps will amplify the error of f_2 by $\kappa_{\text{glob}}^{\tilde{\Theta}(\kappa_1)}$. Under finite-precision, one needs $\tilde{\Theta}(\kappa_1)$ bits of precision to keep track of this growth, which is polynomial in the condition numbers. The same arguments apply if one runs all the $\tilde{\Theta}(\kappa_1\kappa_2)$ little steps first — polynomial bits of precision are needed to secure the progress made by the little steps in f_1 before the big steps bring the error up. In contrast, our recursive BSLS (Algorithm 1) overcomes this issue because the progress of all sub-objectives is balanced thanks to the interlacing step-sizes. We demonstrate this phenomenon in Fig. 3 with a numerical example. We formally prove the stability of (recursive) BSLS in Appendix B.2.

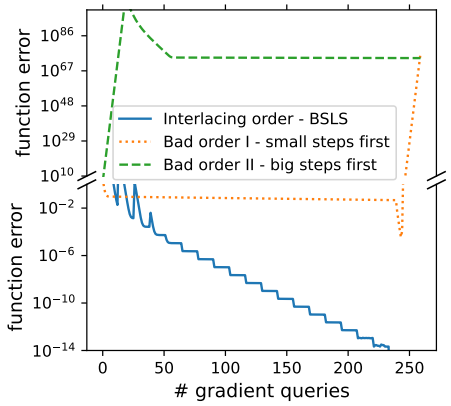


Figure 3: **Importance of interlacing order in BSLS under finite-precision arithmetic.** Our objective is $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{b}^\top\mathbf{x}$, where $\mathbf{A}$ has eigenvalues in $[0.001, 0.002] \cup [0.5, 1]$. We consider three different step-size orderings, each running 20 big steps and 240 little steps in total. The first (solid) line runs $\texttt{BSLS}_1$ (Algorithm 1), namely every big step is followed by 12 little steps. The second (dotted) line runs all little steps before the big steps. The third (dashed) line runs all big steps before the little steps. Observe that only the principled BSLS converges under finite arithmetic (double-precision floating point format).

3.3. Accelerated Big-Step-Little-Step algorithm (AcBSLS)

We provide an accelerated version of BSLS algorithm, namely Accelerated BSLS (AcBSLS), which with $\mathcal{O}\left(\left(\prod_{i\in[m]}\sqrt{\kappa_i}\right)\cdot\log^{m-1}\kappa_{\text{glob}}\cdot\log\left(\frac{f(\mathbf{x}^{(0)})-f^\star}{\epsilon}\right)\right)$ gradient queries solves the multiscale optimization problem (Theorem 1) up-to ϵ-optimality. As we will see below, AcBSLS is optimal across first-order deterministic methods up-to poly-logarithmic factors.

AcBSLS shares the similar motivations of adopting alternating step-sizes as in BSLS. Instead of running GD, AcBSLS runs Accelerated Gradient Descent (AGD) with various "step-sizes". Formally, we use $\texttt{AGD}(\mathbf{x}, \mathbf{v}; L, \mu)$ to denote one-step of AGD with smooth estimate L and convexity estimate μ (see the first block of Algorithm 2 for definitions). The AcBSLS algorithm (the second block of Algorithm 2) then follows a similar recursive structure as in BSLS defined in Algorithm 1.

The major difference between AcBSLS and the (un-accelerated) BSLS lies in the difficulties of fixing the larger (less-smooth) sub-objectives after executing the big step-sizes. To understand this challenge, let us consider the simple case with only $m = 2$ sub-objectives. Recall that in BSLS, after executing one big GD step, we run T_2 little GD steps to fix the surge in f_2. This is backed by the

fact that little GD steps will not increase the smaller (smoother) sub-objective f_1, as suggested by Theorem 12. Unfortunately, this relation does *not* trivially extend to the accelerated setting, because $\texttt{AGD}(\mathbf{x}, \mathbf{v}; L_2, \mu_2)$ may *not* keep the joint progress of $\mathbf{x}, \mathbf{v}$ in f_1. Consequently, we adopt a more sophisticated branching strategy that fixes $\mathbf{x}$ and $\mathbf{v}$ separately, see Line 5 of $\texttt{AcBSLS}_i$ in Algorithm 2. We refer readers to Appendix C.3 for a numerical example on the non-convergence of naive AcBSLS without branching.

Algorithm 2 Accelerated Big-Step-Little-Step Algorithm

Procedure $\texttt{AGD}(\mathbf{x}, \mathbf{v}; L, \mu)$

$\kappa \leftarrow L/\mu$; $\alpha \leftarrow \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$; $\beta \leftarrow 1 - \frac{1}{\sqrt{\kappa}}$

$\mathbf{y} \leftarrow \alpha\mathbf{x} + (1-\alpha)\mathbf{v}$; $\quad \mathbf{v}_+ \leftarrow \beta\mathbf{v} + (1-\beta)(\mathbf{y} - \frac{1}{\mu}\nabla f(\mathbf{y}))$; $\quad \mathbf{x}_+ \leftarrow \mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})$.

return $(\mathbf{x}_+, \mathbf{v}_+)$

Procedure $\texttt{AcBSLS}_i\ (\mathbf{x}^{(0)}, \mathbf{v}^{(0)})$

for $t = 0, 1, \ldots, T_i - 1$ **do**

$T_i \leftarrow \begin{cases} \left\lceil \sqrt{\kappa_1} \log\left(\frac{2(f(\mathbf{x}^{(0)}) - f^\star)}{\epsilon}\right) \right\rceil & \text{if } i = 1, \\ \left\lceil \sqrt{\kappa_i}(\log(4\kappa_{\text{glob}}^4) + 1) \right\rceil & \text{otherwise.} \end{cases}$

$(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \leftarrow \texttt{AGD}(\mathbf{x}^{(t)}, \mathbf{v}^{(t)}; L_i, \mu_i)$

if $i < m$ **then**

$(\mathbf{x}^{(t+1)}, \_) \leftarrow \texttt{AcBSLS}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})$; $\quad (\_, \mathbf{v}^{(t+1)}) \leftarrow \texttt{AcBSLS}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)})$

else

$(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) \leftarrow (\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)})$

return $(\mathbf{x}^{(T_i)}, \mathbf{v}^{(T_i)})$

We specialize the initialization $\mathbf{x}^{(0)}$ and $\mathbf{v}^{(0)}$ to be the same to simplify the exposition of the theorem. We present and prove a general version of Theorem 8 with arbitrary $\mathbf{x}^{(0)}, \mathbf{v}^{(0)}$ in Appendix C.

Theorem 8 (Simplified from Theorem 15) *In the multiscale optimization (Theorem 1), for any $\mathbf{x}^{(0)}$ and $\epsilon > 0$, $\texttt{AcBSLS}(\mathbf{x}^{(0)}, \mathbf{x}^{(0)})$ returns an ϵ-optimal solution with $\mathcal{O}\left(\left(\prod_{i\in[m]} \sqrt{\kappa_i}\right) \cdot \left(\log^{m-1} \kappa_{\text{glob}}\right) \cdot \log\left(\frac{f(\mathbf{x}^{(0)}) - f^\star}{\epsilon}\right)\right)$ gradient queries when $\{(\mu_i, L_i), i \in [m]\}$ are known. Moreover in the case where $\{(\mu_i, L_i), i \in [m]\}$ are unknown and only m, μ_1, L_m and $\pi_\kappa = \prod_{i\in[m]} \kappa_i$ are known, we can achieve the same asymptotic sample complexity (up to constant factors suppressed in the $\mathcal{O}(\cdot)$).*

Similar to (un-accelerated) BSLS, under finite-precision arithmetic, AcBSLS can also attain the same rate of convergence with only logarithmic bits of precision. We defer the formal discussion to Appendix C.2.

3.4. Lower bound for the multi-scale optimization problem

We demonstrate the optimality of AcBSLS (up to poly-logarithmic factors) by establishing the minimax complexity lower bound of the multiscale optimization problem (Theorem 1) across first-order deterministic algorithm. We start by introducing the formal definition of a first-order deterministic algorithm from Carmon et al. (2021).

Definition 9 (Definition of first-order deterministic algorithms from Carmon et al. (2021)) *An algorithm* A *operating on* $f : \mathbb{R}^d \to \mathbb{R}$ *is a* ***first-order deterministic algorithm*** *if it produces iterates* $\{\mathbf{x}^{(t)}\}_{t=1}^{\infty}$ *of the form* $\mathbf{x}^{(t)} = \mathsf{A}^{(t)}(f(\mathbf{x}^{(1)}), \nabla f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(t-1)}), \nabla f(\mathbf{x}^{(t-1)}))$, *where* $\mathsf{A}^{(t)} : \mathbb{R}^{(d+1)(t-1)} \to \mathbb{R}^d$ *is measurable (the dependency on* d *is implicit).*

Note that the algorithm class considered in Theorem 9 is fairly general. For example, the definition does not require the algorithm to query points in the span of the previous gradients as in the some classic literature (Nemirovsky, 1991, 1992; Nesterov, 2018) (we refer readers to Carmon et al. (2020) for more detailed discussions on the generality of this function class). The formal statement of our lower bound is as follows and the proof is relegated to Appendix D.

Theorem 10 (Lower bound of first-order deterministic algorithms for multiscale optimization) *For any* μ_j, L_j *such that* $\min_{j\in[m]} \kappa_j \geq 2$, *for any deterministic first-order algorithm* A *defined in Theorem 9, for any* $t \in \mathbb{N}$, *there exists an objective* f *satisfying Theorem 1 with* $\|\nabla f(\mathbf{0})\|_2 \leq \Delta_{\text{grad}}$ *such that*

$$\min_{\tau\in[t]} \|\nabla f(\mathbf{x}^{(\tau)})\|_2 \geq \exp\left(-\frac{8t}{\sqrt{\prod_{i\in[m]} \kappa_i} \cdot \prod_{i\in[m-1]} \left(0.03 \cdot \log(16\frac{\mu_{i+1}}{L_i})\right)}\right) \Delta_{\text{grad}}.$$

Theorem 10 shows that our proposed `AcBSLS` is optimal up-to a poly-log factor due to the shared polynomial dependency $\Theta(\prod_{i\in[m]} \sqrt{\kappa_i})$. Theorem 10 also reveals the necessity of the poly-logarithmic dependence on κ_{glob}. For example, when the spectrum bands are evenly spaced such that $\frac{\mu_{i+1}}{L_i} \equiv \kappa_{\text{gap}}$ and $\kappa_i \ll \kappa_{\text{gap}}$, then $\prod_{i=1}^{m-1} \log \frac{\mu_{i+1}}{L_i} \approx \log^{m-1}(\kappa_{\text{glob}}^{\frac{1}{m-1}}) = \frac{\log^{m-1}(\kappa_{\text{glob}})}{(m-1)^{m-1}}$, which yields the same asymptotic dependency on κ_{glob} as in the upper bound of `AcBSLS` (Theorem 15).

3.5. Stochastic BSLS

Recall the stochastic version of a quadratic multiscale optimization problem from Theorem 4. We find that a variant of the `BSLS` algorithm efficiently solves this problem. We define the stochastic analog of `BSLS` in Algorithm 3, which we call `StochBSLS`.

Our proofs require that the distribution $\mathcal{D}$ generating samples $(\mathbf{a}, b)$ must satisfy a kind of "second-order independence" in the projected space $\mathbf{Pa}$; for any triple of distinct i, j, k we must have that $\mathbb{E}[(\mathbf{Pa})_i(\mathbf{Pa})_k^2(\mathbf{Pa})_j] = 0$. Note that this assumption is satisfied for natural distributions such as $\mathbf{a} \sim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma})$ whenever $\mathbf{P}$ diagonalizes the covariance $\mathbf{\Sigma}$. For a few more non-trivial examples of distributions which satisfy this assumption see Theorem 95. We define $\mathsf{Kurt}(\mathcal{D})$ as the kurtosis of the distribution, which is the smallest constant such that for any $\mathbf{w} \in \mathbb{R}^d$, $\mathbb{E}[(\mathbf{w}^\top \mathbf{a})^4] \leq \mathsf{Kurt}(\mathcal{D})\mathbb{E}[(\mathbf{w}^\top \mathbf{a})^2]^2$. In the case where $\mathbf{a} \sim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma})$ we have $\mathsf{Kurt}(\mathcal{D}) = 3$. The kurtosis of the distribution will play a role in the necessary number of stochastic gradient queries taken by `StochBSLS`. We establish the following theorem, the proof of which is relegated to Appendix E.

Theorem 11 *Consider the stochastic quadratic multiscale optimization problem from Definition 4. Suppose* $\mathcal{D}$ *is such that for any* $i, j, k \in [d]$, $\mathbb{E}[(\mathbf{Pa})_i(\mathbf{Pa})_k^2(\mathbf{Pa})_j] = 0$, *unless* $i = j$ *and for any* $\mathbf{w} \in \mathbb{R}^d$, $\mathbb{E}[(\mathbf{w}^\top \mathbf{a})^4] \leq \mathsf{Kurt}(\mathcal{D})\mathbb{E}[(\mathbf{w}^\top \mathbf{a})^2]^2$. *If* $m \leq \log(\kappa_{\text{glob}})/3$ *and* $\{\mu_i, L_i\}_{i\in m}$ *are known, then given any* $\mathbf{x}^{(0)}$ *let* $T_1 = \left\lceil \kappa_1 \log(9 \left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2 / \epsilon) \right\rceil$, $\quad T_i = 8 \left\lceil \kappa_i \log(\kappa_{\text{glob}}) \right\rceil$, *for* $i = 2, \ldots m$, *and* $n_{avg} \geq \mathsf{Kurt}(\mathcal{D}) dm^2 \left(\prod_{i\in[m]} T_i\right) \left(\max_{i\in[m]} T_i\right)$. *Then* $\texttt{StochBSLS}_1\left(\mathbf{x}^{(0)}\right)$ *returns*

Algorithm 3 Stochastic Variant of BSLS Algorithm

Procedure $\texttt{SGD}(\mathbf{x}; L)$

$\mathbf{g} \leftarrow \mathbf{0}$
for $i = 1, \ldots, n_{\text{avg}}$ **do**
  Receive $(\mathbf{a}^{(i)}, b^{(i)})$ and update $\mathbf{g} \leftarrow \mathbf{g} + \frac{1}{n_{\text{avg}}}\left((\mathbf{a}^{(i)\top}\mathbf{x}) - b^{(i)}\right)\mathbf{a}^{(i)}$
return $\mathbf{x} - \frac{1}{L} \cdot \mathbf{g}$

Procedure $\texttt{StochBSLS}_i\left(\mathbf{x}^{(0)}\right)$

$T_i \leftarrow \begin{cases} \left\lceil \kappa_1 \log(9 \left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2 / \epsilon) \right\rceil & \text{if } i = 1 \\ 8 \left\lceil \kappa_i \log(\kappa_{\text{glob}}) \right\rceil & \text{otherwise} \end{cases}$
for $t = 0, 1, \ldots, T_i - 1$ **do**
  $\tilde{\mathbf{x}}^{(t)} \leftarrow \texttt{StochBSLS}_{i+1}(\mathbf{x}^{(t)})$ if $i < m$, or $\mathbf{x}_t$ otherwise
  $\mathbf{x}^{(t+1)} \leftarrow \texttt{SGD}(\tilde{\mathbf{x}}^{(t)}; L_i)$
return $\texttt{StochBSLS}_{i+1}(\mathbf{x}^{(T_i)})$ if $i < m$, or $\mathbf{x}^{(T_i)}$ otherwise

an ϵ-optimal solution in expectation using $\tilde{\mathcal{O}}(d)$ space , with a total of $\mathcal{O}\left(n_{avg} \cdot 2^m \cdot \prod_{i \in [m]} T_i\right)$ queries of $(\mathbf{a}, b) \sim \mathcal{D}$. If only μ_1, L_m, and $\prod_{i \in m} \kappa_i$ are known and there exists some K such that

$$\left\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\right\|_2 \leq K \left(\mathbb{E}_{\mathbf{a} \sim \mathcal{D}} \left\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\right\|_2^2\right)^{1/2}, \tag{3.1}$$

then we can solve the stochastic quadratic multiscale optimization problem with an extra multiplicative factor of $\mathcal{O}\left(K^2 d \log \frac{4d}{\delta}\left(1 + \sqrt{\frac{\varepsilon}{\delta}}\right)\right)$ more queries of $(\mathbf{a}, b) \sim \mathcal{D}$.

4. Implications and future directions

We view the multiscale optimization problem and our algorithmic results as promising first steps towards obtaining a more fine-grained complexity of convex optimization which goes beyond condition number. Though we give near-optimal rates for solving a class of smooth strongly-convex optimization problems, our work still leaves a number of open directions. Key among them are whether we can design methods with the full practical flexibility and applicability of methods like non-linear CG and limited-memory Quasi-Newton methods that have theoretical grounding as well, in the sense that they solve the types of problems that this work proposes. For instance, is there a variant of non-linear CG or limited-memory Quasi-Newton methods that provably solves our multiscale optimization problem, or a stochastic version of CG which solves the stochastic quadratic problem? More broadly, our work raises several intriguing questions regarding the role of memory in optimization, and when it is possible to achieve the convergence rates of second-order methods with only linear memory. Further, though we have established lower bounds on multiple modifications of our multiscale optimization problem, there are several natural related problems for which it remains open to develop fast methods—for example, problems for which the Hessian has some sort of consistent multi-scale structure and cases where the problems at different scales interact instead of being completely orthogonal.

We now further elaborate on some of these implications and directions.

Space limited optimization. Recall that both `BSLS` and `StochBSLS` work in $\tilde{\mathcal{O}}(d)$ space, and `AcBSLS` uses $\tilde{\mathcal{O}}(dm)$ space. Despite using linear memory, our algorithms only suffer a polylogarithmic dependence on the overall condition number κ_{glob}. In this context, they serve as a bridge between quadratic-memory second-order methods which achieve a logarithmic dependence on the condition number, and previous linear-memory first-order methods which usually have a worse polynomial dependence on the condition number. For the stochastic case, we are unaware of any previous algorithm which only uses linear memory but still has a polylogarithmic dependence on κ_{glob}. In fact, some recent work (Sharan et al., 2019; Woodworth and Srebro, 2019) conjectures that a polynomial dependence on κ_{glob} is in general unavoidable for sub-quadratic memory algorithms. Our work shows that, at least for the structured problems we consider, it is possible to match the polylogarithmic dependence on κ_{glob} of second-order methods, while only using linear memory, and raises the question of whether this is possible for a larger class of problems.

History and structure in accelerated methods. Our near-optimal accelerated method stores up to $2m$ points at a time; this is in contrast to CG, non-linear CG, and standard accelerated methods (Nesterov, 1983) which store at most two points. It is an interesting open problem as to whether our space bound for accelerated methods could be improved. If not this raises several questions about the power of using additional history and memory in first-order methods.

Stochastic CG. We note that the `StochBSLS` algorithm for the stochastic quadratic version of the multiscale optimization problem does not obtain an accelerated convergence rate. We suspect that the natural stochastic analog of CG where we approximate any matrix-vector products over a sufficiently large set of samples *does* obtain an accelerated convergence rate for the stochastic quadratic problem, and showing this is an interesting direction for future work. This algorithm would additionally have the desirable property of not needing to guess the eigenvalues of the quadratic problem nor requiring a step size schedule.

Optimization problems with diagonal scaling. Another interesting direction is to consider non-linear, convex optimization problems which are diagonally scaled ($\mathbf{x} \to \mathbf{D}\mathbf{x}$ for a diagonal matrix $\mathbf{D}$). This does not directly fit within our framework of Theorem 1 because the different scales could interact, but we believe the ideas in this paper may extend to this setting. We remark that our results do apply to the quadratic version of this problem and believe that methods like Newton-CG may be applicable in the non-quadratic case. Further understanding and extending this setting could pave the way for developing algorithms beyond `AdaGrad` for handling scaling in optimization problems.

Acknowledgements

We thank the anonymous reviewers for their suggestions and comments on earlier versions of this paper. JK was supported in part by NSF awards CCF-1955217, CCF-1565235, and DMS-2022448. AS was supported in part by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research award, and a Sloan Research Fellowship. GV and AM were supported by NSF awards 1704417 and 1813049, ONR Young Investigator Award N00014-18-1-2295 and DOE award DE-SC0019205. HY was supported in part by the TOTAL Innovation Scholars program.

References

N.I. Achieser. über einige Funktionen, die in gegebenen Intervallen am wenigsten von Null abweichen. *Bull. Soc. Phys.-Mathem. Kazan*, III(3), 1928.

N.I. Achieser. über einige Funktionen, welche in zwei gegebenen Intervallen am wenigsten von Null abweichen I. Teil. *Bull. Acad. Sci. URSS*, VII(9), 1932.

N.I. Achieser. über einige Funktionen, welche in zwei gegebenen Intervallen am wenigsten von Null abweichen II. Teil. *Bull. Acad. Sci. URSS*, VII(3), 1933a.

N.I. Achieser. über einige Funktionen, welche in zwei gegebenen Intervallen am wenigsten von Null abweichen III, Teil. *Bull. Acad. Sci. URSS*, VII(4), 1933b.

N.I. Achieser. über eine Eigenschaft der "elliptischen" Polynome. *Comm. Kharkov Math. Soc.*, (9), 1934.

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187, 2017.

Naman Agarwal, Surbhi Goel, and Cyril Zhang. Acceleration via fractal learning rate schedules. In *International Conference on Machine Learning*, pages 87–99. PMLR, 2021.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Gökalp Alpan, Alexander Goncharov, and Burak Hatinoğlu. Some Asymptotics for Extremal Polynomials. In *Computational Analysis*, volume 155. Springer International Publishing, 2016.

V. V. Andrievskii. On the Green Function for a Complement of a Finite Number of Real Intervals. *Constructive Approximation*, 20(4), 2004.

Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

A I Aptekarev. Asymptotic Properties of Polynomials Orthogonal on a System of Contours and Periodic Motions of Toda Lattrices. *Mathematics of the USSR-Sbornik*, 53(1), 1986.

Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

Mark Braverman, Elad Hazan, Max Simchowitz, and Blake Woodworth. The gradient complexity of linear regression. In *Conference on Learning Theory*, pages 627–647. PMLR, 2020.

C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1), 1970.

Yair Carmon and John C. Duchi. Analysis of krylov subspace solutions of regularized non-convex quadratic problems. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, pages 10728–10738, 2018.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1-2), 2020.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: First-order methods. *Mathematical Programming*, 185(1-2), 2021.

Frank E Curtis, Daniel P Robinson, Clément W Royer, and Stephen J Wright. Trust-region newton-cg with strong second-order complexity guarantees for nonconvex optimization. *SIAM Journal on Optimization*, 31(1):518–544, 2021.

Wei Dai, Brian C Rider, and Youjian Liu. Volume growth and general rate quantization on grassmann manifolds. In *IEEE GLOBECOM 2007-IEEE Global Telecommunications Conference*, pages 1441–1445. IEEE, 2007.

Yu-Hong Dai. *Nonlinear Conjugate Gradient Methods*. John Wiley & Sons, Inc., 2011.

William C. Davidon. Variable Metric Method for Minimization. *SIAM Journal on Optimization*, 1 (1), 1991.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

Tobin A. Driscoll, Kim-Chuan Toh, and Lloyd N. Trefethen. From Potential Theory to Matrix Iterations in Six Steps. *SIAM Review*, 40(3), 1998.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Mark Embree and Lloyd N. Trefethen. Green's Functions for Multiply Connected Domains via Conformal Mapping. *SIAM Review*, 41(4), 1999.

Marian J Fabian, René Henrion, Alexander Y Kruger, and Jivrí V Outrata. Error bounds: necessary and sufficient conditions. *Set-Valued and Variational Analysis*, 18(2):121–149, 2010.

Bernd Fischer. Chebyshev polynomials for disjoint compact sets. *Constructive Approximation*, 8(3), 1992.

Bernd Fischer. *Polynomial Based Iteration Methods for Symmetric Linear Systems*. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2011.

R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2), 1964.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz. Solving ridge regression using sketched preconditioned svrg. In *International Conference on Machine Learning*, pages 1397–1405. PMLR, 2016.

Joseph Frank Grcar. *Analyses of the Lanczos Algorithm and of the Approximation Problem in Richardson's Method*. PhD thesis, University of Illinois at Urbana-Champaign, 1981.

A. Greenbaum. Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra and its Applications*, 113, 1989a.

Anne Greenbaum. Behavior of slightly perturbed lanczos and conjugate-gradient recurrences. *Linear Algebra and its Applications*, 113:7–63, 1989b.

Anne Greenbaum. *Iterative methods for solving linear systems*. SIAM, 1997.

Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.

Sergey Guminov and Alexander Gasnikov. Accelerated methods for α-weakly-quasi-convex problems. *arXiv preprint arXiv:1710.00797*, 2017.

William W Hager and Hongchao Zhang. A survey of nonlinear conjugate gradient methods. *Pacific Journal of Optimization*, 2(1), 2006.

M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6), 1952.

Oliver Hinder, Aaron Sidford, and Nimit Sharad Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1894–1938. PMLR, 2020.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Arno B. J. Kuijlaars. Convergence Analysis of Krylov Subspace Iterations with Methods from Potential Theory. *SIAM Review*, 48(1), 2006.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3), 1989.

Ji Liu and Stephen Wright. An accelerated randomized kaczmarz algorithm. *Mathematics of Computation*, 85(297):153–178, 2016.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

D. S. Lubinsky. A survey of general orthogonal polynomials for weights on finite and infinite intervals. *Acta Applicandae Mathematica*, 10(3), 1987.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019.

J. C. Mason and D. C. Handscomb. *Chebyshev Polynomials*. Chapman & Hall/CRC, 2003.

C. Musco, C. Musco, and A. Sidford. Stability of the Lanczos Method for Matrix Function Approximation. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2018a.

Cameron Musco, Praneeth Netrapalli, Aaron Sidford, Shashanka Ubaru, and David P. Woodruff. Spectrum approximation beyond fast matrix multiplication: Algorithms and hardness. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference, ITCS*, volume 94 of *LIPIcs*, pages 8:1–8:21, 2018b.

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Math. Program.*, 155(1-2):549–573, 2016. doi: 10.1007/s10107-015-0864-7.

A.S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

Arkadi S Nemirovsky. On optimality of krylov's information when solving linear operator equations. *Journal of Complexity*, 7(2):121–130, 1991.

Arkadi S Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 269:543–547, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, Cham, second edition, 2018.

Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal-dual accelerated gradient descent with line search for convex and nonconvex optimization problems. *arXiv preprint arXiv:1809.05895*, 2018.

Paul Nevai. Géza Freud, orthogonal polynomials and Christoffel functions. A case study. *Journal of Approximation Theory*, 48(1), 1986.

Jorge Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151), 1980.

Jorge Nocedal. Conjugate gradient methods and nonlinear optimization. *Linear and nonlinear conjugate gradient-related methods*, pages 9–23, 1996.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2nd ed edition, 2006a.

Jorge Nocedal and Stephen J Wright. Conjugate gradient methods. *Numerical optimization*, pages 101–134, 2006b.

Michael L. Overton. *Numerical Computing with IEEE Floating Point Arithmetic*. Society for Industrial and Applied Mathematics, 2001.

Christopher Conway Paige. *The computation of eigenvalues and eigenvectors of very large sparse matrices.* PhD thesis, University of London, 1971.

Jong-Shi Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1): 299–332, 1997.

Franz Peherstorfer. Orthogonal and extremal polynomials on several intervals. *Journal of Computational and Applied Mathematics*, 48(1-2), 1993.

Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

E. Polak and G. Ribiere. Note on the convergence of conjugate direction methods. *ESAIM: Mathematical Modeling and Numerical Analysis - Mathematical Modeling and Numerical Analysis*, 3(R1), 1969.

Boris T Polyak. Introduction to optimization. *Inc., Publications Division, New York*, 1, 1987.

Boris Teodorovich Polyak. The conjugate gradient method in extremal problems. *USSR Computational Mathematics and Mathematical Physics*, 9(4):94–112, 1969.

Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods. *Mathematical Programming*, 174(1):293–326, 2019.

Clément W. Royer, Michael O'Neill, and Stephen J. Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, 2019.

Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

EB Saff. Logarithmic Potential Theory with Applications to Approximation Theory. *Surveys in Approximation Theory*, 5, 2010.

K. Schiefermayr and F. Peherstorfer. Description of Extremal Polynomials on Several Intervals and their Computation. I. *Acta Mathematica Hungarica*, 83(1/2), 1999.

Klaus Schiefermayr. A lower bound for the minimum deviation of the chebyshev polynomial on a compact real set. *East Journal on Approximations*, 14, 2008.

Klaus Schiefermayr. A Lower Bound for the Norm of the Minimal Residual Polynomial. *Constructive Approximation*, 33(3), 2011a.

Klaus Schiefermayr. Estimates for the asymptotic convergence factor of two intervals. *Journal of Computational and Applied Mathematics*, 236(1), 2011b.

Klaus Schiefermayr. An upper bound for the norm of the Chebyshev polynomial on two intervals. *Journal of Mathematical Analysis and Applications*, 445(1), 2017.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *arXiv preprint arXiv:1309.2375*, 2013.

Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Memory-sample tradeoffs for linear regression with small error. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 890–901, 2019.

J. Shen, G. Strang, and A. J. Wathen. The Potential Theory of Several Intervals and Its Applications. *Applied Mathematics and Optimization*, 44(1), 2001.

Max Simchowitz. On the randomized complexity of minimizing a convex quadratic function. *arXiv preprint arXiv:1807.09386*, 2018.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

Thomas Strohmer and Roman Vershynin. A randomized solver for linear systems with exponential convergence. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 499–507. Springer, 2006.

Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, 1997.

Huynh Van Ngai and Jean-Paul Penot. Approximately convex functions and approximately monotonic operators. *Nonlinear Analysis: Theory, Methods & Applications*, 66(3):547–564, 2007.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2019.

Harold Widom. Extremal polynomials associated with a system of curves in the complex plane. *Advances in Mathematics*, 3(2), 1969.

Blake Woodworth. The minimax complexity of distributed optimization. *arXiv preprint arXiv:2109.00534*, 2021.

Blake Woodworth and Nathan Srebro. Lower bound for randomized first order convex optimization. *arXiv preprint arXiv:1709.03594*, 2017.

Blake Woodworth and Nathan Srebro. Open problem: The oracle complexity of convex optimization with limited memory. In *Conference on Learning Theory*, pages 3202–3210. PMLR, 2019.

Peng Xu, Fred Roosta, and Michael W Mahoney. Newton-type methods for non-convex optimization under inexact hessian information. *Mathematical Programming*, 184(1):35–70, 2020.

Maryam Yashtini. On the global convergence rate of the gradient descent method for functions with hölder continuous gradients. *Optimization letters*, 10(6):1361–1370, 2016.

Honglin Yuan and Tengyu Ma. Federated Accelerated Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems 33*, 2020.

Hui Zhang and Lizhi Cheng. Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization. *Optimization Letters*, 9(5):961–979, 2015.

Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. *Advances in Neural Information Processing Systems*, 26:980–988, 2013.

Contents

Appendix A. General notation

Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. We use bold lower-case letter (e.g., $\mathbf{x}$) to denote vectors, bold upper-case letter (e.g., $\mathbf{A}$) to denote matrices. We use $\mathbf{I}$ to denote identity matrix, $\mathbf{1}$ to denote all-1 vector, $\mathbf{0}$ to denote all-zero vectors or matrices, $\mathbf{e}_i$ to denote i-th unit vector (i-th column of $\mathbf{I}$). When comparing two vectors or matrices, the ordinary inequality signs $(\leq, \geq)$ denote element-wise inequality. For example, $\mathbf{A} \geq \mathbf{0}$ means $\mathbf{A}$ is a non-negative matrix. When comparing two matrices, $(\preceq, \succeq)$ denote spectrum inequality. For example, $\mathbf{A} \succeq \mathbf{0}$ means $\mathbf{A}$ is a positive semi-definite matrix. We use $\|\cdot\|_1$ to denote vector ℓ_1 or matrix ℓ_1-operator (row-sum) norm, $\|\cdot\|_2$ to denote vector ℓ_2 norm or matrix ℓ_2-operator (spectrum) norm. For any function f we use $f^\star$ to denote the optimum (minimum) value of f.

Appendix B. BSLS algorithm for multiscale optimization

In this section, we provide the formal proof of Theorem 6 in Appendix B.1, and then formally establish the stability of BSLS in Appendix B.2.

B.1. Proof of Theorem 6 and Theorem 7: BSLS under exact arithmetic

Here we formalize the aforementioned intuition and prove Theorem 6. To begin, we first study the effect of $\mathtt{GD}(\mathbf{x}; L_i)$ on various subspaces $j \in [m]$.

Lemma 12 *In the setting of Theorem 6, for any* $\mathbf{x}$ *and* $i, j \in [m]$,

$$f_j(\mathbf{P}_j \mathtt{GD}(\mathbf{x}; L_i)) - f_j^\star \leq \left(f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star\right) \cdot \begin{cases} 1 & j < i \\ 1 - \kappa_i^{-1} & j = i \\ \kappa_{\text{glob}}^2 & j \geq i. \end{cases}$$

Proof [Proof of Lemma 12] Let $\mathbf{x}_+$ denote the result of $\mathtt{GD}(\mathbf{x}; L_i)$. By L_j-smoothness of f_j, we have

$$f_j(\mathbf{P}_j\mathbf{x}_+) - f_j^\star \leq f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star + \left(-\frac{1}{L_i} + \frac{L_j}{2L_i^2}\right)\|\nabla f_j(\mathbf{P}_j\mathbf{x})\|_2^2. \tag{B.1}$$

Now we consider the three possible cases $j = i$, $j < i$ and $j > i$ separately.

(a) For $j = i$, the inequality Eq. (B.1) becomes

$$f_j(\mathbf{P}_j\mathbf{x}_+) - f_j^\star \leq f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star - \frac{1}{2L_j}\|\nabla f_j(\mathbf{P}_j\mathbf{x})\|_2^2.$$

By μ_j-strong-convexity of f_j we have $\|\nabla f_j(\mathbf{P}_j\mathbf{x})\|_2^2 \geq 2\mu_j(f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star)$. Thus $f_j(\mathbf{P}_j\mathbf{x}_+) - f_j^\star \leq (1 - \kappa_j^{-1})(f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star)$.

(b) For $j < i$, the coefficient of the second term of Eq. (B.1) is non-positive since $L_j \leq L_i$. Hence $f_j(\mathbf{P}_j\mathbf{x}_+) - f_j^\star \leq f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star$.

(c) For $j > i$, first observe that by μ_j-strong convexity and L_j-smoothness of f_j, one has $2L_j(f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star) \geq \|\nabla f_j(\mathbf{P}_j\mathbf{x})\|_2^2 \geq 2\mu_j(f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star)$. Therefore by Eq. (B.1), we have

$$f_j(\mathbf{P}_j\mathbf{x}_+) - f_j^\star \leq \left(1 - \frac{2\mu_j}{L_i} + \frac{L_j^2}{L_i^2}\right)(f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star) \leq \left(-1 + \kappa_{\text{glob}}^2\right)(f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star),$$

where the last inequality is due to $\mu_j \geq L_i$ and $\frac{L_j}{L_i} \leq \kappa_{\text{glob}}$ by definition of κ_{glob}.

Summarizing the above cases completes the proof of Theorem 12. ■

With Lemma 12 at hand we are ready to prove Theorem 6.

Proof [Proof of Theorem 6] By expanding the $\texttt{BSLS}$ procedure, we observe that $\texttt{BSLS}_1(\cdot)$ consists of $T_j \cdot \prod_{k=1}^{j-1}(T_k+1)$ steps of $\texttt{GD}(\cdot; L_j)$ in total, for $j \in [m]$. Therefore, by Theorem 12, for any $i \in [m]$, the following inequality holds

$$f_i(\mathbf{P}_i\texttt{BSLS}_1(\mathbf{x}^{(0)})) - f_i^\star \leq \kappa_{\text{glob}}^{2\sum_{j=1}^{i-1} T_j \prod_{k=1}^{j-1}(T_k+1)} \cdot \left(1 - \kappa_i^{-1}\right)^{T_i \prod_{j=1}^{i-1}(T_j+1)} (f_i(\mathbf{P}_i\mathbf{x}^{(0)}) - f_i^\star)$$

$$\leq \exp\left(\underbrace{2\log\kappa_{\text{glob}} \cdot \sum_{j=1}^{i-1} T_j \prod_{k=1}^{j-1}(T_k+1) - \kappa_i^{-1} T_i \prod_{j=1}^{i-1}(T_j+1)}_{\text{denoted as } \gamma_i}\right)(f_i(\mathbf{P}_i\mathbf{x}^{(0)}) - f_i^\star).$$

(since $1 - x \leq e^{-x}$)

It remains to upper bound γ_i. For $i = 1$, by definition, we have $\gamma_i := -\kappa_1^{-1}T_1 \leq \log\left(\frac{\epsilon}{f(\mathbf{x}^{(0)}) - f^\star}\right)$ due to the choice of T_1. For $i > 1$, we observe that

$$\gamma_i - \gamma_{i-1} = 2\log\kappa_{\text{glob}} \cdot T_{i-1} \cdot \prod_{j=1}^{i-2}(T_j+1) - \kappa_i^{-1}T_i \prod_{j=1}^{i-1}(T_j+1) + \kappa_{i-1}^{-1}T_{i-1}\prod_{j=1}^{i-2}(T_j+1)$$

$$\leq \left(T_{i-1}\prod_{j=1}^{i-2}(T_j+1)\right) \cdot \left(-\kappa_i^{-1}T_i + \kappa_{i-1}^{-1} + 2\log\kappa_{\text{glob}}\right).$$

Since $T_i \geq \kappa_i(2\log\kappa_{\text{glob}} + 1)$ (due to the choice of T_i) we obtain $\gamma_i - \gamma_{i-1} \leq \left(\prod_{j=1}^{i-1} T_j\right) \cdot \left(-1 + \kappa_{i-1}^{-1}\right) \leq 0$. Consequently, $\gamma_m \leq \gamma_{m-1} \leq \cdots \leq \gamma_1 \leq \log\left(\frac{\epsilon}{f(\mathbf{x}^{(0)}) - f^\star}\right)$. Therefore for all $i \in [m]$, $f_i(\mathbf{P}_i\texttt{BSLS}_1(\mathbf{x}^{(0)})) - f_i^\star \leq \exp(\gamma_i)\left(f_i(\mathbf{P}_i\mathbf{x}^{(0)}) - f_i^\star\right) \leq \frac{\epsilon}{f(\mathbf{x}^{(0)}) - f^\star}\left(f_i(\mathbf{P}_i\mathbf{x}^{(0)}) - f_i^\star\right)$. Summing over all $i \in [m]$ gives $f(\texttt{BSLS}_1(\mathbf{x}^{(0)})) - f^\star \leq \epsilon$.

To show the last part of Theorem 6 regarding the setting where the parameters $\{(\mu_i, L_i), i \in [m]\}$ are unknown, we do a black-box reduction from the case where the parameters are known to when only m, μ_1, L_m and π_κ are known.

Proposition 13 *Let $\pi_\kappa = \prod_{i\in[m]} \kappa_i$. An algorithm A which solves the multiscale optimization problem in Definition 1 to sub-optimality ϵ with $T(\pi_\kappa, \kappa_{\text{glob}}, m, \epsilon)$ gradient queries when the parameters (μ_i, L_i) are known, can be used to solve the multiscale optimization problem with $T(\pi_\kappa 2^{5m}, \kappa_{\text{glob}}, m, \epsilon) \cdot O(\log^m(\kappa_{\text{glob}}))$ gradient queries when only m, μ_1, L_m and π_κ are known.*

The proof Proposition 13 works by simply doing a brute force search over all the parameters over a suitable grid and appears in Appendix F.1. The last part of Theorem 6 now follows. ■

The re-ordering Theorem 7 holds because Theorem 6 only leverages the fact that $\texttt{BSLS}_1(\cdot)$ consists of $\Theta(\prod_{k=1}^{j} T_k)$ steps of $\texttt{GD}(\cdot; L_j)$ in total, for $j \in [m]$.

B.2. Theory on the stability of BSLS: why interlacing order matters

Now we verify the intuition above and theoretically justify the stability of BSLS (Algorithm 1). For clarity, let $\widehat{\texttt{GD}}$ be the finite-precision implementation of GD, and $\widehat{\texttt{BSLS}}_i$ be the finite-precision implementation of $\texttt{BSLS}_i$ by replacing GD with $\widehat{\texttt{GD}}$. To understand finite-precision behavior without going into excessive details of low-level implementation, we impose the following Requirement 1 that $\widehat{\texttt{GD}}$ returns a δ-multiplicative approximation of the exact GD. Requirement 1 is reminiscent of the "correct rounding" requirement on basic operations in IEEE standard (c.f. Chap. 6 of Overton (2001)). Technically, if GD operator is well-conditioned (and no overflow or underflow occurs), then Req. 1 can be satisfied by a floating-point system with $\mathcal{O}(\log(1/\delta))$ bits (c.f. Chap. 12 of Overton (2001)).

Requirement 1 *There exists a $\delta < 1$ such that for any $\mathbf{x}$ and i, for $\mathbf{x}_+ \leftarrow \texttt{GD}(\mathbf{x}; L_i)$ and $\widehat{\mathbf{x}_+} \leftarrow \widehat{\texttt{GD}}(\mathbf{x}; L_i)$, it is the case that $|\widehat{\mathbf{x}_+} - \mathbf{x}_+| \le \delta|\mathbf{x}_+|$, where $|\cdot|$ denotes element-wise absolute value.*

In the following Theorem 14, we prove that finite-precision $\widehat{\texttt{BSLS}}_1$ can match the exact arithmetic rate under only logarithmic bits of precision in that δ only has to be polynomially small. As a conclusion, BSLS can be implemented stably with $\tilde{\mathcal{O}}(d)$ bits of memory. We specialize the initialization $\mathbf{x}^{(0)}$ to $\mathbf{0}$ to simplify the exposition of the theorem. In Appendix G, we provide and prove the general version with arbitrary $\mathbf{x}^{(0)}$.

Theorem 14 (BSLS under finite-precision arithmetic) *Consider multiscale optimization problem (Theorem 1), for any $\epsilon > 0$, assuming Requirement 1 with*

$$\delta^{-1} \ge m \cdot (10\kappa_{\text{glob}})^{2m-1} \cdot \frac{(f(\mathbf{0}) - f^\star)}{\epsilon} \cdot \left(\prod_{i \in [m]} T_i \right),$$

then $\min\{f(\mathbf{0}), f(\widehat{\texttt{BSLS}}_1(\mathbf{0}))\} - f^\star \le 3\epsilon$ provided that $T_1, \ldots, T_m$ satisfy

$$T_1 \ge \kappa_1 \log\left(\frac{f(\mathbf{0}) - f^\star}{\epsilon}\right); \qquad T_i \ge \kappa_i(2\log(\kappa_{\text{glob}}) + 1), \quad \text{for } i = 2, \ldots, m,$$

when $\{(\mu_i, L_i), i \in [m]\}$ are known. We can also achieve the same asympototic sample complexity (up to constant factors suppressed in the $\mathcal{O}(\cdot)$) when $\{(\mu_i, L_i), i \in [m]\}$ are unknown and only m, μ_1, L_m and $\pi_\kappa = \prod_{i=1}^{m} \kappa_i$ are known.

The proof of Theorem 14 is relegated to Appendix G.

Appendix C. Accelerated BSLS algorithm for multiscale optimization

In this section, we first state and prove the extended version of Theorem 8 on the complexity of AcBSLS with general $(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})$. Then we establish the stability result of AcBSLS in Appendix C.2. Finally, we discuss in Appendix C.3 on the necessity of branching procedure in AcBSLS.

We will use standard potentials from accelerated GD to monitor the progress of AcBSLS. For any $i \in [m]$ and $\mathbf{x}$, define

$$\Delta_i(\mathbf{x}) \coloneqq f_i(\mathbf{P}_i\mathbf{x}) - f_i^\star, \qquad r_i(\mathbf{x}) \coloneqq \frac{\mu_i}{2}\|\mathbf{P}_i(\mathbf{x} - \mathbf{x}^\star)\|_2^2.$$

For any $\mathbf{x}, \mathbf{v}$ pair, define

$$\psi_i(\mathbf{x}, \mathbf{v}) \coloneqq \Delta_i(\mathbf{x}) + r_i(\mathbf{v}), \qquad \psi(\mathbf{x}, \mathbf{v}) \coloneqq \sum_{i\in[m]} \psi_i(\mathbf{x}, \mathbf{v}).$$

We establish the following theorem.

Theorem 15 (AcBSLS with exact arithmetic) *Consider multiscale optimization problem defined in Theorem 1, for any initialization $(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})$ and $\epsilon > 0$, then $\psi(\texttt{AcBSLS}_1(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})) \le \epsilon$ provided that $T_1, \ldots, T_m$ satisfy*

$$T_1 \ge \sqrt{\kappa_1}\log\left(\frac{\psi(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})}{\epsilon}\right), \qquad T_i \ge \sqrt{\kappa_i}(\log(4\kappa_{\text{glob}}^4) + 1), \quad \text{for } i = 2, \ldots, m, \tag{C.1}$$

when $\{(\mu_i, L_i), i \in [m]\}$ are known, and the total number of gradient queries which AcBSLS makes is $\mathcal{O}(\prod_{i\in[m]} T_i)$. We can also achieve the same asymptotic query complexity for finding an ϵ-optimal solution (up to constant factors suppressed in the $\mathcal{O}(\cdot)$) when $\{(\mu_i, L_i), i \in [m]\}$ are unknown and only m, μ_1, L_m and $\pi_\kappa = \prod_{i=1}^m \kappa_i$ are known.

C.1. Proof of Theorem 15: AcBSLS under exact arithmetic

The proof plan is as follows. We first study the effect of one AGD step with various "step-sizes" on each sub-objective in Appendix C.1.1. Then we inductively bound the progress of $\texttt{AcBSLS}_i$ for all i from m down to 1, with $i = 1$ being the ultimate goal (see Appendix C.1.2). Then we finish the proof of Theorem 15 in Appendix C.1.3. Note that the last part regarding the case where $\{(\mu_i, L_i), i \in [m]\}$ are unknown follows from our black-box reduction in Proposition 13 (in the same way as in the proof of Theorem 6).

C.1.1. Effect of one AGD step with various "step-sizes"

In this subsection, we study the effect of AGD on all sub-objectives f_i's. The main goal is to establish the following Lemma 16.

Lemma 16 (Effect of one AGD step with various "step-sizes") *Consider multiscale optimization (Def. 1), for any $\mathbf{x}, \mathbf{v}$ and $i \in [m]$, consider $(\mathbf{x}_+, \mathbf{v}_+) = \texttt{AGD}(\mathbf{x}, \mathbf{v}; L_i, \mu_i)$, then*

(a) (apply the right step-size) $\psi_i(\mathbf{x}_+, \mathbf{v}_+) \le \left(1 - \frac{1}{\sqrt{\kappa_i}}\right)\psi_i(\mathbf{x}, \mathbf{v})$.

(b) (apply small step-size) For any $j < i$, the following two inequalities hold

(i) $\max\{\Delta_j(\mathbf{x}_+), \Delta_j(\mathbf{v}_+)\} \le \max\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\}$;

(ii) $\max\{r_j(\mathbf{x}_+), r_j(\mathbf{v}_+)\} \le \max\{r_j(\mathbf{x}), r_j(\mathbf{v})\}$.

(c) (apply large step-size) For any $j > i$, the following three inequalities hold

(i) $\max\{r_j(\mathbf{v}_+), r_j(\mathbf{x}_+)\} \le \kappa_{\text{glob}}^2 \max\{r_j(\mathbf{v}), r_j(\mathbf{x})\}$;

(ii) $\max\{\Delta_j(\mathbf{v}_+), \Delta_j(\mathbf{x}_+)\} \le \kappa_{\text{glob}}^2 \max\{\Delta_j(\mathbf{v}), \Delta_j(\mathbf{x})\}$;

(iii) $\psi_j(\mathbf{x}_+, \mathbf{v}_+) \le 2\kappa_j \kappa_{\text{glob}}^2(\mathbf{x}, \mathbf{v})$.

Remark 17 *Lemma 16 is supposed to be the counterpart of Lemma 12 (the progress of one-step* GD *in (un-accelerated)* BSLS*). One may be tempted to establish the following (stronger) version of Lemma 16(b)*

$$\psi_j(\mathbf{x}_+, \mathbf{v}_+) \le \psi_j(\mathbf{x}, \mathbf{v}), \quad \text{if } j < i. \tag{C.2}$$

If this claim Eq. (C.2) were true, we would be able to guarantee the convergence of naive AcBSLS *(akin to* BSLS*) without using the branching procedure. Unfortunately, we can show that Eq. (C.2) is not always true, even for quadratic objective f. That is to say, the potential ψ_j may not be conservative under* $\texttt{AGD}(\cdot, \cdot; L_i, \mu_i)$ *with $i > j$ (a.k.a. AGD with "smaller step-sizes"). We provide more details on this topic in Appendix C.3, including a numerical experiment against naive* AcBSLS*.*

In Lemma 16, we instead show that $\max\{\Delta(\mathbf{x}_+), \Delta(\mathbf{v}_+)\}$, $\max\{r(\mathbf{x}_+), r(\mathbf{v}_+)\}$ *are non-increasing under AGD with smaller step-sizes. Since Lemma 16 (a) and (b) keep track of different quantities, we end up requiring the recursive branching procedure defined in* AcBSLS *(Algorithm 2).*

We now prove Lemma 16.

Proof [Proof of Lemma 16]

(a) The proof of (a) follows by standard accelerated gradient descent analysis Nesterov (2018), which we state here for completeness. For clarity, let $\kappa_i = \frac{L_i}{\mu_i}$, $\alpha_i = \frac{\sqrt{\kappa_i}}{\sqrt{\kappa_i}+1}$, $\beta_i = 1 - \frac{1}{\sqrt{\kappa_i}}$ be the corresponding κ, α, β in applying $\texttt{AGD}(\cdot, \cdot; L_i, \mu_i)$. Let us restate the recursion for clarity (we introduce an auxiliary variable $\mathbf{z}$ for ease of exposition).

$$\mathbf{y} = \alpha_i \cdot \mathbf{x} + (1 - \alpha_i) \cdot \mathbf{v}, \qquad \mathbf{z} = \beta_i \cdot \mathbf{v} + (1 - \beta_i) \cdot \mathbf{y},$$
$$\mathbf{v}_+ = \mathbf{z} - \frac{1 - \beta_i}{\mu_i} \cdot \nabla f(\mathbf{y}), \qquad \mathbf{x}_+ = \mathbf{y} - \frac{1}{L_i} \cdot \nabla f(\mathbf{y}).$$

By definition of $\mathbf{z}$, one has

$$\begin{aligned} \|\mathbf{P}_i(\mathbf{z} - \mathbf{x}^\star)\|_2^2 &\le \beta_i \|\mathbf{P}_i(\mathbf{v} - \mathbf{x}^\star)\|_2^2 + (1 - \beta_i)\|\mathbf{P}_i(\mathbf{y} - \mathbf{x}^\star)\|_2^2 && \text{(by convexity)} \\ &\le \beta_i \|\mathbf{P}_i(\mathbf{v} - \mathbf{x}^\star)\|_2^2 + \frac{2(1 - \beta_i)}{\mu_i}\left[f_i^\star - f_i(\mathbf{P}_i\mathbf{y}) + \langle \mathbf{P}_i \nabla f(\mathbf{y}), \mathbf{P}_i(\mathbf{y} - \mathbf{x}^\star)\rangle\right]. \\ & && \text{(by } \mu_i\text{-strong-convexity of } f_i) \end{aligned}$$

By definition of $\mathbf{v}_+$, one has

$$\begin{aligned}
&\|\mathbf{P}_i\left(\mathbf{v}_+ - \mathbf{x}^\star\right)\|_2^2 = \left\|\mathbf{P}_i\left(\mathbf{z} - \mathbf{x}^\star - \frac{1-\beta_i}{\mu_i}\nabla f(\mathbf{y})\right)\right\|_2^2 \\
=&\|\mathbf{P}_i(\mathbf{z}-\mathbf{x}^\star)\|_2^2 - \frac{2(1-\beta_i)}{\mu_i}\left\langle \mathbf{P}_i\nabla f(\mathbf{y}), \mathbf{P}_i(\mathbf{z}-\mathbf{x}^\star)\right\rangle + \left(\frac{1-\beta_i}{\mu}\right)^2\|\mathbf{P}_i\nabla f(\mathbf{y})\|_2^2 \\
\le&\beta_i\|\mathbf{P}_i(\mathbf{v}-\mathbf{x}^\star)\|_2^2 + \frac{2(1-\beta_i)}{\mu_i}\left[f_i^\star - f_i(\mathbf{P}_i\mathbf{y}) + \left\langle \mathbf{P}_i\nabla f(\mathbf{y}), \mathbf{P}_i(\mathbf{y}-\mathbf{x}^\star)\right\rangle\right] \\
&- \frac{2(1-\beta_i)}{\mu_i}\left\langle \mathbf{P}_i\nabla f(\mathbf{y}), \mathbf{P}_i(\mathbf{z}-\mathbf{x}^\star)\right\rangle + \left(\frac{1-\beta_i}{\mu_i}\right)^2\|\mathbf{P}_i\nabla f(\mathbf{y})\|_2^2 \\
=&\beta_i\|\mathbf{P}_i(\mathbf{v}-\mathbf{x}^\star)\|_2^2 + \frac{2(1-\beta_i)}{\mu_i}\left[f_i^\star - f_i(\mathbf{P}_i\mathbf{y}) + \underbrace{\left\langle \mathbf{P}_i\nabla f(\mathbf{y}), \mathbf{P}_i(\mathbf{y}-\mathbf{z})\right\rangle}_{①}\right] \\
&+ \left(\frac{1-\beta_i}{\mu_i}\right)^2\underbrace{\|\mathbf{P}_i\nabla f(\mathbf{y})\|_2^2}_{②}
\end{aligned} \tag{C.3}$$

Next we bound ① and ② in (C.3). First note that by definition $\mathbf{y}-\mathbf{z} = \beta_i(\mathbf{y}-\mathbf{v}) = \beta_i(\alpha_i(\mathbf{x}-\mathbf{v}))$, and $\mathbf{x}-\mathbf{y} = (1-\alpha_i)(\mathbf{x}-\mathbf{v})$, we have $\mathbf{y}-\mathbf{z} = \frac{\beta_i\alpha_i}{1-\alpha_i}(\mathbf{x}-\mathbf{y})$. Therefore ① is bounded as

$$\left\langle \mathbf{P}_i\nabla f(\mathbf{y}), \mathbf{P}_i(\mathbf{y}-\mathbf{z})\right\rangle = \frac{\beta_i\alpha_i}{(1-\alpha_i)}\left\langle \mathbf{P}_i\nabla f(\mathbf{y}), \mathbf{P}_i(\mathbf{x}-\mathbf{y})\right\rangle \le \frac{\beta_i\alpha_i}{(1-\alpha_i)}\left(f_i(\mathbf{P}_i\mathbf{x}) - f_i(\mathbf{P}_i\mathbf{y})\right), \tag{C.4}$$

where the last inequality is by convexity of f_i.

To bound ②, we note that $\mathbf{x}_+ = \mathbf{y} - \frac{1}{L_i}\nabla f(\mathbf{y})$, which implies (by L_i-smoothness of f_i)

$$f_i(\mathbf{P}_i\mathbf{x}_+) \le f_i(\mathbf{P}_i\mathbf{y}) - \left\langle \nabla f_i(\mathbf{P}_i\mathbf{y}), \frac{1}{L_i}\mathbf{P}_i\nabla f(\mathbf{y})\right\rangle + \frac{L_i}{2}\left\|\frac{1}{L_i}\mathbf{P}_i\nabla f(\mathbf{y})\right\|_2^2 = f_i(\mathbf{P}_i\mathbf{y}) - \frac{1}{2L_i}\|\mathbf{P}_i\nabla f(\mathbf{y})\|_2^2.$$

Thus ② is upper bounded as

$$\|\mathbf{P}_i\nabla f(\mathbf{y})\|_2^2 \le 2L_i\left(f_i(\mathbf{P}_i\mathbf{y}) - f_i(\mathbf{P}_i\mathbf{x}_+)\right). \tag{C.5}$$

Plugging the upper bound (C.4), (C.5) down to (C.3) yields

$$\begin{aligned}
\|\mathbf{P}_i(\mathbf{v}_+ - \mathbf{x}^\star)\|_2^2 \le{}& \beta_i\|\mathbf{P}_i(\mathbf{v}-\mathbf{x}^\star)\|_2^2 + \frac{2(1-\beta_i)}{\mu_i}\left(f_i^\star - f_i(\mathbf{P}_i y)\right) \\
&+ \frac{2(1-\beta_i)}{\mu_i}\frac{\beta_i\alpha_i}{(1-\alpha_i)}\left(f_i(\mathbf{P}_i\mathbf{x}) - f_i(\mathbf{P}_i y)\right) + \frac{2L_i(1-\beta_i)^2}{\mu_i^2}\left(f_i(\mathbf{P}_i y) - f_i(\mathbf{P}_i\mathbf{x}_+)\right).
\end{aligned}$$

Substituting $\alpha_i = \frac{\sqrt{\kappa_i}}{\sqrt{\kappa_i}+1}$ and $\beta_i = 1 - \frac{1}{\sqrt{\kappa_i}}$ gives

$$\|\mathbf{P}_i(\mathbf{v}_+ - \mathbf{x}^\star)\|_2^2 + \frac{2}{\mu_i}\left(f_i(\mathbf{P}_i\mathbf{x}_+) - f_i^\star\right) \le \left(1 - \frac{1}{\sqrt{\kappa_i}}\right)\left(\|\mathbf{P}_i(\mathbf{v}-\mathbf{x}^\star)\|_2^2 + \frac{2}{\mu_i}\left(f_i(\mathbf{P}_i\mathbf{x}) - f_i^\star\right)\right),$$

which implies $\psi_i(\mathbf{x}_+, \mathbf{v}_+) \le \left(1 - \frac{1}{\sqrt{\kappa_i}}\right)\psi(\mathbf{x}, \mathbf{v})$, completing the proof of (a).

(b) Let $\kappa_i = \frac{L_i}{\mu_i}, \alpha_i = \frac{\sqrt{\kappa_i}}{\sqrt{\kappa_i}+1}, \beta_i = 1 - \frac{1}{\sqrt{\kappa_i}}$ be the corresponding κ, α, β in applying $\texttt{AGD}(\cdot, \cdot; L_i, \mu_i)$. For clarity we restate the algorithm AGD with an auxiliary state $\mathbf{w}$

$$\begin{aligned} \mathbf{y} &= \alpha_i \cdot \mathbf{x} + (1-\alpha_i) \cdot \mathbf{v}, & \mathbf{w} &= \mathbf{y} - \frac{1}{\mu_i}\nabla f(\mathbf{y}), \\ \mathbf{v}_+ &= \beta_i \mathbf{v} + (1-\beta_i)\mathbf{w}, & \mathbf{x}_+ &= \mathbf{y} - \frac{1}{L_i} \cdot \nabla f(\mathbf{y}). \end{aligned} \tag{C.6}$$

Since $\alpha \in [0,1]$ we have (by convexity)

$$\Delta_j(\mathbf{y}) = f_j(\mathbf{P}_j\mathbf{y}) - f_j^\star \le \max\left\{f_j(\mathbf{P}_j\mathbf{x}), f_j(\mathbf{P}_j\mathbf{v})\right\} - f_j^\star = \max\left\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\right\}.$$

Since the step-size of the $\mathbf{w}$-step satisfies $\frac{1}{\mu_i} \le \frac{1}{L_j}$ by assumption $j < i$, we obtain

$$f_j(\mathbf{P}_j\mathbf{w}) \le f_j(\mathbf{P}_j\mathbf{y}) - \frac{1}{\mu_i}\left\langle \nabla f_j(\mathbf{P}_j y), \mathbf{P}_j\nabla f_j(\mathbf{y})\right\rangle + \frac{L_j}{2}\left\|\frac{1}{\mu_i}\mathbf{P}_j\nabla f_j(\mathbf{y})\right\|_2^2 \le f_j(\mathbf{P}_j\mathbf{y}).$$

For the same reason we have $f_j(\mathbf{P}_j\mathbf{x}_+) \le f_j(\mathbf{P}_j\mathbf{y})$ since the $\mathbf{x}_+$-step takes an even smaller step-size. These imply $\Delta_j(\mathbf{w}) \le \Delta_j(\mathbf{y}) \le \max\left\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\right\}$ and $\Delta_j(\mathbf{x}_+) \le \Delta_j(\mathbf{y}) \le \max\left\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\right\}$. By convexity we have $\Delta_j(\mathbf{v}_+) \le \max\left\{\Delta_j(\mathbf{v}), \Delta_j(\mathbf{w})\right\} \le \max\left\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\right\}$, which completes the proof of the first inequality. The second inequality holds for the same reason.

(c) Let $\kappa_i = \frac{L_i}{\mu_i}, \alpha_i = \frac{\sqrt{\kappa_i}}{\sqrt{\kappa_i}+1}, \beta_i = 1 - \frac{1}{\sqrt{\kappa_i}}$ be the corresponding κ, α, β in applying $\texttt{AGD}(\cdot, \cdot; L_i, \mu_i)$. For clarity we restate the algorithm AGD with an auxiliary state $\mathbf{w}$, as in (C.6).

First note that $r_j(\mathbf{y}) \le \max\left\{r_j(\mathbf{x}), r_j(\mathbf{v})\right\}$ since $\mathbf{y}$ is a convex combination of $\mathbf{x}$ and $\mathbf{v}$. Now we analyze $r_j(\mathbf{w})$

$$\begin{aligned} &\frac{2}{\mu_j}r_j(\mathbf{w}) = \left\|\mathbf{P}_j\left(\mathbf{y} - \frac{1}{\mu_i}\nabla f(\mathbf{y}) - \mathbf{x}^\star\right)\right\|_2^2 \\ &= \|\mathbf{P}_j(\mathbf{y}-\mathbf{x}^\star)\|_2^2 - \frac{2}{\mu_i}\left\langle \mathbf{P}_j\nabla f(\mathbf{y}), \mathbf{P}_j(\mathbf{y}-\mathbf{x}^\star)\right\rangle + \frac{1}{\mu_i^2}\|\mathbf{P}_j\nabla f(\mathbf{y})\|_2^2 \\ &\le \left(1 - \frac{\mu_j}{\mu_i}\right)\|\mathbf{P}_j(\mathbf{y}-\mathbf{x}^\star)\|_2^2 + \left(-\frac{1}{\mu_i L_j} + \frac{1}{\mu_i^2}\right)\|\mathbf{P}_j\nabla f(\mathbf{y})\|_2^2 \\ &\le \left(1 - \frac{\mu_j}{\mu_i}\right)\|\mathbf{P}_j(\mathbf{y}-\mathbf{x}^\star)\|_2^2 + \left(-\frac{L_j}{\mu_i} + \frac{L_j^2}{\mu_i^2}\right)\|\mathbf{P}_j(\mathbf{y}-\mathbf{x}^\star)\|_2^2 \\ &\le \frac{L_j^2}{\mu_i^2}\|\mathbf{P}_j(\mathbf{y}-\mathbf{x}^\star)\|_2^2 = \frac{L_j^2}{\mu_i^2}r_j(\mathbf{y}), \end{aligned}$$

Since $\mathbf{v}_+$ is a convex combination of $\mathbf{w}$ and $\mathbf{v}$, we obtain

$$r_j(\mathbf{v}_+) \le \max\left\{r_j(\mathbf{w}), r_j(\mathbf{v})\right\} \le \max\left\{\frac{L_j^2}{\mu_i^2}r_j(\mathbf{y}), r_j(\mathbf{v})\right\} \le \frac{L_j^2}{\mu_i^2}\max\left\{r_j(\mathbf{x}), r_j(\mathbf{v})\right\}.$$

Similarly we have

$$r_j(\mathbf{x}_+) \le \frac{L_j^2}{L_i^2}r_j(\mathbf{y}) \le \frac{L_j^2}{L_i^2}\max\left\{r_j(\mathbf{x}), r_j(\mathbf{v})\right\},$$

which yields the first inequality of (c). The second inequality of (c) holds for the same reason. The third inequality holds because

$$\begin{aligned}
&\psi_j(\mathbf{x}_+, \mathbf{v}_+) = \Delta_j(\mathbf{x}_+) + r_j(\mathbf{v}_+) \le \max\{\Delta_j(\mathbf{x}_+), \Delta_j(\mathbf{v}_+)\} + \max\{r_j(\mathbf{x}_+), r_j(\mathbf{v}_+)\} \\
\le &\kappa_{\text{glob}}^2 \left(\max\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\} + \max\{r_j(\mathbf{x}), r_j(\mathbf{v})\}\right) \qquad \text{(by the first two inequalities)} \\
\le &\kappa_{\text{glob}}^2 \left(\Delta_j(\mathbf{x}) + \Delta_j(\mathbf{v}) + r_j(\mathbf{x}) + r_j(\mathbf{v})\right) \\
\le &\kappa_{\text{glob}}^2 (\kappa_j + 1)(\Delta_j(\mathbf{x}) + r_j(\mathbf{v})) \le 2\kappa_{\text{glob}}^2 \kappa_j \psi_j(\mathbf{x}, \mathbf{v}).
\end{aligned}$$

■

C.1.2. ESTIMATING THE PROGRESS OF AcBSLS

Lemma 18 *Under the same settings of Theorem 15, for any* $(\mathbf{x}, \mathbf{v})$ *and* $i \in [m]$*, let* $(\mathbf{x}_+, \mathbf{v}_+) \leftarrow \texttt{AcBSLS}_i(\mathbf{x}, \mathbf{v})$*, then for any* $j < i$*, it is the case that*

(a) $\max\{\Delta_j(\mathbf{x}_+), \Delta_j(\mathbf{v}_+)\} \le \max\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\}$.

(b) $\max\{r_j(\mathbf{x}_+), r_j(\mathbf{v}_+)\} \le \max\{r_j(\mathbf{x}), r_j(\mathbf{v})\}$.

Proof [Proof of Lemma 18] We will fix j and prove both statements by induction on i in descent order (from m to $j+1$). Throughout the proof we denote $(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}, \tilde{\mathbf{x}}^{(0)}, \tilde{\mathbf{v}}^{(0)}, \cdots, \mathbf{x}^{(T_i)}, \mathbf{v}^{(T_i)})$ the sequence generated by running $\texttt{AcBSLS}_i(\mathbf{x}, \mathbf{v})$.

Induction base: for $i = m$, note that $\texttt{AcBSLS}_m(\cdot, \cdot)$ is equivalent to $\texttt{AGD}^{T_m}(\cdot, \cdot; L_m, \mu_m)$. Since $j < m$, Lemma 16(b) suggests $\max\{\Delta_j(\tilde{\mathbf{x}}^{(t)}), \Delta_j(\tilde{\mathbf{v}}^{(t)})\} \le \max\{\Delta_j(\mathbf{x}^{(t)}), \Delta_j(\mathbf{v}^{(t)})\}$. Since $i = m$ we have $\mathbf{x}^{(t+1)} = \tilde{\mathbf{x}}^{(t)}$ and $\mathbf{v}^{(t+1)} = \tilde{\mathbf{v}}^{(t)}$, and consequently $\max\{\Delta_j(\mathbf{x}^{(t+1)}), \Delta_j(\mathbf{v}^{(t+1)})\} \le \max\{\Delta_j(\mathbf{x}^{(t)}), \Delta_j(\mathbf{v}^{(t)})\}$. Telescoping t from 0 to T_i yields $\max\{\Delta_j(\mathbf{x}_+), \Delta_j(\mathbf{v}_+)\} \le \max\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\}$. The same arguments hold for (b) as well.

Now suppose the statements hold for $i+1 \le m$ and we study i. Since $j < i$ we can apply Lemma 16(b) to show that $\max\{\Delta_j(\tilde{\mathbf{x}}^{(t)}), \Delta_j(\tilde{\mathbf{v}}^{(t)})\} \le \max\{\Delta_j(\mathbf{x}^{(t)}), \Delta_j(\mathbf{v}^{(t)})\}$. By induction hypothesis we have $\Delta_j(\mathbf{x}^{(t+1)}) \le \Delta_j(\tilde{\mathbf{x}}^{(t)})$ and $\Delta_j(\mathbf{v}^{(t+1)}) \le \Delta_j(\tilde{\mathbf{v}}^{(t)})$. Consequently $\max\{\Delta_j(\mathbf{x}^{(t+1)}), \Delta_j(\mathbf{v}^{(t+1)})\} \le \max\{\Delta_j(\mathbf{x}^{(t)}), \Delta_j(\mathbf{v}^{(t)})\}$. Telescoping t from 0 to T_i yields $\max\{\Delta_j(\mathbf{x}_+), \Delta_j(\mathbf{v}_+)\} \le \max\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\}$. The same arguments hold for (b) as well. ■

Lemma 19 *Under the same settings of Theorem 15, for any* $(\mathbf{x}, \mathbf{v})$ *and* $i \in [m]$*, let* $(\mathbf{x}_+, \mathbf{v}_+) \leftarrow \texttt{AcBSLS}_i(\mathbf{x}, \mathbf{v})$*, then* $\psi_i(\mathbf{x}_+, \mathbf{v}_+) \le \left(1 - \frac{1}{\sqrt{\kappa_i}}\right)^{T_i} \psi_i(\mathbf{x}, \mathbf{v})$.

Proof [Proof of Lemma 19] Let $(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}, \tilde{\mathbf{x}}^{(0)}, \tilde{\mathbf{v}}^{(0)}, \cdots, \mathbf{x}^{(T_i)}, \mathbf{v}^{(T_i)})$ be the trajectory generated by running $\texttt{AcBSLS}_i(\mathbf{x}, \mathbf{v})$. For $i = m$, $\texttt{AcBSLS}_m(\cdot, \cdot)$ is equivalent to $\texttt{AGD}^{T_m}(\cdot, \cdot; L_m, \mu_m)$. Lemma 16(a) suggests that $\psi_i(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) \le (1 - \frac{1}{\sqrt{\kappa_m}})\psi_i(\mathbf{x}^{(t)}, \mathbf{v}^{(t)})$. Telescoping t from 0 to T_m shows $\psi_i(\mathbf{x}_+, \mathbf{v}_+) \le (1 - \frac{1}{\sqrt{\kappa_m}})^{T_m} \psi_i(\mathbf{x}, \mathbf{v})$.

For $i < m$, we first note that Lemma 16(a) suggests $\psi_i(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \le (1 - \frac{1}{\sqrt{\kappa_m}})\psi_i(\mathbf{x}^{(t)}, \mathbf{v}^{(t)})$. Since $(\mathbf{x}^{(t+1)}, \_) = \texttt{AcBSLS}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})$, Lemma 18 suggests that $\Delta_i(\mathbf{x}^{(t+1)}) \le \Delta_i(\tilde{\mathbf{x}}^{(t)})$. For

the same reason we have $r_i(\mathbf{v}^{(t+1)}) \le r_i(\tilde{\mathbf{v}}^{(t)})$. Consequently $\psi_i(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) \le \psi_i(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \le (1 - \frac{1}{\sqrt{\kappa_m}})\psi_i(\mathbf{x}^{(t)}, \mathbf{v}^{(t)})$. Telescoping t from 0 to T_i completes the proof.

■

Lemma 20 *Under the same settings of Theorem 15, for any* $(\mathbf{x}, \mathbf{v})$ *and* $i \in [m]$*, let* $(\mathbf{x}_+, \mathbf{v}_+) \leftarrow \texttt{AcBSLS}_i(\mathbf{x}, \mathbf{v})$*, then for any* $j \ge i$*, the following inequality holds*

$$\psi_j(\mathbf{x}_+, \mathbf{v}_+) \le \exp\left(-\frac{1}{\sqrt{\kappa_j}}\prod_{k=i}^{j} T_k + \left(\sum_{k=i}^{j-1}\prod_{l=i}^{k} T_l\right)\log\left(4\kappa_j^2\kappa_{\text{glob}}^2\right)\right)\psi_j(\mathbf{x}, \mathbf{v}).$$

Proof [Proof of Lemma 20] We will fix j and prove by induction on i in descent order (from j to 1).

Induction base: for $i = j$, the statement (b) follows by Lemma 19

$$\psi_j(\mathbf{x}_+, \mathbf{v}_+) \le \left(1 - \frac{1}{\sqrt{\kappa_j}}\right)^{T_j} \le \exp\left(-\frac{T_j}{\sqrt{\kappa_j}}\right)\psi_j(\mathbf{x}, \mathbf{v}).$$

Now assume the claim holds for $i + 1 \le j$, and we study the case of i. Denote $(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}, \tilde{\mathbf{x}}^{(0)}, \tilde{\mathbf{v}}^{(0)}, \cdots, \mathbf{x}^{(T_i)}, \mathbf{v}^{(T_i)})$ the sequence generated by running $\texttt{AcBSLS}_i(\mathbf{x}, \mathbf{v})$. Since $(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \leftarrow \texttt{AGD}(\mathbf{x}^{(t)}, \mathbf{v}^{(t)}; L_i, \mu_i)$ and $i < j$, Lemma 16(c) suggests that

$$\max\left\{r_j(\tilde{\mathbf{x}}^{(t)}), r_j(\tilde{\mathbf{v}}^{(t)})\right\} \le \kappa_{\text{glob}}^2 \max\left\{r_j(\mathbf{x}^{(t)}), r_j(\mathbf{v}^{(t)})\right\}.$$

Since f_j is κ_j-conditioned we have $r_j \le \Delta_j \le \kappa_j r_j$, which implies

$$\psi_j(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) = \Delta_j(\tilde{\mathbf{x}}^{(t)}) + r_j(\tilde{\mathbf{v}}^{(t)}) \le \kappa_j r_j(\tilde{\mathbf{x}}^{(t)}) + r_j(\tilde{\mathbf{v}}^{(t)}) \le 2\kappa_j \max\left\{r_j(\mathbf{x}^{(t)}), r_j(\mathbf{v}^{(t)})\right\},$$

and

$$\max\left\{r_j(\mathbf{x}^{(t)}), r_j(\mathbf{v}^{(t)})\right\} \le r_j(\mathbf{x}^{(t)}) + r_j(\mathbf{v}^{(t)}) \le \Delta_j(\mathbf{x}^{(t)}) + r_j(\mathbf{v}^{(t)}) = \psi_j(\mathbf{x}^{(t)}, \mathbf{v}^{(t)}).$$

In summary we have

$$\psi_j(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \le 2\kappa_j\kappa_{\text{glob}}^2\psi_j(\mathbf{x}^{(t)}, \mathbf{v}^{(t)}) \tag{C.7}$$

Since $(\mathbf{x}^{(t+1)}, \_) \leftarrow \texttt{AcBSLS}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})$, by induction hypothesis of (a) we have

$$\Delta_j(\mathbf{x}^{(t+1)}) \le \psi_j(\mathbf{x}^{(t+1)}, \_) \le \exp\left(-\frac{1}{\sqrt{\kappa_j}}\prod_{k=i+1}^{j} T_k + \left(\sum_{k=i+1}^{j-1}\prod_{l=i+1}^{k} T_l\right)\log(4\kappa_j^2\kappa_{\text{glob}}^2)\right)\psi_j(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)}). \tag{C.8}$$

For the same reason we have

$$r_j(\mathbf{v}^{(t+1)}) \le \psi_j(\_, \mathbf{v}^{(t+1)}) \le \exp\left(-\frac{1}{\sqrt{\kappa_j}}\prod_{k=i+1}^{j} T_k + \left(\sum_{k=i+1}^{j-1}\prod_{l=i+1}^{k} T_l\right)\log(4\kappa_j^2\kappa_{\text{glob}}^2)\right)\psi_j(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)}). \tag{C.9}$$

Since $\psi_j(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)}) = \Delta_j(\tilde{\mathbf{x}}^{(t)}) + r_j(\tilde{\mathbf{x}}^{(t)}) \le 2\Delta_j(\tilde{\mathbf{x}}^{(t)})$ and $\psi_j(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) = \Delta_j(\tilde{\mathbf{v}}^{(t)}) + r_j(\tilde{\mathbf{v}}^{(t)}) \le (\kappa_j + 1)r_j(\tilde{\mathbf{v}}^{(t)})$ we have

$$\psi_j(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)}) + \psi_j(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \le 2\kappa_j\psi_j(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \tag{C.10}$$

Combining (C.8) (C.9) and (C.10) gives

$$\begin{aligned}&\psi_j(\mathbf{x}^{(t+1)},\mathbf{v}^{(t+1)}) = \Delta_j(\mathbf{x}^{(t+1)}) + r_j(\mathbf{v}^{(t+1)})\\ \leq& 2\kappa_j \cdot \exp\left(-\frac{1}{\sqrt{\kappa_j}}\prod_{k=i+1}^{j} T_k + \left(\sum_{k=i+1}^{j-1}\prod_{l=i+1}^{k} T_l\right)\log(4\kappa_j^2\kappa_{\text{glob}}^2)\right)\psi_j(\tilde{\mathbf{x}}^{(t)},\tilde{\mathbf{v}}^{(t)}). \qquad \text{(C.11)}\end{aligned}$$

By (C.7) and (C.11) we arrive at

$$\psi_j(\mathbf{x}^{(t+1)},\mathbf{v}^{(t+1)}) \leq 4\kappa_j^2\kappa_{\text{glob}}^2 \cdot \exp\left(-\frac{1}{\sqrt{\kappa_j}}\prod_{k=i+1}^{j} T_k + \left(\sum_{k=i+1}^{j-1}\prod_{l=i+1}^{k} T_l\right)\log(4\kappa_j^2\kappa_{\text{glob}}^2)\right)\psi_j(\mathbf{x}^{(t)},\mathbf{v}^{(t)}).$$

Telescoping t from 0 to T_i yields

$$\psi_j(\mathbf{x}^{(T_i)},\mathbf{v}^{(T_i)}) \leq \exp\left(-\frac{1}{\sqrt{\kappa_j}}\prod_{k=i}^{j} T_k + \left(\sum_{k=i}^{j-1}\prod_{l=i}^{k} T_l\right)\log(4\kappa_j^2\kappa_{\text{glob}}^2)\right)\psi_j(\mathbf{x}^{(0)},\mathbf{v}^{(0)}),$$

completing the induction proof of Lemma 20. ■

C.1.3. FINISHING THE PROOF OF THEOREM 15

With Lemma 20 at hands we are ready to finish the proof of Theorem 15. This part of proof is almost identical to the proof of Theorem 6 presented in Appendix B.1.

Proof [Proof of Theorem 15] Applying Lemma 20 yields (for any $i \in [m]$)

$$\psi_i(\texttt{AcBSLS}_1(\mathbf{x}^{(0)},\mathbf{v}^{(0)})) \leq \exp\left(\underbrace{-\frac{1}{\sqrt{\kappa_i}}\prod_{k=1}^{i} T_k + \left(\sum_{k=1}^{i-1}\prod_{l=1}^{k} T_l\right)\log\left(4\kappa_{\text{glob}}^4\right)}_{\text{denoted as } \gamma_i}\right)\psi_i(\mathbf{x}^{(0)},\mathbf{v}^{(0)}),$$

Observe that for any $i = 2,\ldots,m$,

$$\begin{aligned}\gamma_i - \gamma_{i-1} =& \log(4\kappa_{\text{glob}}^4)\cdot\prod_{j=1}^{i-1} T_j - \kappa_i^{-\frac{1}{2}}\prod_{j=1}^{i} T_j + \kappa_{i-1}^{-\frac{1}{2}}\prod_{j=1}^{i-1} T_j\\ =& \prod_{j=1}^{i-1} T_j \cdot \left(-\kappa_i^{-\frac{1}{2}} T_i + \kappa_{i-1}^{-\frac{1}{2}} + \log(4\kappa_{\text{glob}}^4)\right).\end{aligned}$$

Since $T_i \geq \sqrt{\kappa_i}(\log(4\kappa_{\text{glob}}^4)+1)$ we have

$$\gamma_i - \gamma_{i-1} \leq \prod_{j=1}^{i-1} T_j \cdot \left(-1 + \kappa_{i-1}^{-\frac{1}{2}}\right) \leq 0.$$

For γ_1 we observe that $\gamma_1 = -\frac{1}{\sqrt{\kappa_i}}T_1 \leq \log\frac{\epsilon}{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}$. Hence $\gamma_m \leq \gamma_{m-1} \leq \cdots \leq \gamma_1 \leq \log\frac{\epsilon}{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}$. Therefore for all $i \in [m]$ it is the case that

$$\psi_i(\mathbf{x},\mathbf{v}) \leq \exp(\gamma_i)\cdot\psi_i(\mathbf{x}^{(0)},\mathbf{v}^{(0)}) \leq \frac{\epsilon}{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}\psi_i(\mathbf{x}^{(0)},\mathbf{v}^{(0)})$$

Taking summation over i gives

$$\psi(\texttt{AcBSLS}_1(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})) \le \sum_{i\in[m]} \frac{\epsilon}{\psi(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})} \psi_i(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}) = \epsilon,$$

completing the proof of Theorem 15. ■

C.2. Stability of AcBSLS

Similar to (un-accelerated) BSLS, under finite-precision arithmetic, AcBSLS can also attain the same rate of convergence with only logarithmic bits of precision.

Formally, let $\widehat{\texttt{AGD}}$ be the finite-precision implementation of AGD, and $\widehat{\texttt{AcBSLS}}_i$ be the finite-precision implementation of AcBSLS by replacing AGD with $\widehat{\texttt{AGD}}$. We impose the following requirement such that $\widehat{\texttt{AGD}}$ can return a δ-multiplicative approximation of AGD in both $\mathbf{x}$ and $\mathbf{v}$:

Requirement 2 *There exists a $\delta < 1$ such that for any $\mathbf{x}$, $\mathbf{v}$, and i, considering $(\mathbf{x}_+, \mathbf{v}_+) \leftarrow \texttt{AGD}(\mathbf{x}, \mathbf{v}; L_i, \mu_i)$ and $(\widehat{\mathbf{x}_+}, \widehat{\mathbf{v}_+}) \leftarrow \widehat{\texttt{AGD}}(\mathbf{x}, \mathbf{v}; L_i, \mu_i)$, it is the case that $|\widehat{\mathbf{x}_+} - \mathbf{x}_+| \le \delta |\mathbf{x}_+|$ and $|\widehat{\mathbf{v}_+} - \mathbf{v}_+| \le \delta |\mathbf{v}_+|$. (We use $|\cdot|$ to denote element-wise absolute values).*

We specialized the initializations to $\mathbf{0}$ to simplify the exposition of the theorem. In Appendix H, we provide and prove the general version with arbitrary $\mathbf{x}^{(0)}, \mathbf{v}^{(0)}$.

Theorem 21 (AcBSLS under finite-precision arithmetic) *Consider multiscale optimization problem defined in Theorem 1, for any $\epsilon > 0$, assuming Requirement 2 with*

$$\delta^{-1} \ge 4m \left(\prod_{i\in[m]} T_i\right) \cdot (10\kappa_{\text{glob}}^2)^{2m-1} \cdot \frac{\psi(\mathbf{0}, \mathbf{0})}{\epsilon},$$

then $\min\{\psi(\mathbf{0}, \mathbf{0}), \psi(\widehat{\texttt{AcBSLS}}_1(\mathbf{0}, \mathbf{0}))\} \le 3\epsilon$ provided that $T_1, \ldots, T_m$ satisfy Eq. (C.1) (with $\mathbf{x}^{(0)} = \mathbf{v}^{(0)} = \mathbf{0}$), when $\{(\mu_i, L_i), i \in [m]\}$ are known. We can also achieve the same asymptotic sample complexity (up to constant factors suppressed in the $\mathcal{O}(\cdot)$) when $\{(\mu_i, L_i), i \in [m]\}$ are unknown and only m, μ_1, L_m and $\pi_\kappa = \prod_{i=1}^m \kappa_i$ are known.

The proof of Theorem 21 is deferred to Appendix H.

C.3. Why do we need branching for AcBSLS

In this subsection we demonstrate why naive AcBSLS may not converge. Specifically, we consider the following Algorithm 4. The only difference compared with the principled AcBSLS (defined in Algorithm 2) is the replacement of the branching procedure with a naive recursion.

C.3.1. THEORETICAL EVIDENCE

Following the discussion after Lemma 16, we provide a simple result suggesting the potential for AGD may not be "backward compatible" (specifically, the potential governing small $[\mu_i, L_i]$ may not be conservative under AGD with larger $[\mu, L]$, although the latter takes smaller step.) Therefore one cannot replace Lemma 16 with Eq. (C.2). Formally, we prove the following proposition.

Algorithm 4 Naive Accelerated Big-Step Little-Step Algorithm (may not converge)

Procedure $\texttt{NaiveAcBSLS}_i\ (\mathbf{x}^{(0)}, \mathbf{v}^{(0)})$
1: **for** $t = 0, 1, \ldots, T_i - 1$ **do**
2: $\quad(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \leftarrow \texttt{AGD}(\mathbf{x}_t, \mathbf{v}_t; L_i, \mu_i)$ ▷ AGD is the same as the original one (Algorithm 2)
3: $\quad$**if** $i < m$ **then**
4: $\quad\quad(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) \leftarrow \texttt{AcBSLS}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)})$ ▷ Naively recurse instead of branching
5: $\quad$**else**
6: $\quad\quad(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) \leftarrow (\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)})$
7: **return** $(\mathbf{x}^{(T_i)}, \mathbf{v}^{(T_i)})$

Proposition 22 *There exists a function $f : \mathbb{R}^d \to \mathbb{R}$ that is μ_1-strongly-convex and L_1-smooth, but for certain $(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})$ it is the case that*

$$\psi_1(\texttt{AGD}(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}; L_2, \mu_2)) > \psi_1(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}).$$

for some μ_2, L_2 such that $L_2 > \mu_2 > L_1$. Here $\psi_1(\mathbf{x}, \mathbf{v})$ is the potential associated with f, namely

$$\psi_1(\mathbf{x}, \mathbf{v}) \coloneqq f(\mathbf{x}) - f^\star + \frac{\mu_1}{2}\|\mathbf{v} - \mathbf{x}^\star\|_2^2$$

Remark 23 *Although Proposition 22 does not rule out the possibility of other conservative potentials, we conjecture that such a potential may not exist given the inherent instability of accelerated GD (c.f., Section F of Yuan and Ma (2020)).*

Proof [Proof of Proposition 22] Consider the following objective $f : \mathbb{R}^2 \to \mathbb{R}$: $f(\mathbf{x}) = 0.5x_1^2 + 5x_2^2$. Apparently f is 1-strongly-convex, 10-smooth. Consider initialization $\mathbf{x}^{(0)} = (0, 0)^\top, \mathbf{v}^{(0)} = (1, 1)^\top$. Then one can verify that $\psi_1(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}) = 1$ but $\psi_1(\texttt{AGD}(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}; L_2, \mu_2)) > 1.18$ for $L_2 = 200$ and $\mu_2 = 100$. ■

C.3.2. NUMERICAL EVIDENCE

Next, we provide numerical evidence against the convergence of naive AcBSLS, see Fig. 4. We synthesize a quadratic objective with eigenvalues belonging to $[10^{-4}, 10^{-3}] \cup [0.5, 1]$. We implement both the principled AcBSLS (with branching, see Algorithm 2) and naive AcBSLS (Algorithm 4) with the corresponding μ_1, μ_2, L_1, L_2. We observe that the principled AcBSLS (with branching) converge with $T_2 = 8$, as expected. On the other hand, the naive AcBSLS fails to converge for any $T_2 \in \{8, 16, 32, 64\}$. The implementation details can be found in the accompanying notebook in supplementary materials.

Appendix D. Lower bound for multiscale optimization

In this section, we prove our lower bound results (Theorem 10) of the multi-scale optimization problem.

D.1. Proof structure of Theorem 10

We will separate the proof of Theorem 10 into three parts.

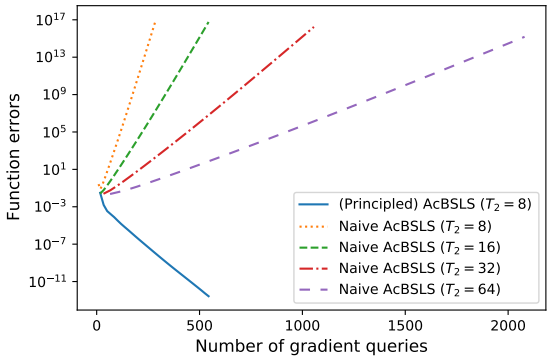


Figure 4: **Numerical evidence that Naive AcBSLS (Algorithm 4) may not converge**. Observe that the principled AcBSLS (with branching) converge with $T_2 = 8$, but the naive AcBSLS fails to converge for any $T_2 \in \{8, 16, 32, 64\}$.

Part I: Reduction to uniform polynomial approximation. In the first part, we reduce the problem of a lower bound over arbitrary first-order deterministic algorithms to a constrained polynomial uniform approximation problem on $S = \bigcup_{i\in[m]}[\mu_i, L_i]$ across $\mathcal{P}_k^0$, where (throughout this section)

$$\mathcal{P}_k^0 = \{p : p \text{ is a polynomial of degree at most } k \text{ and } p(0) = 1\}. \tag{D.1}$$

The result is as follows.

Lemma 24 (Reduction to a uniform polynomial approximation problem) *For any first-order deterministic algorithm* A*, for any $k \in \mathbb{N}$ and $\Delta_{\text{grad}} > 0$, there exists an objective f satisfying Theorem 1 with $\|\nabla f(\mathbf{0})\|_2 \le \Delta_{\text{grad}}$ such that*

$$\min_{\tau\in[k]} \|\nabla f(\mathbf{x}^{(\tau)})\|_2 \ge \left(\min_{p\in\mathcal{P}_k^0} \max_{\lambda\in S} |p(\lambda)|\right) \cdot \Delta_{\text{grad}}.$$

The rough proof idea of Lemma 24 is to 1) first reduce the general first-order deterministic algorithm class to the construction of a tri-diagonal objective for which zero-respecting algorithm (see Carmon et al. (2021) for definition) is hard , then 2) reduce to the problem of discrete weighted ℓ_2 polynomial approximation over S, and finally 3) reduce to uniform polynomial approximation over S. The detailed proof of Lemma 24 is relegated to Appendix D.2.

Part II: Reduction to Green's function. In the second part, we cite classic results from potential theory literature to reduce the uniform polynomial approximation problem raised in Lemma 24 to the estimation of Green's function. The results are as follows.

Lemma 25 (Reduction to Green's function) *Let $S = \bigcup_{i\in[m]}[\mu_i, L_i]$, then for any $k \in \mathbb{N}$, the following inequality holds*

$$\min_{p\in\mathcal{P}_k^0} \max_{\lambda\in S} |p(\lambda)| \ge \exp(-k g_S(0))$$

where $g_S(0)$ is the Green's function associated with S (with pole at ∞), see Theorem 37 in Appendix D.3 for formal definition.

The detailed reference of Lemma 25 is relegated to Appendix D.3.

Part III: Estimating (upper bound) the Green's function. In the last part, we provide a bound of $g_S(0)$ as follows. We identify that this estimate may be of independent interest.

Lemma 26 (Estimating the Green's function) *Let $S = \bigcup_{j=1}^{m}[\mu_j, L_j]$, and assume $\frac{L_j}{\mu_j} \geq 2$ for $j \in [m]$. Then the Green's function associated with S satisfies*

$$g_S(0) \leq \frac{8}{\sqrt{\prod_{i\in[m]} \frac{L_i}{\mu_i}} \cdot \prod_{i\in[m-1]} \left(0.03 \cdot \log(16\frac{\mu_{i+1}}{L_i})\right)}.$$

The proof of Lemma 26 is relegated to Appendix D.4.

The Theorem 10 then follows immediately from Lemmas 24, 25 and 26.

D.2. Proof of Lemma 24: Reduction to uniform polynomial approximation

In this subsection we will prove Lemma 24 on the reduction from the lower bound of arbitrary first-order deterministic algorithms to the uniform polynomial approximation problem.

We will prove Lemma 24 in three steps.

Step 1: Reduction to a first-order zero chain (or hard tri-diagonal quadratic objective). Following the techniques of Carmon et al. (2021), we first reduce the lower bound across all first-order deterministic algorithms to the construction of a "first-order zero chain" Carmon et al. (2021). Specifically, we reduce to the existence of tri-diagonal quadratic objectives with "large" gradients under limited supports.

Lemma 27 (Reduction from arbitrary first-order deterministic algorithms to first-order zero-chains) *Let $S = \bigcup_{i\in[m]}[\mu_i, L_i]$, suppose for some $\Delta_{\text{grad}} > 0$, $\epsilon > 0$ and $k \in \mathbb{N}$, there exists a symmetric tri-diagonal matrix $\mathbf{T} \in \mathbb{R}^{(k+m)\times(k+m)}$ with eigenvalues all belonging to S, and suppose the objective $f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{T}\mathbf{x} + \Delta_{\text{grad}} \cdot \mathbf{e}_1^\top \mathbf{x}$ satisfies*

$$\min_{\mathbf{x}:x_{k+1}=x_{k+2}=\cdots=x_{k+m}=0} \|\nabla f(\mathbf{x})\|_2 \geq \epsilon.$$

Then for any first-order deterministic algorithm A, there exists a function $\tilde{f}$ satisfying Theorem 1 with $\|\nabla f(\mathbf{0})\|_2 \leq \Delta_{\text{grad}}$ such that the trajectory $\{\mathbf{x}^{(t)}\}_{t=1}^{\infty}$ generated by A on $\tilde{f}$ satisfies

$$\min_{\tau\in[k+1]} \|\nabla \tilde{f}(\mathbf{x}^{(\tau)})\|_2 \geq \epsilon.$$

The proof of Lemma 27 is similar to the original proof of lower bounds in Carmon et al. (2021). We first reduce the range of arbitrary deterministic first-order algorithms to zero-respecting algorithms via the equivalency result in Carmon et al. (2021), and then show that any zero-respecting algorithm can only reveal one dimension per step for the tri-diagonal quadratic objective. The detailed proof of Lemma 27 is relegated to Appendix D.2.1.

Step 2: Reduction to discrete weighted ℓ_2 polynomial approximation. Next, we reduce the problem raised in Lemma 27 to the following constrained weighted discrete ℓ_2 polynomial approximation problem.

Lemma 28 (Reduction to discrete weighted ℓ_2 polynomial approximation) *Let $S = \bigcup_{i\in[m]}[\mu_i, L_i]$, then for any $k \in \mathbb{N}$, for any $\Delta_{\text{grad}} > 0$, there exists a symmetric tri-diagonal matrix $\mathbf{T} \in \mathbb{R}^{(k+m)\times(k+m)}$ with eigenvalues all belonging to S such that the objective $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{T}\mathbf{x} + \Delta_{\text{grad}} \cdot \mathbf{e}_1^\top \mathbf{x}$ satisfies*

$$\min_{\mathbf{x}:x_{k+1}=\cdots=x_{k+m}=0} \|\nabla f(\mathbf{x})\|_2 \geq \Delta_{\text{grad}} \cdot \max_{\lambda_1,\ldots,\lambda_{k+m}\in S} \max_{\sum_{i\in[k+m]} v_i^2=1} \min_{p\in\mathcal{P}_k^0} \sqrt{\sum_{i\in[k+m]} (p(\lambda_i)v_i)^2}.$$

Recall $\mathcal{P}_k^0$ is defined in Eq. (D.1) as the set of polynomials p of degree at most k with $p(0) = 1$.

The proof of Lemma 28 is constructive, where we explicitly construct a symmetric tri-diagonal matrix $\mathbf{T}$ with large $\min_{\mathbf{x}:x_{k+1}=\cdots=x_{k+m}=0} \|\nabla f(\mathbf{x})\|_2$. The detailed proof is relegated to Appendix D.2.2.

Step 3: Reduction to uniform polynomial approximation on S. Finally, we reduce the discrete weighted ℓ_2-approximation problem raised in Lemma 28 to a uniform polynomial approximation over S across $\mathcal{P}_k^0$.

Lemma 29 (Reduction to uniform polynomial approximation on S) *Let $S = \bigcup_{i\in[m]}[\mu_i, L_i]$, then for any $k \in \mathbb{N}$, the following inequality holds*

$$\max_{\lambda_1,\ldots,\lambda_{k+m}\in S} \max_{\sum_{i\in[k+m]} v_i^2=1} \min_{p\in\mathcal{P}_k^0} \sqrt{\sum_{i\in[k+m]} (p(\lambda_i)v_i)^2} \geq \min_{p\in\mathcal{P}_k^0} \max_{\lambda\in S} |p(\lambda)|.$$

Recall $\mathcal{P}_k^0$ is defined in Eq. (D.1) as the set of polynomials p of degree at most k with $p(0) = 1$.

The proof of Lemma 29 is based on the fact that the best uniform approximation over S, denoted as $p_k^\star$, is also the best discrete weighted ℓ_2 approximation over the extreme points on $p_k^\star$ with appropriate weights. To this end, we will show that $p_k^\star$ is orthogonal to low-degree polynomials under these weights. The detailed proof of Lemma 29 is relegated to Appendix D.2.3.

The proof of Lemma 24 then follows immediately from Lemmas 27, 28 and 29.

D.2.1. DEFERRED PROOF OF LEMMA 27

Proof [Proof of Lemma 27] Since $\mathbf{T}$ has all its eigenvalues within $S = \bigcup_{i\in[m]}[\mu_i, L_i]$, the objective $f(\mathbf{x})$ satisfies Theorem 1. Since $\mathbf{T}$ is tri-diagonal, for any zero-respecting first-order algorithm A_{zr} (see Carmon et al. (2021) for definition) initialization at $\mathbf{0}$, the first $k+1$ iterates $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k+1)}$ are all supported in the first k coordinates. Hence

$$\min_{\tau\in[k+1]} \|\nabla f(\mathbf{x}^{(\tau)})\|_2 \geq \min_{\mathbf{x}:x_{k+1}=x_{k+2}=\cdots=x_{k+m}=0} \|\nabla f(\mathbf{x})\|_2 \geq \epsilon.$$

Let $\mathcal{F}(\Delta_{\text{grad}})$ denote the union, over $d \in \mathbb{N}$, of the collections of $\mathcal{C}^\infty$ convex functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying Theorem 1 and $\|\nabla f(\mathbf{0})\|_2 \leq \Delta_{\text{grad}}$. Since $\mathcal{F}(\Delta_{\text{grad}})$ is orthogonally invariant, by Proposition 1 of Carmon et al. (2021), the time complexities over all first-order deterministic algorithms are lower bounded by the zero respecting first-order algorithms, completing the proof. ■

D.2.2. DEFERRED PROOF OF LEMMA 28

We introduce the following definition for ease of exposition.

Definition 30 *A symmetric tri-diagonal matrix is **non-degenerate** if none of its sub-diagonal entries are zero.*

We first show that for any distinct $\{\lambda_i\}_{i\in[k+m]}$ and positive $\{v_i\}_{i\in[k+m]}$, one can construct a desired tri-diagonal matrix.

Lemma 31 *Let $\lambda_1, \lambda_2, \ldots, \lambda_{k+m}$ be a set of distinct positive numbers, and $v_1, v_2, \ldots, v_{k+m}$ be another set of positive numbers with $\sum_{i\in[k+m]} v_i^2 = 1$. Then there exists an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{(k+m)\times(k+m)}$ such that*

1. *$\mathbf{Q}\mathbf{e}_1 = \mathbf{v}$, where $\mathbf{v} := [v_1, v_2, \ldots, v_{k+m}]^\top$.*
2. *$\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q}$ is non-degenerate tri-diagonal, where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{k+m})$.*

Proof [Proof of Lemma 31] We construct $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{k+m}]$ column by column as follows:

(a) $\mathbf{q}_1 = \mathbf{v}$.

(b) For any $j = 2, \ldots, k+m$, let $\tilde{\mathbf{q}}_j = (\mathbf{I} - \sum_{i\in[j-1]} \mathbf{q}_i\mathbf{q}_i^\top)\mathbf{\Lambda}\mathbf{q}_{j-1}$, $\mathbf{q}_j = \frac{\tilde{\mathbf{q}}_j}{\|\tilde{\mathbf{q}}_j\|_2}$.

One can verify that

(i) For any $i \in [k+m]$, $\|\mathbf{q}_i\|_2 = 1$.

(ii) For any $i < j$, $\mathbf{q}_i^\top \tilde{\mathbf{q}}_j = 0$ and thus $\mathbf{q}_i^\top \mathbf{q}_j = 0$.

(iii) For any $i < j-1$, $\mathbf{q}_i^\top \mathbf{\Lambda} \mathbf{q}_j = 0$ (To see this, first observe that $\operatorname{span}\langle \mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_j\rangle = \operatorname{span}\langle \mathbf{v}, \mathbf{\Lambda}\mathbf{v}, \ldots, \mathbf{\Lambda}^{j-1}\mathbf{v}\rangle$ for any j. Thus $\mathbf{\Lambda}\mathbf{q}_i \in \operatorname{span}\langle \mathbf{q}_1, \ldots, \mathbf{q}_{i+1}\rangle$. Consequently we have $\mathbf{q}_i^\top \mathbf{\Lambda} \mathbf{q}_j = 0$ by point (ii) above for any $i < j-1$).

By (i) and (ii) we know $\mathbf{Q}$ is orthogonal. By (iii) we know $\mathbf{Q}^\top\mathbf{\Lambda}\mathbf{Q}$ is tri-diagonal. The non-degeneracy of $\mathbf{T}$ follows by the linear independence of $\{\mathbf{v}, \mathbf{\Lambda}\mathbf{v}, \ldots, \mathbf{\Lambda}^{k+m-1}\}$ since $\mathbf{v} > \mathbf{0}$, $\mathbf{\Lambda} > \mathbf{0}$ and the distinctness of $\{\lambda_i\}_{i\in[k+m]}$. ■

Next, following Lemma 31, we show that the tri-diagonal objective has large $\|\nabla f(\mathbf{x})\|_2$ when the last m coordinates of $\mathbf{x}$ are zero.

Lemma 32 *Under the same settings and notation of Lemma 31, let $\mathbf{T} = \mathbf{Q}^\top\mathbf{\Lambda}\mathbf{Q}$, and consider objective $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{T}\mathbf{x} + \Delta_{\text{grad}} \cdot \mathbf{e}_1^\top\mathbf{x}$. Then*

$$\min_{\mathbf{x}:x_{k+1}=\cdots=x_{k+m}=0} \|\nabla f(\mathbf{x})\|_2 = \Delta_{\text{grad}} \cdot \min_{p\in\mathcal{P}_k^0} \|p(\mathbf{\Lambda})\mathbf{Q}\mathbf{e}_1\|_2 = \Delta_{\text{grad}} \cdot \min_{p\in\mathcal{P}_k^0} \sqrt{\sum_{i\in[k+m]} (p(\lambda_i)v_i)^2}.$$

Proof [Proof of Lemma 32] By non-degeneracy of $\mathbf{T}$ we have

$$\{\mathbf{x} : x_{k+1} = \cdots = x_{k+m} = 0\} = \{p(\mathbf{T})\mathbf{e}_1 : \deg p \le k-1\}.$$

Thus the following sets are identical

$$\{\nabla f(\mathbf{x}) : x_{k+1} = \cdots = x_{k+m} = 0\} = \{\mathbf{T}\mathbf{x} + \Delta_{\text{grad}} \cdot \mathbf{e}_1 : x_{k+1} = \cdots = x_{k+m} = 0\}$$
$$= \{\Delta_{\text{grad}} \cdot p(\mathbf{T})\mathbf{e}_1 : p \in \mathcal{P}_k^0\}.$$

It follows that

$$\min_{\mathbf{x}: x_{k+1}=\cdots=x_{k+m}=0} \|\nabla f(\mathbf{x})\|_2 = \Delta_{\text{grad}} \cdot \min_{p\in\mathcal{P}_k^0} \|p(\mathbf{T})\mathbf{e}_1\|_2 = \Delta_{\text{grad}} \cdot \min_{p\in\mathcal{P}_k^0} \|p(\mathbf{\Lambda})\mathbf{Q}\mathbf{e}_1\|_2.$$

The last equality is due to $\mathbf{Q}\mathbf{e}_1 = \mathbf{v}$. ■

Now we finish the proof of Lemma 28.

Proof [Proof of Lemma 28] By Lemmas 31,32, for any distinct $\lambda_1, \ldots, \lambda_{k+m} \in S$ and $v_1, \ldots, v_{k+m} > 0$ such that $\sum_{i\in[k]} v_i^2 = 1$, we have

$$\min_{\mathbf{x}: x_{k+1}=\cdots=x_{k+m}=0} \|\nabla f(\mathbf{x})\|_2 \geq \Delta_{\text{grad}} \cdot \min_{p\in\mathcal{P}_k^0} \sqrt{\sum_{i\in[k+m]} (p(\lambda_i)v_i)^2}.$$

If $\lambda_1, \ldots, \lambda_{k+m}$ are not distinct, then one can find another set of distinct λ_i's such that the RHS is not smaller. The same arguments hold if one of the v_i is zero. Hence

$$\min_{\mathbf{x}: x_{k+1}=\cdots=x_{k+m}=0} \|\nabla f(\mathbf{x})\|_2 \geq \Delta_{\text{grad}} \cdot \max_{\lambda_1,\ldots,\lambda_{k+m}\in S} \max_{\sum_{i\in[k+m]} v_i^2=1} \min_{p\in\mathcal{P}_k^0} \sqrt{\sum_{i\in[k+m]} (p(\lambda_i)v_i)^2}.$$

■

D.2.3. DEFERRED PROOF OF LEMMA 29

We first cite the following Lemma 33 that characterizes the uniform approximation on S within $\mathcal{P}_k^0$. Recall $\mathcal{P}_k^0$ is defined as the set of polynomials p of degree at most k with $p(0) = 1$.

Lemma 33 (Characterization of uniform approximation on S, adapted from Schiefermayr and Peherstorfer (199
Let $S = \bigcup_{i\in[m]}[\mu_i, L_i]$, denote $I_i = [\mu_i, L_i]$ then for any $k \in \mathbb{N}$,

(a) The best uniform approximation $\min_{p\in\mathcal{P}_k^0} \max_{x\in S} |p(x)|$ is attained, denoted as $p_k^\star$. Denote $\|p_k^\star\|_S := \max_{\lambda\in S} |p_k^\star(\lambda)|$ hereinafter.

(b) $|p_k^\star|$ attains $\|p_k^\star\|_S$ in S for $s \in \{k+1, \ldots, k+m\}$ times, denoted as $\lambda_1 < \lambda_2 < \ldots < \lambda_s$. (That is to say $|p_k^\star(\lambda_1)| = |p_k^\star(\lambda_2)| = \cdots = |p_k^\star(\lambda_s)|$ and $\lambda_i \in S$ for $i \in [s]$). The λ_i's are called "e-points" in the literature.

(c) $p_k^\star(\lambda_1), \ldots, p_k^\star(\lambda_s)$ change signs for exactly k times, namely

$$|\{j : \mathbf{sgn}(p_k^\star(\lambda_j)) \neq \mathbf{sgn}(p_k^\star(\lambda_{j+1}))\}| = k$$

(d) If λ_j and λ_{j+1} belong to the same interval I_i, then $\mathbf{sgn}(p_k^\star(\lambda_j)) \cdot \mathbf{sgn}(p_k^\star(\lambda_{j+1})) < 0$.

(e) If λ_j and λ_{j+1} belong to two different intervals I_{i_1}, I_{i_2}, and $\mathbf{sgn}(p_k^\star(\lambda_j)) \cdot \mathbf{sgn}(p_k^\star(\lambda_{j+1})) > 0$, then $\lambda_j = L_{i_1}$, $\lambda_{j+1} = \mu_{i_2}$.

(f) If λ_j and λ_{j+1} belong to two different intervals I_{i_1}, I_{i_2}, and $\mathbf{sgn}(p_k^\star(\lambda_j)) \cdot \mathbf{sgn}(p_k^\star(\lambda_{j+1})) < 0$, then $\max_{\lambda_j \le \lambda \le \lambda_{j+1}} |p_k^\star(\lambda)| = \|p_k^\star\|_S$. Therefore $p_k^\star$ is also the best unifrom approximation within $\mathcal{P}_k^0$ for $I_1 \cup \cdots \cup I_{i_1-1} \cup [\mu_{i_1}, L_{i_2}] \cup I_{i_2+1} \cdots \cup I_m$.

(a-c) extends the well-known Chebyshev equioscillation theorem to the union of multiple intervals. These results were originally developed by Achieser (Achieser, 1928, 1932, 1933a,b, 1934) for the union of two intervals and later generalized by Grcar (1981). We adapt the statements from Schiefermayr (2011a). (d-f) is adapted from Schiefermayr and Peherstorfer (1999). Similar characterizations can also be found in Widom (1969); Nevai (1986); Lubinsky (1987); Fischer (1992); Peherstorfer (1993); Fischer (2011).

Next, we show that the best uniform approximation $p_k^\star \in \mathcal{P}_k^0$ is also the best discrete ℓ_2 approximation on the e-points $\lambda_1, \ldots, \lambda_s$ with a specific set of weights v_i's.

Lemma 34 *Under the same setting and notation of Lemma 33, define*

$$c_j = \begin{cases} \frac{1}{2} & \text{if } \mathbf{sgn}(p_k^\star(\lambda_j)) \cdot \mathbf{sgn}(p_k^\star(\lambda_{j+1})) > 0 \text{ or } j = 1 \text{ or } j = s \\ 1 & \text{otherwise} \end{cases}, \quad v_j = \sqrt{\frac{c_j/\lambda_j}{\sum_{i\in[s]} c_i/\lambda_i}}$$

Then for any $p \in \mathcal{P}_k^0$, the following inequality holds

$$\sum_{j\in[s]} (v_j p(\lambda_j))^2 \ge \|p_k^\star\|_S^2.$$

Proof [Proof of Lemma 34] Repeat the interval-merging procedure in Lemma 33(f) until there isn't any consecutive pair λ_j, λ_{j+1} that belongs to two different intervals but $\mathbf{sgn}(p_k^\star(\lambda_j)) \cdot \mathbf{sgn}(p_k^\star(\lambda_{j+1})) < 0$. After merging, there are exactly $s - k$ intervals left, denoted as $S' = \bigcup_{i\in[s-k]} J_i$.

By definition, $p_k^\star$ is a (un-normalized) T-polynomial on S' (see Schiefermayr and Peherstorfer (1999) for definition). By Theorem 2.3 of Schiefermayr and Peherstorfer (1999), for any polynomial q with $\deg q < k$, it is the case that $\sum_{j\in[s]} c_j p_k^\star(\lambda_j) q(\lambda_j) = 0$. Since both $p, p_k^\star \in \mathcal{P}_k^0$, we know that $\frac{p(\lambda) - p_k^\star(\lambda)}{\lambda}$ is a polynomial with degree $< k$ (since $p(0) = p_k^\star(0) = 1$). Hence

$$\sum_{j\in[s]} v_j^2 p_k^\star(\lambda_j)(p(\lambda_j) - p_k^\star(\lambda_j)) = \frac{1}{\sum_{j\in[s]} \frac{c_j}{\lambda_j}} \sum_{j\in[s]} c_j p_k^\star(\lambda_j) \frac{p(\lambda_j) - p_k^\star(\lambda_j)}{\lambda_j} = 0.$$

Therefore (by orthogonality)

$$\sum_{j\in[s]} (v_j p(\lambda_j))^2 \ge 2 \sum_{j\in[s]} v_j^2 p_k^\star(\lambda_j)(p(\lambda_j) - p_k^\star(\lambda_j)) + \sum_{j\in[s]} (v_j p_k^\star(\lambda_j))^2 = \sum_{j\in[s]} (v_j p_k^\star(\lambda_j))^2 = \|p_k^\star\|_S^2.$$

■

The proof of Lemma 29 is immediate once we have Lemma 33 and Lemma 34.

Proof [Proof of Lemma 29] Apply Lemma 33 and Lemma 34, one has for some $s \in \{k+1, \ldots, k+m\}$.

$$\max_{v_1,\ldots,v_s,\sum_{i\in[s]} v_i^2=1} \max_{\lambda_1,\ldots,\lambda_s\in S} \min_{p\in\mathcal{P}_k^0} \sqrt{\sum_{j\in[s]} (v_j p(\lambda_j))^2} \geq \min_{p\in\mathcal{P}_k^0} \max_{\lambda\in S} |p(\lambda)|.$$

Since $s \leq k+m$, we have therefore

$$\max_{v_1,\ldots,v_{k+m},\sum_{i\in[k+m]} v_i^2=1} \max_{\lambda_1,\ldots,\lambda_{k+m}\in S} \min_{p\in\mathcal{P}_k^0} \sqrt{\sum_{j\in[k+m]} (v_j p(\lambda_j))^2} \geq \min_{p\in\mathcal{P}_k^0} \max_{\lambda\in S} |p(\lambda)|.$$

■

D.3. Reference of Lemma 25: Reduction to the estimation of Green's function

In this section, we will cite literature from potential theory to reduce the uniform approximation problem raised in Lemma 24 to estimating Green's function, as stated in Lemma 25. Most of the results in this subsection are classic (c.f., Widom (1969); Grcar (1981); Aptekarev (1986); Nevai (1986); Lubinsky (1987); Driscoll et al. (1998); Embree and Trefethen (1999); Shen et al. (2001); Andrievskii (2004); Kuijlaars (2006); Saff (2010); Fischer (2011)). We follow the statements from Driscoll et al. (1998).

The following lemma gives the lower bound of uniform approximation by asymptotic convergence factor ρ_S.

Lemma 35 (Asymptotic convergence factor as non-asymptotic lower bound, slightly adapted from Driscoll et al. (*Let S be a compact (possibly not connected) subset of complex planes $\mathbb{C}$. Then the following limit exists*

$$\lim_{k\to\infty} \left(\min_{p\in\mathcal{P}_k^0} \max_{\lambda\in S} |p(\lambda)| \right)^{\frac{1}{k}} = \rho_S \leq 1,$$

*where the limiting value ρ_S is called the **asymptotic convergence factor** of S. Moreover, for any $k \in \mathbb{N}$, the following inequality holds*

$$\min_{p\in\mathcal{P}_k^0} \max_{\lambda\in S} |p(\lambda)| \geq \rho_S^k. \tag{D.2}$$

Recall $\mathcal{P}_k^0$ is defined in Eq. (D.1) as the set of polynomials p with degree at most k and $p(0) = 1$.

Remark 36 *Schiefermayr (2011a) shows that the RHS of inequality (D.2) can be improved to $\frac{2\rho_S^k}{1+\rho_S^{2k}}$ in the case that S is the union of a finite number of real intervals. We will still use the loose bound (D.2) for simplicity since they gave the same order of bound asymptotically.*

The asymptotic convergence factor of S can be analytically represented by the Green's function of S. We formally define the Green's function as follows.

Definition 37 (Definition of Green's function, borrowed from Driscoll et al. (1998)) *Let S be a compact (possibly not-connected) subset of $\mathbb{C}$ with no isolated points. Then the **Green's function** associated with S (with pole at ∞) is the unique $\mathbb{R}$-valued function defined on $\mathbb{C}\backslash S$ such that*

(a) g_S *is harmonic at* $\mathbb{C}\backslash S$.

(b) $g_S(z) \to 0$ *as* $z \to \partial S$.

(c) $g_S(z) - \log|z| \to C$ *as* $|z| \to \infty$ *for some constant* C.

The following result establishes the fundamental connection between Green's function of S and the asymptotic convergence factor of S. This result is classic and we cite the statement from Driscoll et al. (1998).

Lemma 38 (Representation of asymptotic convergence factor via Green's function, slightly adapted from Driscoll *Let S be a compact (possibly not-connected) subset of $\mathbb{C}$ with no isolated points. Let $g_S(z)$ be the Green's function associated with S. Then the asymptotic convergence factor of S is given by $\rho_S = \exp(-g_S(0))$.*

The proof of Lemma 25 then follows immediately from the above two lemmas.

D.4. Proof of Lemma 26: Estimating the Green's function

In this subsection, we will establish Lemma 26 on the upper bound of $g_S(0)$ for $S = \bigcup_{i\in[m]}[\mu_i, L_i]$. Our startpoint is the following classic results due to (Widom, 1969) on the explicit formula of g_S.

Lemma 39 (Green's function with respect to the union of real intervals, adapted from Section 14 of Widom (1969 *Let $S = \bigcup_{j=1}^{m}[\mu_j, L_j] \subset \mathbb{R}$ for some $0 < \mu_1 < L_1 < \cdots < \mu_m < L_m$. Let $q(z)$ be the polynomial*

$$q(z) := \prod_{j\in[m]} (z-\mu_j)(z-L_j).$$

Let $h(z)$ be the unique $(m-1)$-degree monic polynomials satisfying

$$\int_{L_k}^{\mu_{k+1}} \frac{h(\zeta)\mathrm{d}\zeta}{\sqrt{q(\zeta)}} = 0, \quad k = 1, \ldots, m-1.$$

Then the Green's function for S (with pole at ∞) at 0 is given by

$$g_S(0) = (-1)^{m+1} \int_0^{\mu_1} \frac{h(\zeta)\mathrm{d}\zeta}{\sqrt{q(\zeta)}}.$$

Remark 40 *Although Lemma 39 by Widom (1969) gives an exact formula to compute $g_S(0)$ (up to integration), it is hard to read off the dependency of $g_S(0)$ with respect to the condition numbers of the problem (local condition number κ_i and global condition number κ_{glob}). Numerous follow-up works have attempted to establish more concrete estimates of the Green's function when S has two or more intervals (Grcar, 1981; Lubinsky, 1987; Fischer, 1992; Peherstorfer, 1993; Shen et al., 2001; Andrievskii, 2004; Schiefermayr, 2008, 2011a,b; Alpan et al., 2016; Schiefermayr, 2017). Unfortunately, to the best of our knowledge, the existing estimate is either not sharp or not explicit for our purpose.*

We will give an explicit upper bound of $g_S(0)$. This estimate is novel to the best of our knowledge. Starting from Lemma 39, the proof of Lemma 26 relies on the following three technical lemmas. The first lemma upper bounds $g_S(0)$ with the product of the roots of h determined in Lemma 39.

Lemma 41 *Let $S = \bigcup_{j=1}^{m}[\mu_j, L_j] \subset \mathbb{R}$, and assume $\frac{L_j}{\mu_j} \geq 2$ for $j \in [m]$. Let $h(z)$ be the unique polynomial determined in Lemma 39, then $h(z)$ has $m-1$ real roots $r_1, r_2, \ldots, r_{m-1}$ such that $r_k \in [L_k, \mu_{k+1}]$, and the following inequality holds*

$$g_S(0) \leq \frac{7\prod_{k\in[m-1]}\frac{r_k}{\mu_{k+1}}}{\sqrt{\prod_{k\in[m]}\frac{L_k}{\mu_k}}}.$$

Remark 42 *Note that Lemma 41 immediately implies a coarse bound of $g_S(0) \leq \frac{7}{\sqrt{\prod_{k\in[m]}\frac{L_k}{\mu_k}}}$ since $r_k \leq \mu_{k+1}$.*

The second lemma establishes the following upper bound of r_k by the ratio of two integrals.

Lemma 43 *Under the same settings of Lemma 41, the k-th root of polynomial h satisfies the following inequality.*

$$r_k \leq 4 \cdot \frac{\int_{L_k}^{\mu_{k+1}} \frac{\zeta \mathrm{d}\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}}{\int_{L_k}^{\mu_{k+1}} \frac{\mathrm{d}\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}}.$$

The third lemma upper bounds the ratio encountered in Lemma 43.

Lemma 44 *Assume $\frac{L_j}{\mu_j} \geq 2$ (for any $j \in [m]$), then the following inequality holds for any $k \in [m-1]$,*

$$\frac{\int_{L_k}^{\mu_{k+1}} \frac{\zeta \mathrm{d}\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}}{\int_{L_k}^{\mu_{k+1}} \frac{\mathrm{d}\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}} \leq \frac{7\mu_{k+1}}{\log\left(16\frac{\mu_{k+1}}{L_k}\right)}.$$

The proof of Lemmas 41, 43 and 44 are standard yet tedious estimation of definite integrals, which we defer to Appendix I.

The proof of Lemma 26 then follows immediately from Lemmas 41, 43 and 44.

Proof [Proof of Lemma 26] By Lemmas 41, 43 and 44,

$$g_S(0) \leq \frac{7\prod_{k\in[m-1]}\frac{r_k}{\mu_{k+1}}}{\sqrt{\prod_{k\in[m]}\frac{L_k}{\mu_k}}} \leq \frac{7\prod_{k\in[m-1]}\frac{28}{\log(16\frac{\mu_{k+1}}{L_k})}}{\sqrt{\prod_{k\in[m]}\frac{L_k}{\mu_k}}} \leq \frac{7}{\sqrt{\prod_{k\in[m]}\frac{L_k}{\mu_k}\cdot\prod_{k\in[m-1]}\left(0.03\log(16\frac{\mu_{k+1}}{L_k})\right)}}.$$

■

Appendix E. Stochastic BSLS algorithm for quadratic multiscale optimization

In this section we prove Theorem 11, showing that a variant of `BSLS`, which we call `StochBSLS`, efficiently solves the stochastic version of a quadratic multiscale optimization problem from Theorem 4, restated below for convenience.

Definition *[Restated Theorem 4]The* ***stochastic quadratic multiscale optimization problem*** *asks to approximately solve the following problem*

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}} \left[\frac{1}{2}(\mathbf{a}^\top\mathbf{x} - b)^2\right],$$

where $b = \mathbf{a}^\top\mathbf{x}^\star$ for some fixed, unknown $\mathbf{x}^\star$ and the eigenvalues of the covariance matrix $\mathbb{E}_{\mathcal{D}}[\mathbf{a}\mathbf{a}^\top]$ can be partitioned into m "bands" such that for $i = 1, \ldots, m$ and $j = 1, \ldots, d_i$, each eigenvalue λ_{i_j} satisfies $\lambda_{i_j} \in [\mu_i, L_i]$ with $L_i < \mu_{i+1}$ for all $i < m$.

We introduce some additional notation. Let $n_{\max} = \max_{i\in[m]} d_i$. We let $\mathbf{H}_i \coloneqq \operatorname{diag}(\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_{d_i}})$ be the diagonal matrix with eigenvalues that lie in the band $[\mu_i, L_i]$ and $\mathbf{P}_i$ be an orthonormal matrix such that $\mathbf{P}_i\mathbf{\Sigma}\mathbf{P}_i^\top = \mathbf{H}_i$. Let $\mathbf{P} = \left(\mathbf{P}_1^\top, \ldots, \mathbf{P}_m^\top\right)^\top$ and $\mathbf{H} = \operatorname{diag}\left(\mathbf{H}_1, \ldots, \mathbf{H}_m\right)$. We will use the notation that for matrices and vectors, $\mathbf{M}^{(t)}$ or $\mathbf{x}^{(t)}$ refers to the t^{th} element of a sequence, while $\mathbf{M}_{ij}$ or $\mathbf{x}_i$ refers to the index of that matrix or vector. We note that this problem can be translated to the multiscale optimization problem formulation as per Def. 1. Indeed,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{2}(\mathbf{a}^\top\mathbf{x} - b)^2\right] = \frac{1}{2}\mathbf{x}^\top\mathbf{\Sigma}\mathbf{x} - \left(\mathbf{\Sigma}\mathbf{x}^\star\right)^\top\mathbf{x} + \left\|\mathbf{\Sigma}\mathbf{x}^\star\right\|_2^2 \\
&= \sum_{i\in[m]} \left(\frac{1}{2}\left(\mathbf{P}_i\mathbf{x}\right)^\top\mathbf{H}_i\left(\mathbf{P}_i\mathbf{x}\right) - \left(\mathbf{H}_i\mathbf{P}_i\mathbf{x}^\star\right)^\top\left(\mathbf{P}_i\mathbf{x}\right) + \frac{1}{m}\left\|\mathbf{\Sigma}\mathbf{x}^\star\right\|_2^2\right) \\
&= \sum_{i\in[m]} f_i(\mathbf{P}_i\mathbf{x}), \quad \text{for} \quad f_i(\mathbf{v}) \coloneqq \frac{1}{2}\mathbf{v}^\top\mathbf{H}_i\mathbf{v} - \left(\mathbf{H}_i\mathbf{P}_i\mathbf{x}^\star\right)^\top\mathbf{v} + \frac{1}{m}\left\|\mathbf{\Sigma}\mathbf{x}^\star\right\|_2^2,
\end{aligned}$$

where each $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ satisfies the constraints of Def. 1. Therefore the problem from Def. 4 can be thought of as a stochastic version of the general problem from Section B.

E.1. Proof overview of Theorem 11

In what follows we prove Theorem 11, guaranteeing the convergence rate of `StochBSLS` in expectation for the stochastic quadratic multiscale optimization problem (Theorem 4). First, Section E.2 uses our distributional assumptions to establish that if $\texttt{StochBSLS}_1$ takes N_1 steps then

$$\mathbb{E}\left[\left\|\texttt{StochBSLS}(\mathbf{x}^{(0)}) - \mathbf{x}^\star\right\|_2^2\right] = (\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top\mathbf{D}^{(N_1)}(\mathbf{x}^{(0)} - \mathbf{x}^\star),$$

where $\left\{\mathbf{D}^{(t)}\right\}_{t=0}^{N_1}$ is a sequence of matrices with a clean recurrence relation. Next, Section E.3 uses this recurrence relation to bound the spectral norm of each $\mathbf{D}^{(t)}$. This is where the band structure of the eigenvalues plays a role and the stochasticity poses an obstacle. Finally in Theorem 54 we use the previous work to prove Theorem 11 without too much effort since Section E.3 guarantees that $\mathbf{D}^{(N_1)}$ has sufficiently small spectral norm. Finally in Section E.4 we extend our analysis to the setting where only m, μ_1, L_m, and $\prod_{i\in m}\kappa_i$ are known.

To this end, we introduce some notation in addition to the notation from the beginning of Appendix E. We let $\mathbf{A} \coloneqq \frac{1}{n_{\text{avg}}}\sum_{i\in n_{\text{avg}}}\mathbf{a}^{(i)}\mathbf{a}^{(i)\top}$ denote the empirical covariance matrix and $\mathbf{b} \coloneqq \frac{1}{n_{\text{avg}}}\sum_{i\in n_{\text{avg}}} b^{(i)}\mathbf{a}^{(i)}$ be the empirical approximation to $\mathbf{\Sigma}\mathbf{x}^\star$. For convenience we introduce $\delta \coloneqq \mathsf{Kurt}(\mathcal{D})n_{\max}/n_{\text{avg}}$, which (very roughly) corresponds to the noise induced by stochasticity.

Assumption 45 (Distribution assumptions) *For $\mathbf{a} \sim \mathcal{D}$ we assume*

(a) For any $i, j, k \in [d]$ with $i \neq j$ it is the case that $\mathbb{E}[(\mathbf{Pa})_i(\mathbf{Pa})_k^2(\mathbf{Pa})_j] = 0$. (Recall that in this section, $\mathbf{x}_i$ refers to the i^{th} index of vector $\mathbf{x}$.)

(b) There exists a constant $\mathsf{Kurt}(\mathcal{D})$ such that for any $\mathbf{w} \in \mathbb{R}^d$, $\mathbb{E}[(\mathbf{w}^\top \mathbf{a})^4] \leq \mathsf{Kurt}(\mathcal{D})\mathbb{E}[(\mathbf{w}^\top \mathbf{a})^2]^2$.

E.2. Simplifying the stochasticity

With the notation and assumptions in place, we begin with Lemma 46 which bounds the degree to which stochasticity poses an obstacle. For motivation, first suppose that we had no stochasticity in that instead of approximating $\mathbf{\Sigma}$ by $\mathbf{A} = \frac{1}{n}\sum_{i \in n} \mathbf{a}_i \mathbf{a}_i^\top$, we had access to $\mathbf{\Sigma}$ itself. Then in $\texttt{SGD}(\mathbf{x}; L)$ we would have instead

$$\mathbf{g} = \mathbf{\Sigma}(\mathbf{x} - \mathbf{x}^\star),$$

and

$$\texttt{SGD}(\mathbf{x}; L) - \mathbf{x}^\star = \mathbf{x} - \frac{1}{L}\mathbf{\Sigma}(\mathbf{x} - \mathbf{x}^\star) - \mathbf{x}^\star = \left(\mathbf{I} - \frac{1}{L}\mathbf{\Sigma}\right)(\mathbf{x} - \mathbf{x}^\star).$$

Therefore if t denotes the total number of calls to $\texttt{SGD}$ and $\eta^{(t)}$ is the stepsize taken at step t we have

$$\mathbf{x}^{(t)} - \mathbf{x}^\star = \left(\mathbf{I} - \eta^{(t)}\mathbf{\Sigma}\right)\left(\mathbf{x}^{(t-1)} - \mathbf{x}^\star\right) = \left(\mathbf{I} - \eta^{(t)}\mathbf{\Sigma}\right) \cdots \left(\mathbf{I} - \eta^{(1)}\mathbf{\Sigma}\right)\left(\mathbf{x}^{(0)} - \mathbf{x}^\star\right).$$

Critically, since $\mathbf{\Sigma}$ commutes with itself we can simplify the above to

$$\mathbf{x}^{(t)} - \mathbf{x}^\star = \left(\prod_{i=1}^{m}\left(\mathbf{I} - \frac{1}{L_i}\mathbf{\Sigma}\right)^{N_i}\right)(\mathbf{x}^{(0)} - \mathbf{x}^\star).$$

Therefore

$$\left\|\mathbf{x}^{(t)} - \mathbf{x}^\star\right\|_2^2 = \left(\mathbf{P}(\mathbf{x}^{(0)} - \mathbf{x}^\star)\right)^\top \left(\prod_{i=1}^{m}\left(\mathbf{I} - \frac{1}{L_i}\mathbf{H}\right)^{N_i}\right)^2 \mathbf{P}(\mathbf{x}^{(0)} - \mathbf{x}^\star).$$

Then using the fact that for each eigenvalue of $\mathbf{\Sigma}$ we have $\lambda_{i_j} \in [\mu_i, L_i]$ we have,

$$\left\|\mathbf{x}^{(N_1)} - \mathbf{x}^\star\right\|_2^2 \leq \sum_{j=1}^{m}\left\|\mathbf{P}_j^\top\left(\mathbf{x}^{(0)} - \mathbf{x}^\star\right)\right\|_2^2 \cdot \prod_{i=1}^{m}\left(1 - \frac{\mu_j}{L_i}\right)^{2N_i} \leq \sum_{j=1}^{m}\left\|\mathbf{P}_j^\top\left(\mathbf{x}^{(0)} - \mathbf{x}^\star\right)\right\|_2^2 \cdot \left(1 - \frac{1}{\kappa_j}\right)^{N_j} \kappa_{\text{glob}}^{2\sum_{i=j+1}^{m} N_i}.$$

Written this way we see that for some constant C we can bound by $\left\|\mathbf{x}^{(N_1)} - \mathbf{x}^\star\right\|_2^2$ by ϵ if $N_j \geq C\kappa_j \log(\kappa_{\text{glob}}) \sum_{i=j+1}^{m} N_i$ and $N_1 \geq C \log(\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2 / \epsilon)\kappa_1 \log(\kappa_{\text{glob}}) \sum_{i=2}^{m} N_i$. This would give an overall query complexity of $\mathcal{O}\left(\left(\prod_{i \in m} \kappa_i\right) \log^m(\kappa_{\text{glob}}) \log\left(\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2 / \epsilon\right)\right)$. Instead, in the stochastic case we have

$$\left\|\mathbf{x}^{(t)} - \mathbf{x}^\star\right\|_2^2 = \left(\mathbf{x}^{(0)} - \mathbf{x}^\star\right)^\top \left(\prod_{s=1}^{t}\left(\mathbf{I} - \eta^{(s)}\mathbf{A}^{(s)}\right)\right)\left(\prod_{s=1}^{t}\left(\mathbf{I} - \eta^{(t-s)}\mathbf{A}^{(t-s)}\right)\right)\left(\mathbf{x}^{(0)} - \mathbf{x}^\star\right). \tag{E.1}$$

The random instances $\mathbf{A}^{(t)}$ do not necessarily commute with each other and so simplifying their product is not as simple as the non-stochastic case. The following lemma roughly shows we can replace the above $\mathbf{A}^{(t)}$ with a perturbation of $\mathbf{\Sigma}$.

Lemma 46 (Understanding Second Moments) *Recall* $\mathbf{a} \overset{iid}{\sim} \mathcal{D}$ *and* $\mathbf{A}$. *Suppose some matrix* $\mathbf{D}$ *commutes with the covariance matrix* $\mathbf{\Sigma}$. *Then* $\mathbb{E}[\mathbf{ADA}]$ *also commutes with* $\mathbf{\Sigma}$ *and*

$$\mathbb{E}[\mathbf{ADA}] \preceq \frac{n_{avg}-1}{n_{avg}}\mathbf{\Sigma D \Sigma} + \frac{\mathsf{Kurt}(\mathcal{D})}{n_{avg}} \operatorname{tr}(\mathbf{D\Sigma})\,\mathbf{\Sigma}.$$

Next we can simplify Eq. E.1 by sequentially conditioning on $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(t-1)}$ and then invoking Lemma 46 for $\mathbf{A}^{(t)}$. Lemma 47 does this explicitly and in doing so constructs the aforementioned sequence $\left\{\mathbf{D}^{(t)}\right\}_{t=0}^{N_1}$. After Lemma 47 the purpose of the remainder of the proof is only to bound the spectral norm of $\mathbf{D}^{(N_1)}$.

Lemma 47 *Recall* $\mathbf{a} \sim \mathcal{D}$ *and the definitions of* $\mathbf{A}$ *and* $\mathbf{b}$. *Recall we use* $\mathbf{g} = \mathbf{Ax} - \mathbf{b}$ *in the subroutine* $\texttt{SGD}(\mathbf{x}; L)$. *Define*

$$\mathbf{D}^{(t)} := \mathbb{E}[(\mathbf{I} - \eta^{(N-t+1)}\mathbf{A}^{(N-t+1)})\mathbf{D}^{(t-1)}(\mathbf{I} - \eta^{(N-t+1)}\mathbf{A}^{(N-t+1)})], \qquad \mathbf{D}^{(0)} := \mathbf{I}.$$

Then if N *denotes the total number of calls to* $\texttt{SGD}(\mathbf{x}; L)$ *we have*

$$\mathbb{E}\left[\left\|\texttt{StochBSLS}_1(\mathbf{x}^{(0)}) - \mathbf{x}^\star\right\|_2^2\right] = (\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{D}^{(N)}(\mathbf{x}^{(0)} - \mathbf{x}^\star).$$

Proof [Proof of Lemma 47] We begin our proof by noting that for $\mathbf{g}$ as defined in $\texttt{SGD}(\mathbf{x}; L)$ we have

$$\mathbf{g} = \frac{1}{n_{\text{avg}}} \sum_{i \in n_{\text{avg}}} \mathbf{a}^{(i)}\mathbf{a}^{(i)\top}(\mathbf{x} - \mathbf{x}^\star).$$

Thus we have

$$\texttt{SGD}(\mathbf{x}; L) - \mathbf{x}^\star = \mathbf{x} - \frac{1}{L}\mathbf{A}(\mathbf{x} - \mathbf{x}^\star) - \mathbf{x}^\star = \left(\mathbf{I} - \frac{1}{L}\mathbf{A}\right)(\mathbf{x} - \mathbf{x}^\star).$$

Therefore if t denotes the total number of calls to $\texttt{SGD}$, $\mathbf{A}^{(t)}$ denotes the random matrix generated in the t^{th} call to $\texttt{SGD}$, and $\eta^{(t)}$ is the stepsize taken at step t we have

$$\mathbf{x}^{(t)} - \mathbf{x}^\star = \left(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(t)}\right)\left(\mathbf{x}^{(t-1)} - \mathbf{x}^\star\right) = \left(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(t)}\right)\cdots\left(\mathbf{I} - \eta^{(1)}\mathbf{A}^{(1)}\right)\left(\mathbf{x}^{(0)} - \mathbf{x}^\star\right).$$

Therefore,

$$\left\|\mathbf{x}^{(N)} - \mathbf{x}^\star\right\|_2^2 = (\mathbf{x}^{(N)} - \mathbf{x}^\star)^\top(\mathbf{x}^{(N)} - \mathbf{x}^\star) \tag{E.2}$$

$$= (\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \left(\mathbf{I} - \eta^{(1)}\mathbf{A}^{(1)}\right)^\top \cdots \left(\mathbf{I} - \eta^{(N)}\mathbf{A}^{(N)}\right)^\top \left(\mathbf{I} - \eta^{(N)}\mathbf{A}^{(N)}\right) \cdots \left(\mathbf{I} - \eta^{(1)}\mathbf{A}^{(1)}\right)(\mathbf{x}^{(0)} - \mathbf{x}^\star).$$

Let $\mathbf{D}^{(0)} := \mathbf{I}$ and $\mathbf{D}^{(t)} := \mathbb{E}\left[(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(N-t+1)})\mathbf{D}^{(t-1)}(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(N-t+1)})\right]$. For short let $\mathbf{M}^{(t)} := \left(\mathbf{I} - \eta^{(1)}\mathbf{A}^{(1)}\right)\ldots\left(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(t)}\right)$ and $\mathbf{M}_{\text{rev}}^{(t)} := \left(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(t)}\right)\ldots\left(\mathbf{I} - \eta^{(1)}\mathbf{A}^{(1)}\right)$. Using this notation and using the independence of $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}$ we have for any $k \geq 1$,

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{M}^{(N-k+1)}\mathbf{D}^{(k-1)}\mathbf{M}_{\text{rev}}^{(N-k+1)}(\mathbf{x}^{(0)} - \mathbf{x}^\star)\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[(\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{M}^{(N-k+1)}\mathbf{D}^{(k-1)}\mathbf{M}_{\text{rev}}^{(N-k+1)}(\mathbf{x}^{(0)} - \mathbf{x}^\star) \mid \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N-k)}\right]\right] \\
=&\mathbb{E}\left[(\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{M}^{N-k}\mathbb{E}\left[\left(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(N-k+1)}\right)\mathbf{D}^{(k-1)}\left(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(N-k+1)}\right)\right]\mathbf{M}_{\text{rev}}^{N-k}(\mathbf{x}^{(0)} - \mathbf{x}^\star)\right] \\
=&\mathbb{E}\left[(\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{M}^{N-k}\mathbf{D}^{(k)}\mathbf{M}_{\text{rev}}^{N-k}(\mathbf{x}^{(0)} - \mathbf{x}^\star)\right]
\end{aligned}$$

Therefore using Eq. E.2 in the first equality and the above recursion in the second equality we have

$$\mathbb{E}\left[\left\|\mathbf{x}^N - \mathbf{x}^\star\right\|_2^2\right] = \mathbb{E}\left[(\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{M}^N \mathbf{D}^{(0)} \mathbf{M}_{\text{rev}}^N (\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top\right] = (\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{D}^{(N)} (\mathbf{x}^{(0)} - \mathbf{x}^\star).$$

■

E.3. Bounding the spectral norm of $\mathbf{D}^{(t)}$

This section is where we address the difficulty posed by stochasticity. From Lemma 47 we see that it suffices to bound the spectral norm of $\mathbf{D}^{N_1}$. To that end, in the following lemma we construct a clean recursive form to analyze the sequence $\left\{\mathbf{D}^{(t)}\right\}$.

Lemma 48 *For $\mathbf{D}^{(t)}$ as defined in Lemma 47 we have that $\mathbf{D}^{(t)}$ commutes with $\mathbf{\Sigma}$. Moreover we have the following spectral upperbound,*

$$\mathbf{D}^{(t)} \preceq (\mathbf{I} - \eta^{(t)}\mathbf{\Sigma})^2 \mathbf{D}^{(t-1)} + \frac{\mathsf{Kurt}(\mathcal{D})\eta^{(t)2}}{n_{avg}} \operatorname{tr}(\mathbf{D}^{(t-1)}\mathbf{\Sigma})\mathbf{\Sigma}.$$

Proof [Proof of Lemma 48] Recalling the definition of $\mathbf{D}^{(t)}$ from Lemma 47,

$$\mathbf{D}^{(t)} \coloneqq \mathbb{E}[(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(t)})\mathbf{D}^{(t-1)}(\mathbf{I} - \eta^{(t)}\mathbf{A}^{(t)})] \qquad \mathbf{D}^{(0)} \coloneqq \mathbf{I}.$$

we have

$$\mathbf{D}^{(t)} = \left((\mathbf{I} - \eta^{(t)}\mathbf{\Sigma})\mathbf{D}^{(t-1)}(\mathbf{I} - \eta^{(t)}\mathbf{\Sigma})\right) + \eta^{(t)2}\left(\mathbb{E}[\mathbf{A}^{(t)}\mathbf{D}^{(t-1)}\mathbf{A}^{(t)}] - \mathbf{\Sigma}\mathbf{D}^{(t-1)}\mathbf{\Sigma}\right).$$

Lemma 46 allows us to bound $\mathbb{E}[\mathbf{A}^{(t)}\mathbf{D}^{(t-1)}\mathbf{A}^{(t)}]$ from above. Using this we have,

$$\mathbf{D}^{(t)} \preceq \left((\mathbf{I} - \eta^{(t)}\mathbf{\Sigma})\mathbf{D}^{(t-1)}(\mathbf{I} - \eta^{(t)}\mathbf{\Sigma})\right) + \frac{\eta^{(t)2}}{n_{\text{avg}}}\left(\mathsf{Kurt}(\mathcal{D})\operatorname{tr}(\mathbf{D}^{(t-1)}\mathbf{\Sigma})\mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{D}^{(t-1)}\mathbf{\Sigma}\right).$$

By Lemma 46 $\mathbf{\Sigma}$ and $\mathbf{D}^{(t-1)}$ commute and thus we have more simply,

$$\begin{aligned}\mathbf{D}^{(t)} &\preceq (\mathbf{I} - \eta^{(t)}\mathbf{\Sigma})^2\mathbf{D}^{(t-1)} + \frac{\eta^{(t)2}}{n_{\text{avg}}}\left(\mathsf{Kurt}(\mathcal{D})\operatorname{tr}(\mathbf{D}^{(t-1)}\mathbf{\Sigma})\mathbf{\Sigma} - \mathbf{D}^{(t-1)}\mathbf{\Sigma}^2\right) \\ &\preceq (\mathbf{I} - \eta^{(t)}\mathbf{\Sigma})^2\mathbf{D}^{(t-1)} + \frac{\mathsf{Kurt}(\mathcal{D})\eta^{(t)2}}{n_{\text{avg}}}\operatorname{tr}(\mathbf{D}^{(t-1)}\mathbf{\Sigma})\mathbf{\Sigma}. \qquad (\mathbf{D}^{(t-1)}\mathbf{\Sigma}^2 \text{ is PSD})\end{aligned}$$

■

Remark 49 *Recall that $\mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top = \mathbf{H}$ and $\mathbf{H} = \operatorname{diag}(\mathbf{H}_1, \ldots, \mathbf{H}_m)$, where $\mathbf{H}_i \coloneqq \operatorname{diag}(\lambda_{i,1}, \ldots, \lambda_{i,d_i})$ represents the i^{th} eigenvalue band. By Lemma 48 each $\mathbf{D}^{(t)}$ commutes with $\mathbf{\Sigma}$. Therefore if $\tilde{\mathbf{D}}^{(t)} = \mathbf{P}\mathbf{D}\mathbf{P}^\top$ then $\tilde{\mathbf{D}}^{(t)}$ is diagonal and*

$$\tilde{\mathbf{D}}^{(t)} \leq (\mathbf{I} - \eta^{(t)}\mathbf{H})^2\tilde{\mathbf{D}}^{(t-1)} + \frac{\mathsf{Kurt}(\mathcal{D})\eta^{(t)2}}{n_{avg}}\operatorname{tr}(\tilde{\mathbf{D}}^{(t-1)}\mathbf{H})\mathbf{H}. \tag{E.3}$$

Thus the structure on $\mathbf{H}$ induces structure on matrix $\tilde{\mathbf{D}}^{(t)}$,

$$\tilde{\mathbf{D}}^{(t)} = \operatorname{diag}(\tilde{\mathbf{D}}_1^{(t)}, \ldots, \tilde{\mathbf{D}}_m^{(t)}).$$

Lemma 50 *Let* $g_i(\eta) \coloneqq \max\left\{(1-\eta\mu_i)^2, (1-\eta L_i)^2\right\}$. *Define the "update" matrix from stepsize* $\eta^{(t)}$ *as*

$$\left(\mathbf{U}_{\eta^{(t)}}\right)_{ij} \coloneqq \begin{cases} g_i(\eta) & \text{if } i = j, \\ \delta\eta^2 L_i L_j & \text{else.} \end{cases} \tag{E.4}$$

Define the following vector to represent the maximum entry of each $\tilde{\mathbf{D}}_i^{(t)}$,

$$\mathbf{r}_i^{(t)} \coloneqq \left\|\tilde{\mathbf{D}}_i^{(t)}\right\|_\infty,$$

(and note that since $\tilde{\mathbf{D}}^{(0)} = \mathbf{I}$ *then* $\mathbf{r}^{(0)} = \mathbf{1}$*). Then for any* $i = 1, \ldots, m$ *we have for* $t \geq 1$,

$$\mathbf{r}^{(t)} \leq \mathbf{U}_{\eta^{(t)}}\mathbf{r}^{(t-1)}.$$

Proof [Proof of Lemma 50] From Lemma 48 we can bound the growth of $\mathbf{r}^{(t)}$. Let $\sigma(k,\ell) \in \mathbb{Z}$ denote the index corresponding to the ℓ^{th} smallest eigenvalue of the k^{th} band so that $\mathbf{H}_{\sigma(k,\ell)} = \lambda_{k,\ell}$. Letting

$$g_i(\eta) \coloneqq \max\left\{(1-\eta\mu_i)^2, (1-\eta L_i)^2\right\},$$

we have (using Eq. E.3),

$$\begin{aligned}
\mathbf{r}_i^{(t+1)} &= \max_{j=1,\ldots,d_i}\left\{(1-\eta^{(t)}\lambda_{i,j})^2\mathbf{r}_i^{(t)} + \frac{\mathsf{Kurt}(\mathcal{D})\eta^{(t)2}}{n_{\text{avg}}}\sum_{k=1}^{m}\sum_{\ell=1}^{d_k}\mathbf{D}^{(t)}_{\sigma(k,\ell)}\mathbf{H}_{\sigma(k,\ell)}\lambda_{i,j}\right\} \\
&\leq \max_{j=1,\ldots,d_i}\left\{(1-\eta^{(t)}\lambda_{i,j})^2\right\}\mathbf{r}_i^{(t)} + \frac{\mathsf{Kurt}(\mathcal{D})\eta^{(t)2}}{n_{\text{avg}}}\left(\sum_{k=1}^{m} d_k\mathbf{r}_k^{(t)}L_k\right)L_i \\
&\leq \max\left\{(1-\eta^{(t)}\mu_i)^2, (1-\eta^{(t)}L_i)^2\right\}\mathbf{r}_i^{(t)} + \frac{\mathsf{Kurt}(\mathcal{D})\eta^{(t)2}}{n_{\text{avg}}}\left(\sum_{k=1}^{m} d_k\mathbf{r}_k^{(t)}L_k\right)L_i \\
&\leq g_i(\eta^{(t)})\mathbf{r}_i^{(t)} + \frac{\mathsf{Kurt}(\mathcal{D})n_{\max}\eta^{(t)2}}{n_{\text{avg}}}\left(\sum_{k=1}^{m}\mathbf{r}_k^{(t)}L_k\right)L_i.
\end{aligned}$$

Inspecting the definition of $\mathbf{U}_{\eta^{(t)}}$ finishes the proof. ■

Remark 51 *For simplicity, let* $\mathbf{U}_{i_t}$ *denote* $\mathbf{U}_{\eta^{(t)}}$ *where* i_t *is the index belonging to* $\{1, \ldots, m\}$ *such that the stepsize in the* t^{th} *step corresponds to the* i^{th} *eigenvalue band; that is:* $\eta^{(t)} = 1/L_{i_t}$. *Recall that Lemma 50 guarantees that for* $\mathbf{r}_i^{(t)} \coloneqq \left\|\tilde{\mathbf{D}}_i^{(t)}\right\|_\infty$

$$\mathbf{r}^{(t)} \leq \mathbf{U}_{i_t}\mathbf{r}^{(t-1)}.$$

We ultimately want to bound $\left\|\mathbf{r}^{N_1}\right\|_\infty$, *however the evolution of* $\left\{\mathbf{r}^{(t)}\right\}_{t=0}^{N_1}$ *is difficult to track exactly. Instead we can analyze the evolution of* $\left\{\mathbf{u}^{(t)}\right\}_{t=0}^{N}$ *where*

$$\mathbf{u}^{(t)} \coloneqq \mathbf{U}_{i_t}\mathbf{u}^{(t-1)} \qquad \mathbf{u}^{(0)} = \mathbf{r}^{(0)}.$$

Taking this another step further, for convenience we define

$$(\mathbf{V}_i)_{jk} := \begin{cases} \rho^{N_{i+1}}\left(\mathbf{u}^{(0)}\right), \text{ if } j = k \text{ and } j < i, \\ \gamma_i, \text{ if } j = k \text{ and } j = i, \\ \frac{L_j}{L_i}\rho\left(\mathbf{u}^{(0)}\right)^{N_{i+1}+1}, \text{ if } j > i \text{ and } k = i, \\ 0, \text{ else.} \end{cases}$$

and

$$(\mathbf{W}_i)_{jk} := \begin{cases} 1, \text{ if } j = k = i, \\ (\mathbf{V}_i)_{jk}, \text{ else.} \end{cases}$$

Suppose we now re-define $\left\{\mathbf{u}^{(t)}\right\}_{t=0}^{N}$ *where either*

$$\mathbf{u}^{(t)} := \max\left\{\mathbf{U}_{i_t}\mathbf{u}^{(t-1)}, \mathbf{V}_{i_t}\mathbf{u}^{(t-1)}\right\} \qquad \mathbf{u}^{(0)} = \mathbf{r}^{(0)}, \tag{E.5}$$

or

$$\mathbf{u}^{(t)} := \max\left\{\mathbf{U}_{i_t}\mathbf{u}^{(t-1)}, \mathbf{W}_{i_t}\mathbf{u}^{(t-1)}\right\} \qquad \mathbf{u}^{(0)} = \mathbf{r}^{(0)}. \tag{E.6}$$

Then since

$$\mathbf{r}^{(t)} \leq \mathbf{u}^{(t)}$$

we can analyze $\left\{\mathbf{u}^{(t)}\right\}_{t=0}^{N_1}$ *and bound* $\left\|\mathbf{u}^{(N_1)}\right\|_\infty$ *to get a bound on* $\left\|\mathbf{r}^{N_1}\right\|_\infty$*. This is convenient because the evolution of* $\mathbf{u}^{(t)}$ *is easier to track while capturing the critical behavior of the evolution of* $\mathbf{r}^{(t)}$*. Towards this end, we introduce Algorithm 5 which we call as* `StochBSLSRes` *(*`Res` *for "residuals") and which mirrors the structure of* `StochBSLS`*. Lemma 52, which bounds* $\left\|\mathbf{u}^{(N_1)}\right\|_\infty$ *from Algorithm 5, is the heart of the proof of Theorem 11.*

Algorithm 5 BSLS Residuals [For analysis of the stochastic variant]

Procedure $\texttt{StochBSLSRes}_i\ (\mathbf{u})$

for $t = 0, 1, \ldots, T_i - 1$ **do**
  if $i < m$ **then**
    $\tilde{\mathbf{u}}^{(t)} \leftarrow \texttt{StochBSLSRes}_{i+1}(\mathbf{u}^{(t)})$
  else
    $\tilde{\mathbf{u}}^{(t)} \leftarrow \mathbf{u}^{(t)}$
  if $t > \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$ and $i \geq 2$ **then**
    $\mathbf{u}^{(t+1)} \leftarrow \max\left\{\mathbf{U}_i\tilde{\mathbf{u}}^{(t)}, \mathbf{W}_i\tilde{\mathbf{u}}^{(t)}\right\}$ (for $\mathbf{W}_i$ as defined in Eq. E.6)
  else
    $\mathbf{u}^{(t+1)} \leftarrow \max\left\{\mathbf{U}_i\tilde{\mathbf{u}}^{(t)}, \mathbf{V}_i\tilde{\mathbf{u}}^{(t)}\right\}$ (for $\mathbf{U}_i$ and $\mathbf{V}_i$ as defined in Eqs. E.4, E.5 respectively)
return $\texttt{StochBSLSRes}_{i+1}(\mathbf{u}^{(T_i)})$

Lemma 52 *For any $i = 1, \ldots, m$ define $\gamma_i \coloneqq 1 - (2\kappa_i)^{-1}$. Fix some $i \in \{2, \ldots, m\}$ and let $T_i = \lceil 8\kappa_i \log(\kappa_{\text{glob}}) \rceil$. Let $N_i = \prod_{j=i}^{m}(2T_i + 1)$ and $T_{max} = \max_i T_i$. Define*

$$\beta_0(\mathbf{u}) \coloneqq \inf\left\{ t \,\middle|\, \mathbf{u}_k \le t \max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\} \mathbf{u}_j \text{ for any } j, k \in [m] \right\}$$
$$\rho(\mathbf{u}) \coloneqq 1 + 3\delta m \cdot \beta_0(\mathbf{u})$$
$$\beta_{total}(\mathbf{u}) \coloneqq \rho(\mathbf{u})^{T_{max}(N_1+2)^2+1} \cdot \max_{\ell \in [m]}\left\{\frac{1}{\gamma_\ell^2}\right\}$$

Suppose that

$$\beta_0(\mathbf{u})\, \beta_{total}^{m-i+1}(\mathbf{u}) \le \min\left\{ \frac{\kappa_{\text{glob}}}{\rho(\mathbf{u})^{N_1}}, \frac{1}{144 N_1 T_{max}\delta m}, \frac{1}{2\max_{\ell\in[m]}\kappa_\ell} \frac{1}{6(N_1+1)\delta m} \right\}. \tag{E.7}$$

Further suppose that

$$n_{avg} \ge \mathsf{Kurt}(\mathcal{D}) m^2 n_{max} \left(\prod_{i\in[m]} \kappa_i\right)\left(\max_{i\in[m]}\kappa_i\right) \log\left(\frac{9\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2}{\varepsilon^2}\right) \log^m(\kappa_{\text{glob}}).$$

Then if $\tilde{\mathbf{u}} = \texttt{StochBSLSRes}_i(\mathbf{u})$ we have that for all $j \ge i$,

$$\tilde{\mathbf{u}}_j \le \frac{L_{i-1}}{L_j}\mathbf{u}_{i-1},$$

and for all $j < i$,

$$\tilde{\mathbf{u}}_j \le \rho^{N_i}(\mathbf{u}) \cdot \mathbf{u}_j.$$

The proof of Lemma 52 requires careful and somewhat tedious analysis of the evolution of $\mathbf{u}^{(t)}$. The difficulty lies in controlling the error induced by stochasticity. For a full proof see Appendix J.2. With Lemma 52, we can now easily bound the convergence of $\mathbf{u}^{(t)}$ which then allows us to bound the spectral norm of $\tilde{\mathbf{D}}^{(t)}$.

Lemma 53 *Suppose that $m \le \log(\kappa_{\text{glob}})/3$ and*

$$n_{avg} \ge \mathsf{Kurt}(\mathcal{D}) m^2 n_{max} \left(\prod_{i\in[m]} \kappa_i\right)\left(\max_{i\in[m]}\kappa_i\right) \log\left(9\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2/\varepsilon\right) \log^m(\kappa_{\text{glob}}).$$

For $i = 2, \ldots, m$ let T_i be as in Lemma 52 and let $T_1 = \left\lceil 2\kappa_1 \log\left(\frac{9\|\mathbf{x}^{(0)}-\mathbf{x}^\star\|_2^2}{\epsilon}\right)\right\rceil$. Then if $\tilde{\mathbf{u}} = \texttt{StochBSLSRes}_1(\mathbf{1})$ we have for all i

$$\tilde{\mathbf{u}}_i \le \frac{\epsilon}{\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2}.$$

Proof [Proof of Lemma 53] First we show that Eq. E.7 holds for $\mathbf{u}^{(0)} = \mathbf{1}$. We bound $\beta_0(\mathbf{1}), \rho(\mathbf{1}),$ and $\beta_{\text{total}}(\mathbf{1})$. Using that $\max_{\ell\in[m]}\left\{\frac{1}{\gamma_\ell^2}\right\} \leq 4$ to bound $\beta_{\text{total}}(\mathbf{1})$ we have

$$\begin{aligned}
&\beta_0(\mathbf{1}) \leq 1\\
&\rho(\mathbf{1}) = 1 + 3\delta m\left(\max_{i\leq m-1}\frac{L_i}{L_{i+1}}\right) \leq 1 + 3\delta m\\
&\beta_{\text{total}}(\mathbf{1}) \leq 4\left(1+3\delta m\right)^{T_{\max}(N_1+2)^2+1}.
\end{aligned}$$

To show Eq. E.7 holds we must show

$$4^m(1+3\delta m)^{m\left(T_{\max}(N_1+2)^2+1\right)} \leq \min\left\{\frac{\kappa_{\text{glob}}}{(1+3\delta m)^{N_1}}, \frac{1}{144N_1T_{\max}\delta m}, \frac{1}{2\max_{\ell\in[m]}\{\kappa_\ell\}}\frac{1}{6(N_1+1)\delta m}\right\}. \tag{E.8}$$

Since

$$n_{\text{avg}} \geq \mathsf{Kurt}(\mathcal{D})m^2n_{\max}\left(\prod_{i\in[m]}\kappa_i\right)\left(\max_{i\in[m]}\kappa_i\right)\log\left(\frac{9\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|_2^2}{\varepsilon^2}\right)\log^m(\kappa_{\text{glob}}),$$

then recalling that $\delta = \mathsf{Kurt}(\mathcal{D})n_{\max}/n_{\text{avg}}$ and noting that $N_1 \leq 2\kappa_1\log\left(9\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|_2^2/\varepsilon^2\right)\prod_{i=2}^m 8\kappa_i\log(\kappa_{\text{glob}})$ we have

$$\begin{aligned}
\delta \leq \min\Bigg\{&\frac{1}{3m}\frac{1}{T_{\max}(N_1+2)^2+1},\\
&\frac{1}{6m(4^m)}\left(48mT_{\max}^2(N_1+2)^3\right)^{-1/2},\\
&\frac{1}{6m(4^m)}\left(\max_{\ell\in[m]}\{\kappa_\ell\}\cdot mT_{\max}^2(N_1+2)^3\right)^{-1/2}\Bigg\}.
\end{aligned}$$

This guarantees that Eq. E.8 holds; to see the details please refer to Appendix J.1. Next we show

$$T_1 \geq \log_{(1/\gamma_1)}\left(\frac{9\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|_2^2}{\epsilon}\right).$$

Indeed, using that $\log(1/(1-x)) \geq x$ and $\gamma_1 = 1 - \frac{1}{2C_1\kappa_1}$,

$$\log_{(1/\gamma_1)}\left(\frac{9\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|_2^2}{\epsilon}\right) = \frac{\log\left(\frac{9\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|_2^2}{\epsilon}\right)}{\log\left(1/\gamma_1\right)} \leq 2C_1\kappa_1\log\left(\frac{9\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|_2^2}{\epsilon}\right).$$

Next recall Claim 98 from the proof of Lemma 52 which shows that

$$\mathbf{u}_1^{(T_1)} \leq \gamma_1^{T_1}\mathbf{u}_1^{(0)}.$$

The proof holds in this case as well and so we have that

$$\mathbf{u}_1^{(T_1)} \leq \gamma_1^{T_1}\mathbf{u}_1^{(0)} \leq \frac{\epsilon}{9\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|_2^2}.$$

Note that $\texttt{StochBSLSRes}_1(\mathbf{1}) = \texttt{StochBSLSRes}_2(\mathbf{u}^{(T_1)})$. So we have that if $\tilde{\mathbf{u}} = \texttt{StochBSLSRes}_1(\mathbf{1})$ then

$$\tilde{\mathbf{u}}_1 \leq \frac{\epsilon}{\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2}.$$

Finally we use that for any $j \geq 2$,

$$\tilde{\mathbf{u}}_j \leq \frac{L_1}{L_j}\mathbf{u}_1^{(T_1)} \leq \frac{\epsilon}{\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2}.$$

■

Finally we can combine the previous results to give the proof of Theorem 11.

Theorem 54 *Suppose Assumption 45 holds. For $i = 2, \ldots, m$ let $T_i = \lceil 8\kappa_i \log(\kappa_{\text{glob}}) \rceil$ and let $T_1 = \left\lceil 2\kappa_1 \log\left(9\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2/\epsilon\right)\right\rceil$. Let n_{max} denote the maximum number of eigenvalues lying in any single band region $[\mu_i, L_i]$ and further suppose*

$$n_{avg} \geq \mathsf{Kurt}(\mathcal{D}) m^2 n_{max} \left(\prod_{i\in[m]} T_i\right)\left(\max_{i\in[m]} T_i\right).$$

Then

$$\mathbb{E}\left[\left\|\left(\texttt{StochBSLS}_1(\mathbf{x}^{(0)}) - \mathbf{x}^\star\right)\right\|_2^2\right] \leq \epsilon.$$

Therefore since $\texttt{StochBSLS}_1$ requires $\mathcal{O}\left(n_{avg} 2^m \prod_{i\in[m]} T_i\right)$ queries of $(\mathbf{a}^{(i)}, b^{(i)})$ we conclude that $\texttt{StochBSLS}_1$ can return in expectation a ϵ-optimal solution with

$$\mathcal{O}\left(n_{max}\mathsf{Kurt}(\mathcal{D})\left(\prod_{i\in[m]}\kappa_i^2\right)\left(\max_{i\in[m]}\{\kappa_i\}\right)\log^{2m}(\kappa_{\text{glob}})\log^2\left(L_m\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2/\epsilon\right)\right)$$

first order queries.

Proof [Proof of Theorem 54] By Lemma 46 and Lemma 47 we have that for $N_1 = \prod_{i=1}^m (2T_i + 1)$ and for

$$\mathbf{D}^{(t)} := \mathbb{E}[(\mathbf{I} - \eta^{(N_1-t+1)}\mathbf{A}^{(N_1-t+1)})\mathbf{D}^{(t-1)}(\mathbf{I} - \eta^{(N_1-t+1)}\mathbf{A}^{(N_1-t+1)})] \qquad \mathbf{D}^{(0)} := \mathbf{I},$$

then

$$\mathbb{E}\left[\left\|\texttt{StochBSLS}_1(\mathbf{x}^{(0)}) - \mathbf{x}^\star\right\|_2^2\right] = (\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{D}^{(N_1)}(\mathbf{x}^{(0)} - \mathbf{x}^\star). \tag{E.9}$$

Recall that $\tilde{\mathbf{D}} := \mathbf{P}\mathbf{D}\mathbf{P}^\top$. As in Lemma 50 we define the following vector to represent the maximum entry of each $\tilde{\mathbf{D}}_i^{(t)}$,

$$\mathbf{r}_i^{(t)} := \left\|\tilde{\mathbf{D}}_i^{(t)}\right\|_\infty \qquad \mathbf{r}^{(0)} = \mathbf{1}.$$

By Lemma 50 we have $\mathbf{r}^{(0)} = \mathbf{1}$ and

$$\mathbf{r}^{(t)} \leq \mathbf{U}_{\eta^{(t)}}\mathbf{r}^{(t-1)}.$$

By Remark 51 it suffices to argue about the convergence of $\texttt{StochBSLSRes}(\mathbf{1})$. Applying Lemma 53 with error ϵ/L_m we have that

$$\left\|\texttt{StochBSLSRes}_1(\mathbf{1})\right\|_\infty \leq \frac{\epsilon}{L_m \left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2}.$$

From Remark 51 this implies that if $\mathbf{r}^{(N_1)}$ is our residuals vector at the end of $\texttt{StochBSLS}_1$ we have,

$$\left\|\mathbf{r}^{(N_1)}\right\|_\infty \leq \frac{\epsilon}{L_m \left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2}.$$

Therefore using that $\mathbf{P}$ is an orthonormal matrix,

$$\left\|\mathbf{D}^{(N_1)}\right\|_\infty = \left\|\mathbf{P}^\top \tilde{\mathbf{D}}^{(N_1)} \mathbf{P}\right\|_\infty \leq \left\|\tilde{\mathbf{D}}^{(N_1)}\right\|_\infty \leq \left\|\mathbf{r}^{(N_1)}\right\|_\infty \leq \frac{\epsilon}{L_m \left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2^2}.$$

Thus by Eq. E.9,

$$\mathbb{E}\left[\left\|\texttt{StochBSLS}_1(\mathbf{x}^{(0)}) - \mathbf{x}^\star\right\|_2^2\right] = (\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top \mathbf{D}^{(N_1)} (\mathbf{x}^{(0)} - \mathbf{x}^\star) \leq \epsilon/L_m.$$

Note that

$$\mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[\frac{1}{2}\left(\mathbf{a}^\top \mathbf{x} - b\right)^2\right] = \frac{1}{2}(\mathbf{x} - \mathbf{x}^\star)^\top \mathbf{\Sigma} (\mathbf{x} - \mathbf{x}^\star) \leq L_m \mathbb{E}\left[\|\mathbf{x} - \mathbf{x}^\star\|_2^2\right] \leq \epsilon.$$

This concludes the proof. ■

E.4. Setting where only m, μ_1, L_m, and $\prod_{i\in m} \kappa_i$ are known

To extend to this setting we have the following proposition, similar to Theorem 13,

Proposition 55 *Let $\pi_\kappa = \prod_{i\in m} \kappa_i$. Suppose with failure probability at most p, we can evaluate f up to a multiplicative constant factor C with $\tilde{T}(p, C)$ oracle queries; that is we can construct some $\hat{f}$ such that for any $\mathbf{x}$,*

$$f(\mathbf{x})/C \leq \hat{f}(\mathbf{x}) \leq C f(\mathbf{x}).$$

A randomized algorithm A which, in expectation, solves the stochastic multiscale optimization problem in Definition 4 to sub-optimality ϵ with $T(\pi_\kappa, \kappa_{\text{glob}}, m, \epsilon)$ gradient queries when the parameters (μ_i, L_i) are known, can be used along with the approximate function evaluation to solve the stochastic multiscale optimization problem with failure probability at most $C^2\epsilon + p$ with $\tilde{T}(p, C) \cdot T(\pi_\kappa 2^{5m}, \kappa_{\text{glob}}, m, \epsilon^2) \cdot O(\log^m(\kappa_{\text{glob}}))$ oracle queries when only m, μ_1, L_m and π_κ are known.

To apply this proposition to $\texttt{StochBSLS}$ requires guaranteeing first that $f(\texttt{StochBSLS}(\mathbf{x}^{(0)})) < \epsilon$ with good probability (for this we use Markov's inequality since we have bounded $\mathbb{E} f(\texttt{StochBSLS}(\mathbf{x}^{(0)}))$) and second that we can estimate $f(\mathbf{x})$ up to a multiplicative constant factor (this is why we must include the assumption from Eq. (3.1)). This results in the following corollary,

Corollary 56 *Assume the setting from Theorem 11 except that only m, μ_1, L_m and π_k are known. Suppose that $\mathcal{D}$ is such that for $\mathbf{a} \sim \mathcal{D}$ there exists some K where*

$$\left\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\right\|_2 \leq K\left(\mathbb{E}\left\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\right\|_2^2\right)^{1/2}.$$

Then with failure probability at most δ, `StochBSLS` can be used to solve the stochastic quadratic multiscale optimization problem from Definition 4 with $\tilde{\mathcal{O}}(d)$ space and an extra multiplicative factor of $\mathcal{O}\left(K^2 d \log\frac{4d}{\delta}\left(1+\sqrt{\varepsilon/\delta}\right)\right)$ queries of $(\mathbf{a}, b) \sim \mathcal{D}$.

The proofs of Theorem 55 and Theorem 56 are in Appendix F.1.

Appendix F. Extended results regarding the multiscale optimization problem

F.1. Black-box reduction from unknown (μ_i, L_i) to known (μ_i, L_i)

In this subsection, we show that the assumption in `BSLS` that the μ_i, L_i parameters are known is essentially without loss of generality, since we can reduce from the case where these are unknown to the case where they are known without changing the asymptotic complexity. The reduction is black-box and does not utilize any special properties of our algorithm.

Proposition 57 (Restated Theorem 13) *Let $\pi_\kappa = \prod_{i\in m}\kappa_i$. Suppose with failure probability at most p, we can evaluate f up to a multiplicative constant factor C with $\tilde{T}(p, C)$ oracle queries; that is we can construct some $\hat{f}$ such that for any $\mathbf{x}$,*

$$f(\mathbf{x})/C \leq \hat{f}(\mathbf{x}) \leq Cf(\mathbf{x}).$$

A randomized algorithm A which, in expectation, solves the stochastic multiscale optimization problem in Definition 4 to sub-optimality ϵ with $T(\pi_\kappa, \kappa_{\text{glob}}, m, \epsilon)$ gradient queries when the parameters (μ_i, L_i) are known, can be used along with the approximate function evaluation to solve the stochastic multiscale optimization problem with failure probability at most $C^2\epsilon + p$ with $\tilde{T}(p, C)\cdot T(\pi_\kappa 2^{5m}, \kappa_{\text{glob}}, m, \epsilon^2)\cdot O(\log^m(\kappa_{\text{glob}}))$ oracle queries when only m, μ_1, L_m and π_κ are known.

Proof Let $\{(\mu_i, L_i), i \in [m]\}$ be the original parameters of the multiscale optimization problem. The proof relies on a simple brute force search over these parameters over a suitable grid. In the first step, we will do a brute force search for the parameters $\kappa_i \;\forall\; i \in [m]$. Then, we do a brute force search over the the parameters $(\mu_i, L_i)\;\forall\; i \in [m]$ and run the algorithm with every instance of these parameters. One of these choices will be guaranteed to work because of the guarantees of the algorithm. The full procedure is given in Algorithm 6.

Algorithm 6 Brute force search over (μ_i, L_i) parameters

Procedure Search(m, μ_1, L_m, π_κ, ϵ)

$\pi_{\kappa,\log} \leftarrow \lceil \log_2(\pi_\kappa) \rceil + 4m$

$\mu_{1,\log} \leftarrow \lfloor \log_2(\mu_1) \rfloor$, $L_{m,\log} \leftarrow \lceil \log_2(L_m) \rceil$

for all $\{\kappa_{i,\log} : i \in [m], \kappa_{i,\log} \in [\pi_{\kappa,\log}]\}$ such that $\sum_{i=1}^m \kappa_{i,\log} = \pi_{\kappa,\log}$ **do**

  for all $\{\mu_{i,\log} : i \in \{2, \dots, m\}, \mu_{i,\log} \in \{\mu_{1,\log}, \dots, L_{m,\log}\}, \mu_{i,\log} \leq \mu_{i+1,\log} \; \forall \; i \leq m-1\}$ **do**

    $\forall \; i \in [m]$, $L_{i,\log} \leftarrow \mu_{i,\log} + \kappa_{i,\log}$

    $(m', \{\mu_{i,\log}, L_{i,\log}, i \in [m']\}) \leftarrow$ MergeOverlapping$(m, \{\mu_{i,\log}, L_{i,\log}, i \in [m]\})$

    $\forall \; i \in [m']$, $\mu_{i'} \leftarrow 2^{\mu_{i,\log}}$, $L_{i'} \leftarrow 2^{L_{i',\log}}$.

    $\pi'_\kappa \leftarrow 2^{\pi_{\kappa,\log}}$

    Let $\mathbf{x}'$ be result of running Algorithm A with parameters $\{(\mu_{i'}, L_{i'}), i \in [m']\}$ for $T(\pi'_\kappa, \kappa_{\text{glob}}, m', \epsilon)$ gradient steps, and ϵ' be the function error of $\mathbf{x}'$.

    if $\epsilon' < \epsilon$ **then**

      return $\mathbf{x}'$.

return $\emptyset$.

Procedure MergeOverlapping$(m, \{\mu_{i,\log}, L_{i,\log}, i \in [m]\})$

$m' \leftarrow m$

for all $i \in [m'-1]$ **do**

  if $L_{i,\log} \geq \mu_{i+1,\log}$ **then**

    $L_{i,\log} \leftarrow L_{i+1,\log}$

    for all $i+1 \leq i' \leq m-1$ **do**

      $\mu_{i',\log} \leftarrow \mu_{i'+1,\log}$

      $L_{i',\log} \leftarrow L_{i'+1,\log}$

    $m' \leftarrow m' - 1$

return $(m', \{\mu_{i,\log}, L_{i,\log}, i \in [m']\})$

We first remark that at least one of the runs of Algorithm A has the property that for all $i \in [m]$ there exists some $i' \in [m']$ such that $\mu_{i'} \leq \mu_i$ and $L_{i'} \geq L_i$, i.e. the original function $f(\mathbf{x})$ is a multiscale optimization problem with parameters $\{(\mu_{i'}, L_{i'}), i \in [m']\}$. Note that it is sufficient to show that this is true for the choice of parameters before the MergeOverlapping function is called, since the MergeOverlapping function will preserve this property. To verify that the property is true before the MergeOverlapping function is called, note that one of the choices in the brute force search satisfies (a) $\forall \; i \in [m]$, $\lceil \log_2(\kappa_i) \rceil + 1 \leq \kappa_{i,\log}$, (b) $\forall \; i \in [m], \mu_{i,\log} = \lfloor \log_2(\mu_i) \rfloor$. (a) and (b) together ensure that $L_{i,\log} \geq \log_2(L_i)$, which verifies that $\forall \; i \in [m], \mu_{i,\log} \leq \log_2(\mu_i)$ and $L_{i,\log} \geq \log_2(L_i)$.

Finally, we claim that Algorithm 6 runs with at most $T(\pi_\kappa 2^{5m}, \kappa_{\text{glob}}, m, \epsilon) \cdot O(\log^m(\kappa_{\text{glob}}))$ gradient evaluations. This follows because (a) each run of Algorithm A runs for $T(\pi'_\kappa, \kappa_{\text{glob}}, m', \epsilon)$ steps where $\pi'_\kappa \leq 2^{5m}\pi_\kappa$ and $m' \leq m$, and (b) there are at most $O(\log^m(\kappa_{\text{glob}}))$ choices for the brute force search over the parameters.

■

Next we extend Theorem 13 to the stochastic setting. Recall Theorem 55, restated here for convenience:

Proposition 58 (Restated Theorem 55) *Let $\pi_\kappa = \prod_{i \in m} \kappa_i$. Suppose with failure probability at most p, we can evaluate f up to a multiplicative constant factor C with $\tilde{T}(p, C)$ oracle queries; that is we can construct some $\hat{f}$ such that for any $\mathbf{x}$,*

$$f(\mathbf{x})/C \leq \hat{f}(\mathbf{x}) \leq Cf(\mathbf{x}).$$

A randomized algorithm A which, in expectation, solves the stochastic multiscale optimization problem in Definition 4 to sub-optimality ϵ with $T(\pi_\kappa, \kappa_{\text{glob}}, m, \epsilon)$ gradient queries when the parameters (μ_i, L_i) are known, can be used along with the approximate function evaluation to solve the stochastic multiscale optimization problem with failure probability at most $C^2\epsilon + p$ with $\tilde{T}(p, C) \cdot T(\pi_\kappa 2^{5m}, \kappa_{\text{glob}}, m, \epsilon^2) \cdot O(\log^m(\kappa_{\text{glob}}))$ oracle queries when only m, μ_1, L_m and π_κ are known.

Proof [Proof of Theorem 55] Consider Algorithm 6 with the change in line 9 of $Search$ that the function error is estimated using $\hat{f}$. First note that if Search(m, μ_1, L_m, π_κ, ϵ^2) returns any $\mathbf{x}'$, it satisfies that $\hat{f}(\mathbf{x}) < \epsilon/C$ and so with failure probability at most p $f(\mathbf{x}') < \epsilon$. Next we want to show it will return an $\mathbf{x}'$ with probability at least $1 - C^2\epsilon$. By the proof of Theorem 13 we know there is at least one run of algorithm A with parameters $\{(\mu_i', L_i'), i \in [m']\}$ such that the original function $f(\mathbf{x})$ is a multiscale optimization problem with respect to these parameters and $\pi_\kappa' \leq 2^{5m}\pi_\kappa$. Thus with $T(\pi_\kappa', \kappa_{\text{glob}}, m', \epsilon^2)$ many oracle queries, algorithm A returns some $\mathbf{x}'$ such that in expectation (over the randomness of the algorithm's output $\mathbf{x}$) $f(\mathbf{x}') < \epsilon^2$. Then by Markov's Inequality,

$$P(f(\mathbf{x}') \geq \epsilon/C^2) \leq \frac{\mathbb{E}[f(\mathbf{x}')]}{\epsilon/C^2} \leq C^2\epsilon.$$

Therefore with failure probability at most $C^2\epsilon$, $f(\mathbf{x}') < \epsilon/C^2$. Then since for any $\mathbf{x}$,

$$f(\mathbf{x})/C \leq \hat{f}(\mathbf{x}) \leq Cf(\mathbf{x}),$$

we have that with $\hat{f}(\mathbf{x}') < \epsilon/C$ and so Search(m, μ_1, L_m, π_κ, ϵ^2) will return this $\mathbf{x}'$ if it hasn't already returned another $\mathbf{x}'$. ■

Next recall Theorem 56, restated here for convenience:

Corollary 59 (Restated Theorem 56) *Assume the setting from Theorem 11 except that only m, μ_1, L_m and π_k are known. Suppose that $\mathcal{D}$ is such that for $\mathbf{a} \sim \mathcal{D}$ there exists some K where*

$$\left\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\right\|_2 \leq K\left(\mathbb{E}\left\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\right\|_2^2\right)^{1/2}.$$

Then with failure probability at most δ, `StochBSLS` can be used to solve the stochastic quadratic multiscale optimization problem from Definition 4 with $\tilde{\mathcal{O}}(d)$ space and an extra multiplicative factor of $\mathcal{O}\left(K^2 d \log\frac{4d}{\delta}\left(1 + \sqrt{\varepsilon/\delta}\right)\right)$ queries of $(\mathbf{a}, b) \sim \mathcal{D}$.

In the proof of Theorem 56 we will make use of the following Theorem 5.6.1 from Vershynin (2019) which we state for the reader's convenience.

Theorem 60 *Let* $\mathbf{x}$ *be a random vector in* $\mathbb{R}^d$, $d \geq 2$. *Let* $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ *and* $\hat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{i\in n}\mathbf{x}_i\mathbf{x}_i^\top$ *for i.i.d.* $\mathbf{x}_i$. *Assume that for some* $K \geq 1$,

$$\|\mathbf{x}\|_2 \leq K(\mathbb{E}[\|\mathbf{x}\|_2^2])^{1/2} \text{ almost surely.}$$

Then, for every positive integer n *and any* $t \geq 0$,

$$\left\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\| \leq \left(\sqrt{\frac{K^2 d(\log d + t)}{n}} + \frac{2K^2 d(\log d + t)}{n}\right)\|\mathbf{\Sigma}\|,$$

with probability at least $1 - 2e^{-t}$.

Proof [Proof of Theorem 56] We will show that for $\tilde{T}(p, C) = \frac{8K^2 d\log(2d/p)}{\left(1-\frac{1}{C}\right)^2}$ oracle queries of $(\mathbf{a}, b) \sim \mathcal{D}$ we can construct $\hat{f}$ to estimate f up to multiplicative error C. Recall

$$f(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[(\mathbf{a}^\top\mathbf{x} - b\right] = \frac{1}{2}(\mathbf{x} - \mathbf{x}^\star)^\top\mathbf{\Sigma}(\mathbf{x} - \mathbf{x}^\star).$$

We construct $\hat{f}_n(\mathbf{x})$ as

$$\hat{f}_n(\mathbf{x}) \coloneqq \frac{1}{2n}\sum_{i\in n}\frac{1}{2}(\mathbf{a}_i^\top\mathbf{x} - b_i)^2 = \frac{1}{2}(\mathbf{x} - \mathbf{x}^\star)^\top\hat{\mathbf{\Sigma}}_n(\mathbf{x} - \mathbf{x}^\star).$$

For $\mathbf{a} \sim \mathcal{D}$, consider the random vector $\tilde{\mathbf{a}} = \mathbf{\Sigma}^{-1/2}\mathbf{a}$. Note that by assumption

$$\left\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\right\|_2 \leq K\left(\mathbb{E}\left\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\right\|_2^2\right)^{1/2},$$

almost surely. Suppose for $C \geq 1$

$$n = \frac{8K^2 d\log(2d/p)}{\left(1 - \frac{1}{C}\right)^2}.$$

Then by Theorem 60, with failure probability at most p,

$$\left\|\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}_n\mathbf{\Sigma}^{-1/2} - \mathbf{I}\right\| \leq \frac{1 - \frac{1}{C}}{2} + \frac{\left(1 - \frac{1}{C}\right)^2}{4} \leq 1 - \frac{1}{C}. \tag{F.1}$$

Note that Eq. (F.1) holds if and only if

$$\frac{1}{C}\mathbf{\Sigma} \preceq \hat{\mathbf{\Sigma}}_n \preceq \left(2 + \frac{1}{C}\right).$$

Therefore for $C \geq 3$ since $2 + \frac{1}{C} \leq C$,

$$\frac{1}{C}f(\mathbf{x}) \leq \hat{f}(\mathbf{x}) \leq Cf(\mathbf{x}).$$

To conclude the proof of Theorem 56 we simply recall Theorem 11 and apply Theorem 55. ■

F.2. If decomposition is known, then can solve with $\tilde{\mathcal{O}}(\sum_{i\in[m]}\sqrt{\kappa_i})$ gradient queries

In this subsection we show that when the gradient of sub-objectives are known, then the multiscale optimization problem (Theorem 1) can be solved in $\tilde{\mathcal{O}}(\sum_{i\in[m]}\sqrt{\kappa_i})$ queries. To prove this claim, consider the algorithm that run accelerated gradient descent on each sub-objective f_j independently. Since each sub-objective f_j takes $\mathcal{O}(\sqrt{\kappa_j}\log(m/\epsilon))$ to converge to $\frac{\epsilon}{m}$-optimality, we only need a total of $\sum_{j\in[m]}\mathcal{O}(\sqrt{\kappa_j}\log(m/\epsilon))$ gradients for f to converge to ϵ-optimality.

However, as we will see in the next subsection (Appendix F.3), recovering the projection $\mathbf{P}_i$ is costly.

F.3. Recovering the projections $\mathbf{P}_i$ is costly

In this subsection we show that recovering the projections $\mathbf{P}_i$ in the multiscale optimization problem (Theorem 1) requires $\Omega(d)$ gradient evaluations in the worst-case.

Proposition 61 *Consider the multiscale optimization problem in Theorem 1 for $m = 2$. There exist functions f_1, f_2 such that recovering $\mathbf{P}_1, \mathbf{P}_2$ requires $\Omega(d)$ gradient evaluations in the worst-case, even if f_1, f_2 are known.*

Proof [Proof of Proposition 61] Let $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{R}^{d/4\times d}$. Consider the following multiscale optimization problem:

$$f_1(\mathbf{x}) = \|\mathbf{P}_1\mathbf{x}\|^2, \quad f_2(\mathbf{x}) = (1/\kappa_{\text{glob}})\|\mathbf{P}_2\mathbf{x}\|^2, \quad f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}).$$

Note that by our choice of f_1 and f_2, $\kappa_1 = \kappa_2 = 1$ and κ_{glob} is the overall condition number. Now we can write,

$$\nabla f(\mathbf{x}) = 2\mathbf{P}_1^\top\mathbf{P}_1\mathbf{x} + (2/\kappa_{\text{glob}})\mathbf{P}_2^\top\mathbf{P}_2\mathbf{x}.$$

Let S_1, S_2 be the $d/4$ dimensional subspaces spanned by $\mathbf{P}_1^\top$ and $\mathbf{P}_2^\top$ respectively. Note that $\mathbf{P}_1^\top\mathbf{P}_1\mathbf{x}$ is the projection of $\mathbf{x}$ onto S_1, and similarly for S_2. Therefore $\nabla f(\mathbf{x})$ is a linear combination of the projections onto S_1 and S_2, and with every gradient evaluation we get a single vector in the span of S_1 and S_2. Since the union of S_1, S_2 is a $d/2$ dimensional subspace, any algorithm needs $d/2$ vectors from the subspace to learn it. Hence any algorithm needs at least $d/2$ gradient evaluations to learn S_1, S_2, and hence also to learn $\mathbf{P}_1, \mathbf{P}_2$. We note that it could be possible to extend this argument to approximately learning the subspaces S_1 and S_2 using bounds on quantization on the Grassmann manifold Dai et al. (2007). ■

F.4. GD with exact line search or constant step-sizes cannot match the guarantee of BSLS

In this subsection we prove that gradient descent with exact line search or any constant step-size cannot match the guarantee of BSLS (Algorithm 1) given by Theorem 6.

Formally, gradient descent with exact line search has the form:

$$\mathbf{x}^{(t+1)} \leftarrow \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) \middle| \mathbf{x} = \mathbf{x}^{(t)} - \eta\nabla f(\mathbf{x}^{(t)}) \text{ for some } \eta \in \mathbb{R} \right\}.$$

Proposition 62 *Consider the objective* $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$ *with*

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \qquad \mathbf{b} = \mathbf{1}.$$

Let $\lambda_1 < \lambda_2$ *and note that* $\kappa_{\text{glob}} = \lambda_2/\lambda_1$. *Then*

(a) Gradient descent with exact line search initialized at $\mathbf{x}^{(0)} = \mathbf{0}$ *requires at least* $\left\lfloor \frac{\kappa_{\text{glob}}}{8} \log\left(\frac{(1/\lambda_1)+(1/\lambda_2)}{2\epsilon}\right)\right\rfloor$ *gradient queries to attain* ϵ*-optimality.*

(b) If $\kappa_{\text{glob}} \geq 2$, *gradient descent with a constant step size requires at least* $\left\lfloor \frac{\kappa_{\text{glob}}}{2} \log\left(\frac{1}{\lambda_1 \epsilon}\right)\right\rfloor$ *gradient queries to attain* ϵ*-optimality.*

Remark 63 *Theorem 6 states that* `BSLS` *only requires* $(2\log(\kappa_{\text{glob}}+1)\log\left(\frac{(1/\lambda_1)+(1/\lambda_2)}{2\epsilon}\right)$ *gradient queries which compares favorable to gradient descent with exact line search and constant stepsize.*

Proof [Proof of Proposition 62 - Exact Line Search] Assume we initialize $\mathbf{x}^{(0)} = \mathbf{0}$. For exact line search we update $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - s_t \mathbf{g}^{(t)}$ with $s_t = \frac{\mathbf{g}^{(t)\top}\mathbf{g}^{(t)}}{\mathbf{g}^{(t)\top}\mathbf{A}\mathbf{g}^{(t)}}$. It is a fact that

$$f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^\star) = \left(1 - \frac{1}{\kappa_t}\right)\left(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star)\right), \tag{F.2}$$

where if $\mathbf{g}^{(t)} := \nabla f(\mathbf{x}^{(t)})$,

$$\kappa_t := \frac{\mathbf{g}^{(t)\top}\mathbf{A}\mathbf{g}^{(t)}}{\mathbf{g}^{(t)\top}\mathbf{g}^{(t)}} \frac{\mathbf{g}^{(t)\top}\mathbf{A}^{-1}\mathbf{g}^{(t)}}{\mathbf{g}^{(t)\top}\mathbf{g}^{(t)}}.$$

Our goal is to show that a large portion of the gradients $\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(T)}$ are such that κ_t is close to κ. Note $\mathbf{g}^{(t)} = \mathbf{A}(\mathbf{x}^{(t)} - \mathbf{x}^\star)$. For simplicity define

$$u_t := \mathbf{g}_1^{(t)} = \lambda_1\left(\mathbf{x}_1^{(t-1)} - \mathbf{x}_1^\star\right) \qquad v_t := \mathbf{g}_2^{(t)} = \lambda_2\left(\mathbf{x}_2^{(t-1)} - \mathbf{x}_2^\star\right). \tag{F.3}$$

We can rewrite κ_t as,

$$\kappa_t = \frac{(\lambda_1 u_t^2 + \lambda_2 v_t^2)}{u_t^2 + v_t^2}\frac{\left(\frac{1}{\lambda_1}u_t^2 + \frac{1}{\lambda_2}v_t^2\right)}{u_t^2 + v_t^2} = \frac{u_t^4 + v_t^4 + \kappa_{\text{glob}} u_t^2 v_t^2 + \frac{1}{\kappa_{\text{glob}}}u_t^2 v_t^2}{(u_t^2 + v_t^2)^2} = \frac{\left(\frac{u_t^2}{v_t^2} + \frac{v_t^2}{u_t^2}\right) + \kappa_{\text{glob}} + \frac{1}{\kappa_{\text{glob}}}}{\left(\frac{u_t^2}{v_t^2} + \frac{v_t^2}{u_t^2}\right) + 2}. \tag{F.4}$$

This motivates us to understand the ratio u_t^2/v_t^2 or rather $\left(\mathbf{x}_1^{(t)} - \mathbf{x}_1^\star\right)^2 / \left(\mathbf{x}_2^{(t)} - \mathbf{x}_2^\star\right)^2$. Since

$$\mathbf{x}^{(t+1)} - \mathbf{x}^\star = (\mathbf{I} - s_t\mathbf{A})\mathbf{x}^{(t)},$$

we have

$$\frac{\left(\mathbf{x}_1^{(t)} - \mathbf{x}_1^\star\right)^2}{\left(\mathbf{x}_2^{(t)} - \mathbf{x}_2^\star\right)^2} = \frac{\prod_{\ell=1}^t (1 - s_\ell\lambda_1)^2\left(\mathbf{x}_1^{(0)} - \mathbf{x}_1^\star\right)^2}{\prod_{\ell=1}^t (1 - s_\ell\lambda_2)^2\left(\mathbf{x}_2^{(0)} - \mathbf{x}_2^\star\right)^2}. \tag{F.5}$$

Define

$$p_t \coloneqq \prod_{\ell=1}^{t} \frac{(1-s_\ell\lambda_1)^2}{(1-s_\ell\lambda_2)^2}.$$

Notice that

$$\mathbf{g}^{(t+1)} = (\mathbf{I} - s_t\mathbf{A})\mathbf{g}^{(t)}.$$

Therefore since $\mathbf{g}^{(0)=} - \mathbf{b} = -\mathbf{1}$,

$$\begin{aligned}
s_{t+1} &= \frac{\mathbf{g}_1^{(t+1)2} + \mathbf{g}_2^{(t+1)2}}{\lambda_1\mathbf{g}_1^{(t+1)2} + \lambda_2\mathbf{g}_2^{(t+1)2}} \\
&= \frac{\left(\prod_{\ell=1}^{t}(1-s_\ell\lambda_1)^2\right) + \left(\prod_{\ell=1}^{t}(1-s_\ell\lambda_2)^2\right)}{\lambda_1\left(\prod_{\ell=1}^{t}(1-s_\ell\lambda_1)^2\right) + \lambda_2\left(\prod_{\ell=1}^{t}(1-s_\ell\lambda_2)^2\right)} \\
&= \frac{(p_t+1)\left(\prod_{\ell=1}^{t}(1-s_\ell\lambda_2)^2\right)}{(\lambda_1 p_t + \lambda_2)\left(\prod_{\ell=1}^{t}(1-s_\ell\lambda_2)^2\right)} \\
&= \frac{(p_t+1)}{(\lambda_1 p_t + \lambda_2)}.
\end{aligned}$$

Therefore,

$$\frac{1-s_t\lambda_1}{1-s_t\lambda_2} = \frac{1-\lambda_1\frac{(p_t+1)}{(\lambda_1 p_t+\lambda_2)}}{1-\lambda_2\frac{(p_t+1)}{(\lambda_1 p_t+\lambda_2)}} = \frac{(\lambda_1 p_t+\lambda_2)-\lambda_1(p_t+1)}{(\lambda_1 p_t+\lambda_2)-\lambda_2(p_t+1)} = \frac{\lambda_2-\lambda_1}{p_t(\lambda_1-\lambda_2)} = -\frac{1}{p_t}.$$

Then since

$$p_{t+1} = p_t\left(\frac{1-s_t\lambda_1}{1-s_t\lambda_2}\right)^2,$$

we have

$$p_{t+1} = \frac{1}{p_t}.$$

Finally since $s_1 = 2/(\lambda_1+\lambda_2)$ we have $p_1 = 1$ and therefore for any t, $p_t = 1$. Thus, recalling Eq. F.5 we have

$$\frac{\left(\mathbf{x}_1^{(t)} - \mathbf{x}_1^\star\right)^2}{\left(\mathbf{x}_2^{(t)} - \mathbf{x}_2^\star\right)^2} = \frac{\left(\mathbf{x}_1^{(0)} - \mathbf{x}_1^\star\right)^2}{\left(\mathbf{x}_2^{(0)} - \mathbf{x}_2^\star\right)^2} = \frac{\mathbf{x}_1^{*2}}{\mathbf{x}_2^{*2}} = \kappa_{\text{glob}}^2.$$

Recalling the definitions of u_t and v_t in Eq. F.3 we have

$$\frac{u_t}{v_t} = \frac{\lambda_1^2\left(\mathbf{x}_1^{(t-1)} - \mathbf{x}_1^\star\right)^2}{\lambda_2^2\left(\mathbf{x}_2^{(t-1)} - \mathbf{x}_2^\star\right)^2} = 1.$$

Finally, recalling Eq. F.4 we have

$$\kappa_t = \frac{2 + \kappa_{\text{glob}} + \frac{1}{\kappa_{\text{glob}}}}{4} \geq \kappa_{\text{glob}}/4.$$

Therefore we can lower bound the progress made by exact line search. Using Eq. F.2 we find

$$f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^\star) \geq \left(1 - \frac{4}{\kappa_{\text{glob}}}\right)\left(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star)\right).$$

Thus exact line search requires at least $\left\lfloor \frac{\kappa_{\text{glob}}}{8} \log\left(\frac{f(\mathbf{0})-f(\mathbf{x}^\star)}{\epsilon}\right)\right\rfloor$ gradient queries. Since $f(\mathbf{0}) - f(\mathbf{x}^\star) = \frac{1}{2}((1/\lambda_1)+(1/\lambda_2))$ we conclude exact line search requires at least $\left\lfloor \frac{\kappa_{\text{glob}}}{8} \log\left(\frac{(1/\lambda_1)+(1/\lambda_2)}{2\epsilon}\right)\right\rfloor$ gradient queries. ■

Proof [Proof of Proposition 62 - Constant step-sizes] We make use of the equality

$$f(\mathbf{x}) - f(\mathbf{x}^\star) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{A}}^2, \tag{F.6}$$

where $\|\mathbf{x}\|_{\mathbf{A}}^2 \coloneqq \mathbf{x}^\top \mathbf{A}\mathbf{x}$. Since

$$\nabla f(\mathbf{x}) = \mathbf{A}(\mathbf{x} - \mathbf{x}^\star),$$

we have that the gradient descent algorithm with constant stepsize α produces the recursion,

$$\mathbf{x}^{(t+1)} - \mathbf{x}^\star = \mathbf{x}^{(t)} - \alpha\nabla f(\mathbf{x}) = (\mathbf{I} - \alpha\mathbf{A})(\mathbf{x}^{(t)} - \mathbf{x}^\star) = (\mathbf{I} - \alpha\mathbf{A})^{t+1}(\mathbf{x}^{(0)} - \mathbf{x}^\star).$$

Therefore using Eq. F.6 we find,

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star) = (\mathbf{x}^{(0)} - \mathbf{x}^\star)^\top(\mathbf{I} - \alpha\mathbf{A})^t\mathbf{A}(\mathbf{I} - \alpha\mathbf{A})^t(\mathbf{x}^{(0)} - \mathbf{x}^\star).$$

Using the definition of $\mathbf{A}$, the fact that $\mathbf{x}^{(0)} = \mathbf{0}$, and finally the fact that $\mathbf{x}^\star = \begin{bmatrix}1/\lambda_1 & 1/\lambda_2\end{bmatrix}^\top$, we have

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star) = \frac{1}{\lambda_1}(1 - \alpha\lambda_1)^{2t} + \frac{1}{\lambda_2}(1 - \alpha\lambda_2)^{2t}.$$

In order to have the function error decrease we must choose $\alpha \in [0, 2/\lambda_2]$. For α in this range we have,

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star) \geq \frac{1}{\lambda_1}(1 - \alpha\lambda_1)^{2t} \geq \frac{1}{\lambda_1}\left(1 - \frac{2\lambda_1}{\lambda_2}\right)^{2t}.$$

Suppose $t < \frac{\kappa_{\text{glob}}}{4}\log\left(\frac{1}{2\lambda_1\epsilon}\right)$. Using that $\left(1 - \frac{1}{x}\right)^x \geq \frac{1}{2}$ for all $x \geq 2$ and our assumption that $\kappa_{\text{glob}} \geq 2$ we have

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star) \geq \frac{1}{\lambda_1}\left(1 - \frac{2\lambda_1}{\lambda_2}\right)^{2t} \geq \frac{1}{\lambda_1}\left(1 - \frac{2\lambda_1}{\lambda_2}\right)^{\frac{\kappa_{\text{glob}}}{2}\log\left(\frac{1}{2\lambda_1\epsilon}\right)} \geq 2\epsilon.$$

This concludes the proof. ■

F.5. The orthogonality assumption in Theorem 1 is necessary

In this section we show that without the assumption in Theorem 1 that

$$\mathbf{P}_i\mathbf{P}_j^\top = \begin{cases} \mathbf{I}_{d_i\times d_i} & \text{if } i=j, \\ \mathbf{0}_{d_i\times d_j} & \text{otherwise,} \end{cases}$$

Theorem 2 does not necessarily hold. Indeed, in Theorem 64 we show that any first-order method (even if randomized) must have query complexity at least $\Omega(\sqrt{\kappa_{\text{glob}}}/\operatorname{polylog}(d))$.

Proposition 64 *There exists a distribution over instances*

$$f(\mathbf{x}) = \sum_{i=1}^{4} f_i(\mathbf{P}_i\mathbf{x}),$$

where $\mathbf{P}_i \in \mathbb{R}^{d/2\times d}$ *is such that*

$$\mathbf{P}_i\mathbf{P}_j^\top \neq \mathbf{0}$$

for $i\neq j$ *and each* f_i *is well conditioned with* $\kappa_i \le 10$ *and* $\kappa_{\text{glob}} = \Theta(d^2)$ *such that any first-order method (even if randomized) which returns a* $\frac{1}{10\kappa_{\text{glob}}}$*-optimal solution with probability at least* 0.9 *needs at least* $\Omega(\sqrt{\kappa_{\text{glob}}}/\operatorname{polylog}(d))$ *first-order queries.*

Proof We use the following result from Braverman et al. (2020) which establishes a hardness result for solving linear systems. Let $\kappa(\mathbf{M})$ denote the condition number of any matrix $\mathbf{M}$.

Theorem 65 *(Theorem 6 of Braverman et al. (2020)) Let* d_0 *be a universal constant. Let* $d\ge d_0$ *be any ambient dimension and let* $\mathcal{A}$ *be any linear system algorithm. Suppose* $\mathcal{A}$ *is such that for all positive semi-definite matrices* $\mathbf{A}$ *with condition number* $\kappa(\mathbf{A})\le d^2$ *and for all initial vectors* $\mathbf{x}_0\in\mathbb{R}^d$ *and* $\mathbf{b}\in\mathbb{R}^d$*,*

$$\Pr\left[\left\|\hat{\mathbf{x}}-\mathbf{A}^{-1}\mathbf{b}\right\|_{\mathbf{A}}^2 \le \frac{1}{10d^2}\right] \ge 1-\frac{1}{e}.$$

Then $\mathcal{A}$ *must have query complexity at least* $\Omega(\kappa(\mathbf{A})/\operatorname{polylog}(d))$.

We prove our lower bound by showing that the hard instance in Braverman et al. (2020) admits a decomposition as a multi-scale optimization problem.

The hard distribution over matrices $\mathbf{A}$ is $\mathbf{A} = (\gamma-1)\mathbf{I}+(1/5)\mathbf{W}$, where $\mathbf{W}$ is sampled from the Wishart distribution, i.e. $\mathbf{W}=\mathbf{X}\mathbf{X}^\top$ where $\mathbf{X}\in\mathbb{R}^{d\times d}$ and $\mathbf{X}_{i,j}$ is distribution as i.i.d. $N(0,1/d)$, and $\gamma = 1+\Theta(1/d^2)$. We show that with high probability $\mathbf{A}$ admits a simple decomposition into the sum of four well-conditioned matrices, hence proving our bound.

Let $\mathbf{X}_1\in\mathbb{R}^{d\times d/2}$ be the submatrix of $\mathbf{X}$ corresponding to its first $d/2$ columns and $\mathbf{W}_1 = \mathbf{X}_1\mathbf{X}_1^\top$. Similarly, let $\mathbf{X}_2\in\mathbb{R}^{d\times d/2}$ be the submatrix of $\mathbf{X}$ corresponding to its last $d/2$ columns and $\mathbf{W}_2=\mathbf{X}_2\mathbf{X}_2^\top$. Note that $\mathbf{W}=\mathbf{W}_1+\mathbf{W}_2$, therefore,

$$\mathbf{A} = \underbrace{\frac{\gamma-1}{2}\mathbf{I}+\frac{1}{5}\mathbf{W}_1}_{\mathbf{U}_1} + \underbrace{\frac{\gamma-1}{2}\mathbf{I}+\frac{1}{5}\mathbf{W}_2}_{\mathbf{U}_2}.$$

Let $\mathcal{E}_1$ be the event that all the non-zero eigenvalues of $\mathbf{W}_1$ and $\mathbf{W}_2$ lie in the interval $[0.2, 2]$. By concentration bounds for the spectrum of Wishart matrices (see for example Corollary 5.35 of Vershynin (2010)), for sufficiently large d, $\mathcal{E}_1$ happens with probability at least 0.99. Let$\mathcal{E}_2$ be the event that $\kappa(\mathbf{A} = \Theta(d^2)$. By similar concentration bounds for Wishart matrices (for example Corollary 12 of Braverman et al. (2020)), $\mathcal{E}_2$ happens with probability at least 0.99. We condition on the events $\mathcal{E}_1$ and $\mathcal{E}_2$ for the rest of the proof.

Let $\mathbf{W}_1 = \mathbf{P}_1^\top \mathbf{\Sigma}_1 \mathbf{P}_1$ denote the singular-value decomposition of $\mathbf{W}_1$. Let $\mathbf{P}_2 \in \mathbb{R}^{d/2 \times d}$ be the matrix whose rows form an orthonormal basis for the orthogonal space to the column space of $\mathbf{W}_1$. Then we can decompose $\mathbf{U}_1$ as $\mathbf{U}_1 = \mathbf{A}_1 + \mathbf{A}_2$, where $\mathbf{A}_1 = \mathbf{P}_1^\top (\frac{\gamma-1}{2}\mathbf{I} + \frac{1}{5}\mathbf{\Sigma}_1)\mathbf{P}_1$ and $\mathbf{A}_2 = \frac{\gamma-1}{2}\mathbf{P}_2^\top \mathbf{P}_2$. Let $\tilde{\mathbf{A}}_1 = \frac{\gamma-1}{2}\mathbf{I} + \frac{1}{5}\mathbf{\Sigma}_1$ and $\tilde{\mathbf{A}}_1 = \frac{\gamma-1}{2}\mathbf{I}$. Note that the eigenvalues of $\tilde{\mathbf{A}}_1$ lie in the interval $[0.2 + \frac{\gamma-1}{2}, 2 + \frac{\gamma-1}{2}]$ and all eigenvalues of $\tilde{\mathbf{A}}_2$ are $\frac{\gamma-1}{2}$. Therefore, $\kappa(\tilde{\mathbf{A}}_1) \leq 10$, and $\kappa(\tilde{\mathbf{A}}_2) = 1$.

Similarly, let $\mathbf{W}_2 = \mathbf{P}_3^\top \mathbf{\Sigma}_3 \mathbf{P}_3$ denote the singular-value decomposition of $\mathbf{W}_2$. Let $\mathbf{P}_4 \in \mathbb{R}^{d/2 \times d}$ be the matrix whose rows form an orthonormal basis for the orthogonal space to the column space of $\mathbf{W}_2$. Then we can decompose $\mathbf{U}_2$ as $\mathbf{U}_2 = \mathbf{A}_3 + \mathbf{A}_4$, where $\mathbf{A}_3 = \mathbf{P}_3^\top \tilde{\mathbf{A}}_3 \mathbf{P}_3$ and $\mathbf{A}_4 = \mathbf{P}_4^\top \tilde{\mathbf{A}}_4 \mathbf{P}_4$, where $\tilde{\mathbf{A}}_3 = \frac{\gamma-1}{2}\mathbf{I} + \frac{1}{5}\mathbf{\Sigma}_3$ and $\tilde{\mathbf{A}}_4 = \frac{\gamma-1}{2}\mathbf{I}$. As before $\kappa(\tilde{\mathbf{A}}_3) \leq 10$, and $\kappa(\tilde{\mathbf{A}}_4) = 1$.

For any matrix $\mathbf{A}$ and vectors $\mathbf{x}, \mathbf{b}$, let $g(\mathbf{A}, \mathbf{b}, \mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} - 2\mathbf{b}^\top \mathbf{x}$. Using the decomposition of $\mathbf{A} = \sum_{i=1}^4 \mathbf{P}_i^\top \tilde{\mathbf{A}}_i \mathbf{P}_i$ and the fact that $\mathbf{x} = \mathbf{P}_1\mathbf{x} + \mathbf{P}_2\mathbf{x} = \mathbf{P}_3\mathbf{x} + \mathbf{P}_4\mathbf{x}$ we can write,

$$f(\mathbf{x}) = g(\mathbf{A}, \mathbf{b}, \mathbf{x}) = g(\tilde{\mathbf{A}}_1, \mathbf{b}/2, \mathbf{P}_1\mathbf{x}) + g(\tilde{\mathbf{A}}_2, \mathbf{b}/2, \mathbf{P}_2\mathbf{x}) + g(\tilde{\mathbf{A}}_3, \mathbf{b}/2, \mathbf{P}_3\mathbf{x}) + g(\tilde{\mathbf{A}}_4, \mathbf{b}/2, \mathbf{P}_4\mathbf{x}).$$

Note that for any $1 \leq i \leq 4$ the condition number of $g(\tilde{\mathbf{A}}_i, \mathbf{b}/2, \mathbf{P}_i\mathbf{x})$ is $\kappa(\tilde{\mathbf{A}}_i) \leq 10$. The condition number of $g(\mathbf{A}, \mathbf{b}, \mathbf{x})$ is $\kappa(\mathbf{A}) = \Theta(d^2)$.

Since $\mathcal{E}_1 \cap \mathcal{E}_2$ happens with probability at least 0.98, we have the above decomposition with probability at least 0.98. Note that any $\hat{\mathbf{x}}$ which is a $\frac{1}{10d^2}$-optimal solution to the above problem satisfies,

$$\left\|\hat{\mathbf{x}} - \mathbf{A}^{-1}\mathbf{b}\right\|_{\mathbf{A}}^2 \leq \frac{1}{10d^2}.$$

Therefore, any algorithm which solves the multi-scale optimization problem probability at least 0.9, also solves the hard instance of $\mathbf{A}$ in Braverman et al. (2020) with probability at least 0.8. By Theorem 65, this requires at least $\Omega(\kappa(\mathbf{A})/\operatorname{polylog}(d))$ gradient queries. ■

F.6. Complexity of conjugate gradient for quadratic multiscale optimization

In this section, we give a simple proof that the conjugate gradient algorithm can stably solve the multiscale optimization problem in the special case when f *is quadratic* in a number of iterations that is comparable to what our accelerated BSLS algorithm requires. More precisely, we show:

Theorem 66 (Complexity of conjugate gradient for multiscale quadratic optimization) *Consider an instance of the multiscale optimization problem (Def. 1) in which each f_i is quadratic. For any $\mathbf{x}^{(0)}$ and $\epsilon > 0$, the conjugate gradient method started at $\mathbf{x}^{(0)}$, can return an ϵ-optimal solution with $\left(\prod_{i\in[m]} \mathcal{O}(\sqrt{\kappa_i})\right) \cdot \mathcal{O}\left(\left(\log^{m-1} \kappa_{\text{glob}}\right) \cdot \log\left(\frac{f(\mathbf{x}^{(0)}) - f^\star}{\epsilon}\right)\right)$ gradient queries. This remains true if all operations are performed using a number of bits of precision that is logarithmic in the problem parameters.*

The conjugate gradient method is usually discussed as an algorithm for solving linear systems, so we begin by rephrasing our quadratic optimization problem in this form.

We can write our quadratic objective function as $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} - 2\mathbf{b}^\top \mathbf{x}$ for some $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{b} \in \mathbb{R}^d$. The assumption that f is strongly convex corresponds to the requirement that the matrix $\mathbf{A}$ be positive definite, and the assumption about the existence of a decomposition of f in terms of f_i with the given smoothness and convexity properties corresponds to the assumption that the eigenvalues of $\mathbf{A}$ all lie in the set $S = \bigcup_{i\in[m]}[\mu_i, L_i]$.

Since $\nabla f(x) = 2(\mathbf{A}\mathbf{x} - \mathbf{b})$, f is minimized at $\mathbf{x}^\star = \mathbf{A}^{-1}\mathbf{b}$, and the function error at some other point $\mathbf{x}$ is given by

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}^\star) &= (\mathbf{x}^\top \mathbf{A}\mathbf{x} - 2\mathbf{b}^\top \mathbf{x}) - (\mathbf{x}^{\star\top} \mathbf{A}\mathbf{x}^\star - 2\mathbf{b}^\top \mathbf{x}^\star) \\ &= (\mathbf{x}^\top \mathbf{A}\mathbf{x} - 2\mathbf{x}^{\star\top} \mathbf{A}\mathbf{x}) - (\mathbf{x}^{\star\top} \mathbf{A}\mathbf{x}^\star - 2\mathbf{x}^{\star\top} \mathbf{A}\mathbf{x}^\star) \\ &= \mathbf{x}^\top \mathbf{A}\mathbf{x} - 2\mathbf{x}^{\star\top} \mathbf{A}\mathbf{x} + \mathbf{x}^{\star\top} \mathbf{A}\mathbf{x}^\star \\ &= (\mathbf{x} - \mathbf{x}^\star)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^\star) := \|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{A}}^2. \end{aligned}$$

Minimizing f is thus equivalent to solving the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, and the function error at a point equals its distance from the optimal solution in the $\mathbf{A}$-norm. To prove Theorem 66, it thus suffices to bound the convergence rate in the $\mathbf{A}$-norm of the conjugate gradient method applied to matrices with eigenvalues in S.

Our proof relies on the connection between the performance of the conjugate gradient algorithm and polynomial approximation. If the algorithm uses exact arithmetic, the classical analysis of the conjugate gradient algorithm asserts that, after k iterations, the algorithm returns a vector $\mathbf{x}^{(k)}$ such that

$$\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_{\mathbf{A}} \le \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_{\mathbf{A}} \cdot \min_{p\in\mathcal{P}_k^0} \max_{i\in[d]} |p(\lambda_i(\mathbf{A})|,$$

where $\mathcal{P}_k^0$ denotes the set of polynomials of degree at most k with $p(0) = 1$.

To prove Theorem 66 under exact arithmetic, it thus suffices to construct a polynomial $p \in \mathcal{P}_k^0$ for k less than or equal to the given bound on the number of gradient queries with $|p(x)| \le \epsilon$ for all $x \in S$.

For finite-precision arithmetic, we apply the following theorem of Greenbaum, which says that the convergence rate of the conjugate gradient method applied to a matrix $\mathbf{A}$ with eigenvalues in S using precision that is logarithmic in the problem parameters can be bounded in terms of its convergence rate under *exact* arithmetic on a matrix with eigenvalues in a slightly enlarged set $S' \supseteq S$.

Theorem 67 (Greenbaum (1989b), as simplified in Musco et al. (2018a)) *Given a positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{b} \in \mathbb{R}^n$, let $\mathbf{x}$ be the result of running the conjugate gradient method for k iterations on the linear system $A\mathbf{x} = \mathbf{b}$ with all operations performed using $\Omega\left(\log \frac{nk(\|\mathbf{A}\|+1)}{\min(\eta,\lambda_{\min}(\mathbf{A})}\right)$ bits of precision.*

Let $\Delta = \min\left(\eta, \frac{\lambda_{\min}(\mathbf{A})}{5}\right)$. There exists a matrix $\tilde{\mathbf{A}}$ with eigenvalues in $S' := \bigcup_{i=1}^n [\lambda_i(\mathbf{A}) - \Delta, \lambda_i(\mathbf{A}) + \Delta]$ and a vector $\tilde{\mathbf{b}}$ with $\|\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{b}}\|_{\tilde{\mathbf{A}}} = \|\mathbf{A}^{-1}\mathbf{b}\|_{\mathbf{A}}$ such that, if $\tilde{\mathbf{x}}$ is the result of running the conjugate gradient method for k iterations on the linear system $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ in exact *arithmetic, then*

$$\|\mathbf{A}^{-1}\mathbf{b} - \mathbf{x}\|_{\mathbf{A}} \le 1.2\|\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{b}} - \tilde{\mathbf{x}}\|_{\tilde{\mathbf{A}}}.$$

Replacing S with S' does not change the asymptotic behavior of the bound asserted in Theorem 66, so it suffices to show the existence of polynomials with the properties required in the exact case. The following Theorem asserts the existence of such polynomials, from which Theorem 66 follows.

Theorem 68 (Existence of good polynomials for unions of intervals) *Let $S = \bigcup_{i\in[m]}[\mu_i, L_i]$, where $\mu_1 < L_1 < \mu_2 < L_2 < \ldots$. For any $\epsilon > 0$, there exists a polynomial P such that $P(0) = 1$, $|P(x)| \leq \epsilon$ for all $x \in S$, and $\deg(P) \leq \left(\prod_{i\in[m]} \mathcal{O}(\sqrt{\kappa_i})\right) \cdot \mathcal{O}\left(\log^{m-1}(\kappa_{\text{glob}}) \cdot \log(1/\epsilon)\right)$.*

We devote the remainder of this section to constructing the polynomials required by this theorem. Note that our goal here is to present a simple construction that has the desired asymptotic behavior rather than to optimize the constants, and the polynomials given are not the exactly optimal polynomials for S.

F.6.1. CONSTRUCTING GOOD POLYNOMIALS FOR UNIONS OF INTERVALS

The basic building blocks of our construction are Chebyshev polynomials. Simple transformations of Chebyshev polynomials give optimal polynomials for individual intervals. We construct good polynomials for unions of intervals by multiplying such polynomials together. The main technical difficulty is that the polynomial for one interval can be quite large on another interval, so naively multiplying together the polynomials for the individual intervals will not produce something that is small on all of S. Instead we will carefully choose the degrees of the polynomials on the different intervals to manage the error caused by these interactions.

We begin by reviewing the definition of Chebyshev polynomials and providing some standard bounds on their magnitude. Let $T_d(x) = \frac{1}{2}\left(x + \sqrt{x^2-1}\right)^d + \frac{1}{2}\left(x - \sqrt{x^2-1}\right)^d$ be the degree-d Chebyshev polynomial (of the first kind). This defines a degree-d polynomial with the following well-known properties:[3]

1. If $|x| \leq 1$, $|T_d(x)| \leq 1$.
2. If $|x| \geq 1$,
$$\frac{1}{2}\left(1 + \sqrt{2(|x|-1)}\right)^d \leq |T_d(x)| \leq |2x|^d,$$
and $|T_d(x)|$ for such x is a monotonically increasing function of $|x|$.

To construct a polynomial p with $p(0) = 1$ that is small on a single interval, we can simply compose Chebyshev polynomials with a linear function that maps our interval onto $[-1, 1]$ and then normalize to get $p(0) = 1$.

To this end, let
$$\ell_{[a,b]}(x) = \frac{b + a - 2x}{b - a}$$
be the linear function that maps $[a, b]$ to $[-1, 1]$ with $\ell_{[a,b]}(a) = 1$ and $\ell_{[a,b]}(b) = -1$, and define
$$p^{(d)}_{[a,b]}(x) = \frac{T_d(\ell_{[a,b]}(x))}{T_d(\ell_{[a,b]}(0))}.$$

3. For an introduction to Chebyshev polynomials and their basic properties, see Mason and Handscomb (2003)

Lemma 69 *For any $b \geq 2a > 0$, if $p^{(d)}_{[a,b]}(x)$ is the polynomial defined above, and $\kappa = b/a$, then*

$$p^{(d)}_{[a,b]}(0) = 1$$

and

$$\left|p^{(d)}_{[a,b]}(x)\right| \leq \begin{cases} 1 & \text{if } x \in [0,a] \\ 2\left(1 + \frac{2}{\sqrt{\kappa}}\right)^{-d} & \text{if } x \in [a,b] \\ \left(\frac{8x}{b}\right)^d & \text{if } x \geq b. \end{cases}$$

Proof The fact that $p^{(d)}_{[a,b]}(0) = 1$ follows immediately from the definition.

For the bound on $|p^{(d)}_{[a,b]}(x)$, lets first suppose that $x \in [0,a]$. In this case, $\ell_{[a,b]}(x) \in \left[1, \ell_{[a,b]}(0)\right]$ with $\ell_{[a,b]}(0) \geq 1$. By the monotonicity assertion in property 2 of Chebyshev polynomials, $T_d(\ell_{[a,b]}(x)) \leq T_d(\ell_{[a,b]}(0))$, so $\left|p^{(d)}_{[a,b]}(x)\right| \leq \left|p^{(d)}_{[a,b]}(0)\right| = 1$, as claimed.

Now, suppose $x \in [a,b]$. In this case, note that $\ell_{[a,b]}(x) \in [-1,1]$, so $|T_d(\ell_{[a,b]}(x))| \leq 1$ by property 1 of Chebyshev polynomials. For the denominator, we have

$$\ell_{[a,b]}(0) = \frac{b+a}{b-a} = 1 + \frac{2a}{b-a} = 1 + \frac{2}{\kappa - 1},$$

so, by property 2 of Chebyshev polynomials and the fact that $T_d(x) \geq 0$ for $x \geq 1$,

$$T_d(\ell_{[a,b]}(0)) = T_d\left(1 + \frac{2}{\kappa-1}\right) \geq \frac{1}{2}\left(1 + \sqrt{2\left(\frac{2}{\kappa-1}\right)}\right)^d \geq \frac{1}{2}\left(1 + \frac{2}{\sqrt{\kappa}}\right)^d.$$

Combining our bounds on the numerator and denominator gives the asserted bound on $\left|p^{(d)}_{[a,b]}(x)\right|$ for $x \in [a,b]$.

Finally, suppose $x \geq b$, and let $\gamma = x/b \geq 1$. We have $|\ell_{[a,b]}(x)| \geq 1$, so, by property 2 of Chebyshev polynomials,

$$|T_d(\ell_{[a,b]}(x))| \leq \left|2\ell_{[a,b]}(x))\right|^d = \left|2\left(\frac{b+a-2x}{b-a}\right)\right|^d = \left|2\left(\frac{\kappa+1-2x/a}{\kappa-1}\right)\right|^d$$
$$= \left|2\left(\frac{\kappa+1-2\kappa\gamma}{\kappa-1}\right)\right|^d = \left|-2\left(1 + 2(\gamma-1)\left(1 + \frac{1}{\kappa-1}\right)\right)\right|^d = \left|2 + 4(\gamma-1)\left(1 + \frac{1}{\kappa-1}\right)\right|^d.$$

Our assumption that $b \geq 2a$ implies that $\kappa \geq 2$, so we have

$$|T_d(\ell_{[a,b]}(x))| \leq |2 + 8(\gamma-1)|^d = |8\gamma - 6|^d = \left|\frac{8x}{b} - 6\right|^d \leq \left(\frac{8x}{b}\right)^d.$$

Combining this with the fact that $T_d(\ell_{[a,b]}(0)) \geq 1$ by the monotonicity in property 2 gives the desired bound on $\left|p^{(d)}_{[a,b]}(x)\right|$ for $x \geq b$. ■

Proof [Proof of Theorem 68] To simplify the calculations, we assume that $\kappa_i \geq 2$ for all i. By enlarging and combining our intervals as necessary, we can easily reduce the general theorem to this case.

Let $d_\kappa(\epsilon) = \sqrt{\kappa}\lceil \log(2/\epsilon) \rceil$, and note that, for $\kappa \geq 2$ and $\epsilon > 0$,

$$2\left(1 + \frac{2}{\sqrt{\kappa}}\right)^{-d_\kappa(\epsilon)} < \epsilon.$$

Let $S = \bigcup_{i\in[m]}[\mu_i, L_i]$, where $\mu_1 < L_1 < \mu_2 < L_2 < \ldots$, and assume that $\kappa_i := L_i/\mu_i \geq 2$ for all i.

We will obtain a good polynomial for S by multiplying together the polynomials for the different intervals with carefully-chosen degrees. To this end, let

$$P_{d_1,\ldots d_m}(x) = \prod_{i\in[m]} p^{(d_i)}_{[\mu_i,L_i]}(x),$$

and note that $P_{d_1,\ldots d_m}(0) = 1$.

By Lemma 69, for $x \in [\mu_j, L_j]$, we have

$$\begin{aligned}
|P_{d_1,\ldots d_m}(x)| &= \prod_{i\in[m]} \left|p^{(d_i)}_{[\mu_i,L_i]}(x)\right| \leq \prod_{i\in[m]} \begin{cases} 1 & \text{if } x \in [0,\mu_i] \\ 2\left(1+\frac{2}{\sqrt{\kappa_i}}\right)^{-d_i} & \text{if } x \in [\mu_i, L_i] \\ \left(\frac{8x}{L_i}\right)^{d_i} & \text{if } x \geq L_i \end{cases} \\
&= \prod_{i<j}\left(\frac{8x}{L_i}\right)^{d_i} \cdot 2\left(1+\frac{2}{\sqrt{\kappa_j}}\right)^{-d_j} \cdot \prod_{i>j} 1 = 2\left(1+\frac{2}{\sqrt{\kappa_j}}\right)^{-d_j} \prod_{i<j}\left(\frac{8x}{L_i}\right)^{d_i}.
\end{aligned}$$

If we want $|P_{d_1,\ldots d_m}(x)| \leq \epsilon$ for all $x \in S$, it thus suffices to choose the d_j so that, for all j,

$$2\left(1+\frac{2}{\sqrt{\kappa_j}}\right)^{-d_j} \leq \epsilon \cdot \prod_{i<j}\left(\frac{8x}{L_i}\right)^{-d_i}.$$

We can achieve this by setting $d_1 = d_{\kappa_1}(\epsilon) = \sqrt{\kappa_1}\lceil \log(2/\epsilon)\rceil$ and then recursively setting

$$\begin{aligned}
d_j &= d_{\kappa_j}\left(\epsilon \cdot \prod_{i<j}\left(\frac{8x}{L_i}\right)^{-d_i}\right) \\
&= \sqrt{\kappa_j}\left\lceil \log\left((2/\epsilon)\cdot\prod_{i<j}\left(\frac{8x}{L_i}\right)^{d_i}\right)\right\rceil = \sqrt{\kappa_j}\left\lceil \log(2/\epsilon) + \sum_{i<j} d_i \log\left(\frac{8x}{L_i}\right)\right\rceil
\end{aligned}$$

Since $x/L_i \leq \kappa_{\text{glob}}/\kappa_1$ and $d_1 \geq \sqrt{\kappa_1}\log(2/\epsilon)$, and using the fact that $\sqrt{\kappa_1}\log(9\kappa_1/8) \geq 1$ for $\kappa_1 \geq 2$, we have the bound

$$\begin{aligned}
d_j &\leq \sqrt{\kappa_j}\left\lceil \log\frac{2}{\epsilon} + \log\frac{8\kappa_{\text{glob}}}{\kappa_1}\sum_{i<j} d_i\right\rceil = \sqrt{\kappa_j}\left\lceil \log\frac{2}{\epsilon} - \log\frac{9\kappa_1}{8}\sum_{i<j} d_i + \log(9\kappa_{\text{glob}})\sum_{i<j} d_i\right\rceil \\
&\leq \sqrt{\kappa_j}\left\lceil \log\frac{2}{\epsilon} - \log\frac{9\kappa_1}{8}\sqrt{\kappa_1}\log\frac{2}{\epsilon} + \log(9\kappa_{\text{glob}})\sum_{i<j} d_i\right\rceil \leq \sqrt{\kappa_j}\,\lceil \log(9\kappa_{\text{glob}})\rceil \sum_{i<j} d_i.
\end{aligned}$$

Since our recurrence guarantees that $d_{i+1} \geq 2d_i$, we have $\sum_{i<j} d_i \leq 2d_{j-1}$, so our bound becomes

$$d_j \leq \sqrt{\kappa_j} \cdot 2\left\lceil \log(9\kappa_{\text{glob}}) \right\rceil d_{j-1}.$$

Applying this recursively gives

$$d_j \leq \left(\prod_{i=2}^{j} \sqrt{\kappa_j}\right) \left(2\left\lceil \log(9\kappa_{\text{glob}}) \right\rceil\right)^{j-1} d_1 = \left(\prod_{i=2}^{j} \sqrt{\kappa_j}\right) \left(2\left\lceil \log(9\kappa_{\text{glob}}) \right\rceil\right)^{j-1} \left(\sqrt{\kappa_1}\left\lceil \log(2/\epsilon) \right\rceil\right)$$
$$= \left(\prod_{i=1}^{j} \sqrt{\kappa_j}\right) \left(2\left\lceil \log(9\kappa_{\text{glob}}) \right\rceil\right)^{j-1} \left\lceil \log(2/\epsilon) \right\rceil,$$

and the total degree is then bounded by $2d_m$, which obeys the desired asymptotic bound. ■

Appendix G. Proof of BSLS under finite-precision arithmetic

In this section we prove BSLS with finite-precision arithmetic (subject to Requirement 1).

In Theorem 14, we specialized our initialization of $\mathbf{x}^{(0)}$ to $\mathbf{0}$ to simplify the exposition of the theorem. In fact, we can (and will) prove the following general (but less clean) version with arbitrary $\mathbf{x}^{(0)}$.

Theorem 70 (BSLS under finite-precision arithmetic, general initialization) *Consider multiscale optimization problem defined in Theorem 1, for any initialization $\mathbf{x}^{(0)}$ and $\epsilon > 0$, assuming Requirement 1 with*

$$\delta^{-1} \geq \left(\prod_{i\in[m]} T_i\right) \max\left\{(10\kappa_{\text{glob}})^{2m-1}m, (10\kappa_{\text{glob}})^{2m-1}m \cdot \frac{f(\mathbf{x}^{(0)}) - f^\star}{\epsilon}, 4^{m+1}mL_m\frac{\|\mathbf{x}^\star\|_2^2}{\epsilon}\right\}$$

then $f(\widehat{\text{BSLS}}_1(\mathbf{x}^{(0)})) - f^\star \leq 3\epsilon$ provided that $T_1, \ldots, T_m$ satisfy

$$T_1 \geq \kappa_1 \log\left(\frac{f(\mathbf{x}^{(0)}) - f^\star}{\epsilon}\right); \qquad T_i \geq \kappa_i(2\log(\kappa_{\text{glob}}) + 1), \quad \text{for } i = 2, \ldots, m. \tag{G.1}$$

We can also achieve the same asymptotic sample complexity (up to constant factors suppressed in the $\mathcal{O}(\cdot)$) when $\{(\mu_i, L_i), i \in [m]\}$ are unknown and only m, μ_1, L_m and $\pi_\kappa = \prod_{i=1}^m \kappa_i$ are known.

Theorem 14 is clearly a corollary of Theorem 70.

Proof [Proof of Theorem 14 based on Theorem 70] Follows by the fact that

$$4^{m+1}mL_m\frac{\|\mathbf{x}^\star\|_2^2}{\epsilon} \leq 4^{m+2}m\kappa_{\text{glob}} \cdot \frac{f(\mathbf{0}) - f^\star}{\epsilon} \leq m(10\kappa_{\text{glob}})^{2m-1} \cdot \frac{f(\mathbf{0}) - f^\star}{\epsilon}.$$

■

From now on we focus on the proof of Theorem 70. The proof of Theorem 70 is structured as follows. We first study the progress of one (inexact) $\widehat{\text{GD}}$ step in Appendix G.1, and inductively estimate the progress of $\widehat{\text{BSLS}}_i$ by matrix inequalities for all $i \in [m]$ in descent order (see Appendix G.2). The proof of Theorem 70 is then finished in Appendix G.3. Note that the last part regarding the case where $\{(\mu_i, L_i), i \in [m]\}$ are unknown follows from our black-box reduction in Proposition 13 (in the same way as in the proof of Theorem 6).

Additional notation. We introduce notation to simplify the exposition. For any $\mathbf{x}$ and $i \in [m]$, define function $\Delta_i(\mathbf{x}) \coloneqq f_i(\mathbf{P}_i\mathbf{x}) - f_i^\star$. Define vector potentials

$$\mathbf{\Delta}(\mathbf{x}) \coloneqq \left[\Delta_1(\mathbf{x}), \Delta_2(\mathbf{x}), \cdots, \Delta_m(\mathbf{x})\right]^\top \in \mathbb{R}^m.$$

We will monitor the progress via this vector potential $\mathbf{\Delta}$. Recall that when comparing two vectors or matrices, we use plain inequalities $(\leq, \geq)$ to denote element-wise inequality.

G.1. Progress of one GD step under finite arithmetic

In this subsection, we study the effect of one (inexact) gradient step $\widehat{\texttt{GD}}$ on the vector potential $\mathbf{\Delta}$ under finite arithmetic (subject to Requirement 1). The goal is to establish the following Lemma 71.

Lemma 71 (Progress of one $\widehat{\texttt{GD}}$ step under finite arithmetic) *Consider multiscale optimization (Def. 1), and assuming Requirement 1, then for any $\mathbf{x}$ and $i \in [m]$, the following inequality holds*

$$\mathbf{\Delta}(\widehat{\texttt{GD}}(\mathbf{x}; L_i)) \leq \left(\mathbf{I} + 5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top\right)\mathbf{D}_i\mathbf{\Delta}(\mathbf{x}) + 2\delta\|\mathbf{x}^\star\|_2^2 L_m \cdot \mathbf{1}, \tag{G.2}$$

where $\mathbf{D}_i$ is an $m \times m$ diagonal matrix defined by

$$(\mathbf{D}_i)_{jj} = \begin{cases} 1 & \text{if } j < i, \\ 1 - \kappa_i^{-1} & \text{if } j = i, \\ \kappa_{\text{glob}}^2 & \text{if } j > i. \end{cases} \tag{G.3}$$

To simplify the notation we will define (throughout this section) that

$$\widehat{\mathbf{D}_i} \coloneqq \left(\mathbf{I} + 5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top\right)\mathbf{D}_i. \tag{G.4}$$

Then Eq. (G.2) becomes

$$\mathbf{\Delta}(\widehat{\texttt{GD}}(\mathbf{x}; L_i)) \leq \widehat{\mathbf{D}_i}\mathbf{\Delta}(\mathbf{x}) + 2\delta\|\mathbf{x}^\star\|_2^2 L_m \cdot \mathbf{1}.$$

Remark 72 *The key observation from Lemma 71 is that under finite-precision arithmetic, the function error in j-th subspace (i.e., Δ_j) also depends on the the errors from other subspace, as well as an constant additive term. If the error in one of the subspaces is too large, it could flow into the other subspaces and ruins the progress elsewhere. Consequently, the order of step-size schedule is crucial in finite-precision arithmetic.*

To prove Lemma 71, we first study the sensitivity of potential $\mathbf{\Delta}$ under multiplicative perturbtaion.

Lemma 73 (Sensitivity of vector potential $\mathbf{\Delta}$ under multiplicative perturbation) *Assuming $\widehat{\mathbf{x}}, \mathbf{x}$ satisfies*

$$|\widehat{\mathbf{x}} - \mathbf{x}| \leq \delta|\mathbf{x}| \tag{G.5}$$

for some $\delta < 1$, then for any $j \in [m]$,

$$\Delta_j(\widehat{\mathbf{x}}) \leq \Delta_j(\mathbf{x}) + 5\delta\kappa_{\text{glob}}\sum_{k=1}^m \Delta_k(\mathbf{x}) + 2\delta L_j\|\mathbf{x}^\star\|_2^2.$$

In vector form we have (in a looser form)

$$\mathbf{\Delta}(\widehat{\mathbf{x}}) \leq (I + 5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top)\mathbf{\Delta}(\mathbf{x}) + 2\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{1}.$$

Proof [Proof of Lemma 73] For any $j \in [m]$,

$$\begin{aligned}
&\Delta_j(\widehat{\mathbf{x}}) = f_j(\mathbf{P}_j\widehat{\mathbf{x}}) - f_j^\star && \text{(by definition of } \Delta_j\text{)}\\
\le& f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star + \langle \nabla f_j(\mathbf{P}_j\mathbf{x}), \mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x})\rangle + \frac{L_j}{2}\|\mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x})\|_2^2 && \text{(by } L_j\text{-smoothness of } f_j\text{)}\\
\le& f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star + \frac{\delta}{2L_j}\|\nabla f_j(\mathbf{P}_j\mathbf{x})\|_2^2 + \frac{L_j}{2\delta}\|\mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x})\|_2^2 + \frac{L_j}{2}\|\mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x})\|_2^2 \\
& && \text{(by Cauchy-Schwartz inequality)}\\
\le& f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star + \frac{\delta}{2L_j}\|\nabla f_j(\mathbf{P}_j\mathbf{x})\|_2^2 + \frac{L_j}{\delta}\|\mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x})\|_2^2 && \text{(since } \delta \le 1\text{)}\\
\le& (1+\delta)(f_j(\mathbf{P}_j\mathbf{x}) - f_j^\star) + \frac{L_j}{\delta}\|\mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x})\|_2^2. && \text{(by } L_j\text{-smoothness of } f_j\text{)}\\
\le& (1+\delta)\Delta_j(\mathbf{x}) + \frac{L_j}{\delta}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 && \text{(by definition of } \Delta_j\text{)}
\end{aligned}$$

By assumption Eq. (G.5)

$$\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \le \delta^2\|\mathbf{x}\|_2^2 \le 2\delta^2\|\mathbf{x} - \mathbf{x}^\star\|_2^2 + 2\delta^2\|\mathbf{x}^\star\|_2^2, \qquad \text{(by Cauchy-Schwartz inequality)}$$

and strong convexity of f_i's

$$\|\mathbf{x} - \mathbf{x}^\star\|_2^2 = \sum_{i\in[m]}\|\mathbf{P}_i(\mathbf{x} - \mathbf{x}^\star)\|_2^2 \le \sum_{i\in[m]}\frac{2}{\mu_i}\Delta_i(\mathbf{x}),$$

we arrive at

$$\frac{L_j}{\delta}\|(\widehat{\mathbf{x}} - \mathbf{x})\|_2^2 \le 2\delta L_j\|\mathbf{x}^\star\|_2^2 + \sum_{i\in[m]}\frac{4\delta L_j}{\mu_i}\Delta_i(\mathbf{x}) \le 2\delta L_j\|\mathbf{x}^\star\|_2^2 + 4\delta\kappa_{\text{glob}}\sum_{i\in[m]}\Delta_i(\mathbf{x}),$$

where the last inequality is due to $\frac{L_j}{\mu_i} \le \kappa_{\text{glob}}$ by definition of κ_{glob}. In summary

$$\begin{aligned}
\Delta_j(\widehat{\mathbf{x}}) &\le (1+\delta)\Delta_j(\mathbf{x}) + 4\delta\kappa_{\text{glob}}\sum_{i\in[m]}\Delta_i(\mathbf{x}) + 2\delta L_j\|\mathbf{x}^\star\|_2^2\\
&\le \Delta_j(\mathbf{x}) + 5\delta\kappa_{\text{glob}}\sum_{i\in[m]}\Delta_i(\mathbf{x}) + 2\delta L_j\|\mathbf{x}^\star\|_2^2.
\end{aligned}$$

In vector form we have (since $L_1 \le L_2 \le \cdots \le L_m$)

$$\boldsymbol{\Delta}(\widehat{\mathbf{x}}) \le (I + 5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top)\boldsymbol{\Delta}(\mathbf{x}) + 2\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{1},$$

completing the proof. ■

With Lemma 73 at hands we are ready to prove Lemma 71:

Proof [Proof of Lemma 71] Apply Lemma 73, we have

$$\boldsymbol{\Delta}(\widehat{\texttt{GD}}(\mathbf{x}; L_i)) \le (\mathbf{I} + 5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top)\boldsymbol{\Delta}(\texttt{GD}(\mathbf{x}; L_i)) + 2\delta L_m\|\mathbf{x}^\star\|_2^2 \cdot \mathbf{1}.$$

By Lemma 12 from exact BSLS analysis we have

$$\boldsymbol{\Delta}(\texttt{GD}(\mathbf{x}; L_i)) \le \mathbf{D}_i\boldsymbol{\Delta}(\mathbf{x}).$$

Combining the two inequalities above yields Lemma 73. ■

G.2. Inductively bound the progress of inexact BSLS by matrix inequalities

In the following Lemma 74, we iteratively construct the bound of $\mathbf{\Delta}$ after executing $\widehat{\texttt{BSLS}_i}$.

Lemma 74 (Estimate the progress of $\widehat{\texttt{BSLS}_i}$ by matrix inequalities) *Considering multiscale optimization problem (Theorem 1), and assuming Requirement 1, define the following three sequences of $m \times m$ matrices $\{\mathbf{F}_i\}_{i=1}^{m+1}$, $\{\mathbf{E}_i\}_{i=1}^{m+1}$, $\{\mathbf{Z}_i\}_{i=1}^{m+1}$ as follows:*

$$\mathbf{F}_{m+1} \coloneqq \mathbf{I}, \qquad \mathbf{E}_{m+1} \coloneqq \mathbf{0}, \qquad \mathbf{Z}_{m+1} \coloneqq \mathbf{0}$$

and for $i = m, m-1$ down to 1, define

$$\mathbf{F}_i \coloneqq (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\mathbf{F}_{i+1}, \quad \mathbf{E}_i \coloneqq \left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i}\right)^{T_i}(\mathbf{F}_{i+1}+\mathbf{E}_{i+1}) - (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\mathbf{F}_{i+1}.$$

and

$$\mathbf{Z}_i \coloneqq ((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i})^{T_i}\mathbf{Z}_{i+1} + \sum_{t_i=0}^{T_i-1}\left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i}\right)^{t_i}(\mathbf{Z}_{i+1}+\mathbf{F}_{i+1}+\mathbf{E}_{i+1}),$$

where $\mathbf{D}_i$ and $\widehat{\mathbf{D}_i}$ were defined in Eqs. (G.3) and (G.4). Then,

(a) For any $i \in [m+1]$, $\mathbf{F}_i$, $\mathbf{E}_i$ and $\mathbf{Z}_i$ are non-negative matrices.

(b) For any $i \in [m]$, the following bound holds

$$\mathbf{\Delta}(\widehat{\texttt{BSLS}_i}(\mathbf{x})) \le (\mathbf{F}_i+\mathbf{E}_i)\mathbf{\Delta}(\mathbf{x}) + 2\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{Z}_i\mathbf{1}.$$

Proof [Proof of Lemma 74]

(a) We first prove (a) by induction in reverse order (from $m+1$ down to 1).

For $i = m+1$ the statement apparently holds. Now assume (a) holds for $i+1$, then we study the case of i. For $\mathbf{F}_i$ we have $\mathbf{F}_i = (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\mathbf{F}_{i+1} \ge \mathbf{0}$ since both $\mathbf{F}_{i+1}$ and $\mathbf{D}_i$ are non-negative. For $\mathbf{E}_i$ we have

$$\mathbf{E}_i = ((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})(\mathbf{I}+5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i)^{T_i}(\mathbf{F}_{i+1}+\mathbf{E}_{i+1}) - (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\mathbf{F}_{i+1} \ge \mathbf{0}.$$

The non-negativity of $\mathbf{Z}_i$ is obvious from the non-negativity of $\mathbf{F}_i$ and $\mathbf{E}_i$.

(b) Next, we prove (b) by induction in reverse order. To simplify the induction let us define $\widehat{\texttt{BSLS}_{m+1}} \coloneqq \mathrm{Id}$ and prove (b) for all $i \in [m+1]$. The inequality holds trivially for $i = m+1$. Now assume the inequality holds for the case of $i+1$, then we study the case of i.

By definition of $\widehat{\texttt{BSLS}_i}$ (including $i = m$),

$$\widehat{\texttt{BSLS}_i} = \left(\widehat{\texttt{BSLS}_{i+1}} \circ \widehat{\texttt{GD}}(\cdot; L_i)\right)^{T_i} \circ \widehat{\texttt{BSLS}_{i+1}} = \underbrace{\widehat{\texttt{BSLS}_{i+1}} \circ \widehat{\texttt{GD}}(\cdot; L_i) \circ \cdots \circ \widehat{\texttt{BSLS}_{i+1}} \circ \widehat{\texttt{GD}}(\cdot; L_i)}_{T_i \text{ iterations of } \widehat{\texttt{BSLS}_{i+1}} \circ \widehat{\texttt{GD}}(\cdot; L_i)} \circ \widehat{\texttt{BSLS}_{i+1}}$$

By induction hypothesis, for any $\mathbf{x}$,

$$\mathbf{\Delta}(\widehat{\texttt{BSLS}_{i+1}}(\mathbf{x})) \le (\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\mathbf{\Delta}(\mathbf{x}) + 2\delta L_m \|\mathbf{x}^\star\|_2^2 \cdot \mathbf{Z}_{i+1}\mathbf{1}.$$

For $\widehat{\texttt{GD}}(\cdot; L_i)$ step we have by Lemma 71 (for any $\mathbf{x}$)

$$\mathbf{\Delta}(\texttt{GD}(\mathbf{x}; L_i)) \le \widehat{\mathbf{D}_i}\mathbf{\Delta}(\mathbf{x}) + 2\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{1}.$$

Combining the above two inequalities, we obtain

$$\mathbf{\Delta}(\widehat{\texttt{BSLS}_{i+1}}(\widehat{\texttt{GD}}(\mathbf{x}; L_i))) \le (\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i}\mathbf{\Delta}(\mathbf{x}) + 2\delta L_m \|\mathbf{x}^\star\|_2^2 (\mathbf{Z}_{i+1}+\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\,\mathbf{1}.$$

Telescoping

$$\begin{aligned}
&\mathbf{\Delta}(\widehat{\texttt{BSLS}_i}(\mathbf{x})) \\
\le& \left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i}\right)^{T_i}(\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\mathbf{\Delta}(\mathbf{x}) \\
&+ 2\delta L_m\|\mathbf{x}^\star\|_2^2 \left[((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i})^{T_i}\mathbf{Z}_{i+1} + \sum_{t_i=0}^{T_i-1}\left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i}\right)^{t_i}(\mathbf{Z}_{i+1}+\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\right]\mathbf{1} \\
=&(\mathbf{F}_i+\mathbf{E}_i)\mathbf{\Delta}(\mathbf{x}) + 2\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{Z}_i\mathbf{1}. \qquad \text{(by definition of } \mathbf{F}_i, \mathbf{E}_i \text{ and } \mathbf{Z}_i)
\end{aligned}$$

■

Next, we estimate the upper bounds of $\|\mathbf{F}_i\|_1$ (in Appendix G.2.1), $\|\mathbf{E}_i\|_1$ (in Appendix G.2.2), and $\|\mathbf{Z}_i\|_1$ (in Appendix G.2.3).

G.2.1. UPPER BOUND OF $\mathbf{F}$

We first bound $\|\mathbf{F}_i\|_1$ with the following lemma.

Lemma 75 (Upper bound of $\|\mathbf{F}_i\|_1$) *Using the same notation as in Lemma 74, and in addition assuming $T_1, \ldots, T_m$ satisfies Eq. (G.1), then the following statements hold*

(a) For any $i \in [m+1]$, $\mathbf{F}_i$ is a diagonal matrix of the form $\prod_{j=i}^m \left(\mathbf{D}_j^{T_j \cdot \prod_{k=i}^{j-1}(T_k+1)}\right)$.

(b) For any $i \in [m]$, $\|\mathbf{F}_{i+1}\mathbf{D}_i\|_1 \le 1$.

(c) For any $i \in [m]$, $\|\mathbf{F}_i\|_1 \le 1$.

(d) $\|\mathbf{F}_1\|_1 \le \frac{\epsilon}{f(\mathbf{x}^{(0)})-f^\star}$.

Proof [Proof of Lemma 75]

(a) The first statement (a) follows immediately by definition of $\mathbf{F}_i$'s. We prove by induction in reverse order from $m+1$ down to 1. For $i = m+1$ we have $\mathbf{F}_{m+1} = \mathbf{I}$ which is consistent. Now assume the statement holds for the case of $i+1$, then the case of i also holds in that

$$\mathbf{F}_i = (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\mathbf{F}_{i+1} = \left(\prod_{j=i+1}^m \left(\mathbf{D}_j^{T_j\prod_{k=i+1}^{j-1}(T_k+1)}\right)\cdot \mathbf{D}_i\right)^{T_i} \prod_{j=i+1}^m \left(\mathbf{D}_j^{T_j\prod_{k=i+1}^{j-1}(T_k+1)}\right) = \prod_{j=i}^m \left(\mathbf{D}_j^{T_j\cdot\prod_{k=i}^{j-1}(T_k+1)}\right),$$

where the last equality is due to the commutability among diagonal matrices.

(b) Follows by the same analysis as in the exact arithmetic proof in Theorem 6.

(c) By (b), $\|\mathbf{F}_i\|_1 = \|(\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\mathbf{F}_{i+1}\|_1 \le \|(\mathbf{F}_{i+1}\mathbf{D}_i)\|_1^{T_i}\|\mathbf{F}_{i+1}\|_1 \le 1$.

(d) Follows by the same analysis as in the exact arithmetic proof in Theorem 6.

■

G.2.2. UPPER BOUND OF $\mathbf{E}$

Before we state the upper bound of $\|\mathbf{E}_i\|_1$, we first establish the following Lemma 76, which is essential towards the bound for $\mathbf{E}_i$'s and $\mathbf{Z}_i$'s.

Lemma 76 *Using the same notation of Lemma 74 and assuming the same assumptions of Lemma 75, and in addition assume $\delta \le \frac{1}{10m\kappa_{\text{glob}}}$, then for any $t \ge 0$, the following inequality holds*

$$\left\|((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i})^t - (\mathbf{F}_{i+1}\mathbf{D}_i)^t\right\|_1 \le \begin{cases}\varphi(5t\delta m\kappa_{\text{glob}}) & i=m\\ \varphi(2t\kappa_{\text{glob}}^2\|\mathbf{E}_{i+1}\|_1 + 5t\delta m\kappa_{\text{glob}}^3) & i<m,\end{cases}$$

where $\varphi(x) \coloneqq xe^x$.

Proof [Proof of Lemma 76] Denote $\mathbf{\Xi}_i \coloneqq (\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i} - \mathbf{F}_{i+1}\mathbf{D}_i$, then

$$\left\|((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i})^t - (\mathbf{F}_{i+1}\mathbf{D}_i)^t\right\| = \left\|((\mathbf{F}_{i+1}\mathbf{D}_i+\mathbf{\Xi}_i)^t - (\mathbf{F}_{i+1}\mathbf{D}_i)^t\right\|_1$$

(by definition of $\mathbf{\Xi}_i$)

$$\le \sum_{s=1}^{t}\binom{t}{s}\|\mathbf{\Xi}_i\|_1^s\|\mathbf{F}_{i+1}\mathbf{D}_i\|_1^{t-s} \le \sum_{s=1}^{t}\binom{t}{s}\|\mathbf{\Xi}_i\|_1^s \qquad \text{(since } \|\mathbf{F}_{i+1}\mathbf{D}_i\|_1 \le 1 \text{ by Lemma 75)}$$

$$\le \|\mathbf{\Xi}_i\|_1 t\sum_{s=0}^{t-1}\binom{t-1}{s}\|\mathbf{\Xi}_i\|_1^s \qquad \text{(by helper Theorem 100)}$$

$$= \|\mathbf{\Xi}_i\|_1 t\,(1+\|\mathbf{\Xi}_i\|_1)^{t-1} \le \|\mathbf{\Xi}_i\|_1 t\exp(\|\mathbf{\Xi}_i\|_1 t). \tag{G.6}$$

It remains to bound $\|\mathbf{\Xi}_i\|_1$. For $i=m$ we have $\mathbf{E}_{m+1}=0$, $\mathbf{F}_{m+1}=\mathbf{I}$, $\|\mathbf{D}_m\|_1 \le 1$, which suggests

$$\|\mathbf{\Xi}_m\|_1 = \|\widehat{\mathbf{D}_m} - \mathbf{D}_m\|_1 \le 5\delta\kappa_{\text{glob}}\|\mathbf{1}\mathbf{1}^\top\|_1 = 5\delta m\kappa_{\text{glob}}.$$

For other $i<m$, note that $\|\mathbf{D}_i\|_1 \le \kappa_{\text{glob}}^2$, $\|\mathbf{F}_{i+1}\|_1 \le 1$ (by Lemma 75), $\|\mathbf{1}\mathbf{1}^\top\|_1 = m$, we have

$$\|\mathbf{\Xi}_i\|_1 = \left\|(\mathbf{F}_{i+1}+\mathbf{E}_{i+1})(\mathbf{I}+5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i - \mathbf{F}_{i+1}\mathbf{D}_i\right\|_1$$

$$\le \|5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top\mathbf{D}_i\|_1 + \|\mathbf{E}_{i+1}(\mathbf{I}+5\delta\kappa_{\text{glob}}\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i\|_1 \le 5\delta m\kappa_{\text{glob}}^3 + 1.5\kappa_{\text{glob}}^2\|\mathbf{E}_{i+1}\|_1.$$

(since $\delta \le \frac{1}{10m\kappa_{\text{glob}}}$)

Substituting back to Eq. (G.6) completes the proof. ■

Lemma 77 (Upper bound of $\|\mathbf{E}_i\|_1$) *Using the same notation of Lemma 74 and assuming the same assumptions of Lemma 75, and in addition assume*

$$\delta \le \frac{1}{m(10\kappa_{\text{glob}})^{2m-1}\prod_{i\in[m]} T_i}, \tag{G.7}$$

then the following inequality holds for any $i \in [m]$,

$$\|\mathbf{E}_i\|_1 \le \delta m \cdot (10\kappa_{\text{glob}})^{2(m-i)+1} \cdot \prod_{j=i}^{m} T_j.$$

Proof [Proof of Lemma 77] First observe that

$$\begin{aligned}
\|\mathbf{E}_i\|_1 &= \left\| \left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i}\right)^{T_i} (\mathbf{F}_{i+1}+\mathbf{E}_{i+1}) - (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\,\mathbf{F}_{i+1} \right\|_1 \\
&\le \left\| \left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i}\right)^{T_i} (\mathbf{F}_{i+1}+\mathbf{E}_{i+1}) - (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\,(\mathbf{F}_{i+1}+\mathbf{E}_{i+1}) \right\|_1 + \left\| (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\,\mathbf{E}_{i+1} \right\|_1 \\
&\le \left\| \left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i}\right)^{T_i} - (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i} \right\|_1 (1+\|\mathbf{E}_{i+1}\|_1) + \|\mathbf{E}_{i+1}\|_1,
\end{aligned}$$

where in the last inequality we applied the fact that $\|\mathbf{F}_{i+1}\mathbf{D}_i\|_1 \le 1$.

Next, we prove by induction in reverse order from m down to 1.

For $i = m$, by definition of $\mathbf{E}_m$,

$$\begin{aligned}
\|\mathbf{E}_m\|_1 &\le \left\| ((\mathbf{F}_{m+1}+\mathbf{E}_{m+1})\widehat{\mathbf{D}_m})^{T_m} - (\mathbf{F}_{m+1}\mathbf{D}_m)^{T_m} \right\|_1 \\
&\le 5\delta m\kappa_{\text{glob}} T_m \exp(5\delta m\kappa_{\text{glob}} T_m) && \text{(by Lemma 76)} \\
&\le 5\sqrt{e}\delta m\kappa_{\text{glob}} T_m \le 10\delta m\kappa_{\text{glob}} T_m. && \text{(since } \delta \le \tfrac{1}{10m\kappa_{\text{glob}} T_m} \text{ by assumption Eq. (G.7))}
\end{aligned}$$

Now suppose the statement holds for the case of $i+1$, we then study the case of i. By definition of $\mathbf{E}_i$ we have

$$\begin{aligned}
\|\mathbf{E}_i\|_1 &\le \left\| ((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}_i})^{T_i} - (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i} \right\|_1 (1+\|\mathbf{E}_{i+1}\|_1) + \|\mathbf{E}_{i+1}\|_1 \\
&\le \|\mathbf{E}_{i+1}\|_1 + (1+\|\mathbf{E}_{i+1}\|_1)(2\|\mathbf{E}_{i+1}\|_1 + 5\delta m\kappa_{\text{glob}})\,\kappa_{\text{glob}}^2 T_i \cdot \exp\left((2\|\mathbf{E}_{i+1}\|_1 + 5\delta m\kappa_{\text{glob}})\,\kappa_{\text{glob}}^2 T_i\right) \\
& && \text{(by Lemma 76)}
\end{aligned}$$

By induction hypothesis $\|\mathbf{E}_{i+1}\|_1 \le (10\kappa_{\text{glob}})^{2(m-i)-1} \cdot \delta m \prod_{j=i+1}^{m} T_j$, we obtain

$$\begin{aligned}
&\|\mathbf{E}_{i+1}\|_1 + (1+\|\mathbf{E}_{i+1}\|_1)\,(2\|\mathbf{E}_{i+1}\|_1 + 5\delta m\kappa_{\text{glob}})\,\kappa_{\text{glob}}^2 T_i \\
\le\; & 4\kappa_{\text{glob}}^2 \left((10\kappa_{\text{glob}})^{2(m-i)-1} \cdot \delta m \prod_{j=i+1}^{m} T_j \right) T_i + 5\delta m\kappa_{\text{glob}}^3 T_i \\
\le\; & 10\kappa_{\text{glob}}^2 \cdot (10\kappa_{\text{glob}})^{2(m-i)-1} \delta m \cdot \prod_{j=i}^{m} T_j.
\end{aligned}$$

Also by δ bound Eq. (G.7)

$$\exp\left(\left(2\|\mathbf{E}_{i+1}\|_1 + 5\delta m\kappa_{\text{glob}}\right)\kappa_{\text{glob}}^2 T_i\right) \leq \exp\left(9\kappa_{\text{glob}}^2 \cdot (10\kappa_{\text{glob}})^{2(m-i)-1}\delta m \cdot \prod_{j=i}^{m} T_j\right) \leq e^{0.07}.$$

Since $9e^{0.07} < 10$ we have

$$\|\mathbf{E}_i\|_1 \leq \left\|((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i)^{T_i} - (\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\right\|_1 \leq 0.1(10\kappa_{\text{glob}})^{2(m-i)+1}\delta m \cdot \prod_{j=i}^{m} T_j. \tag{G.8}$$

completing the induction proof. ■

G.2.3. UPPER BOUND OF $\mathbf{Z}$

Finally we bound $\|\mathbf{Z}_i\|_1$ with the following Lemma 78.

Lemma 78 (Upper bound of $\|\mathbf{Z}_i\|_1$) *Using the same notation of Lemma 74 and assuming the same assumptions of Lemma 77, then the following inequality holds for any $i \in [m]$*

$$\|\mathbf{Z}_i\|_1 \leq 2^{m-i+1}\prod_{j=i}^{m}(T_j+1) \leq 4^{m-i+1}\prod_{j=1}^{m} T_j$$

Proof [Proof of Lemma 78] Recall the definition of $\mathbf{Z}_i$

$$\mathbf{Z}_i \coloneqq ((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i)^{T_i}\mathbf{Z}_{i+1} + \sum_{t_i=0}^{T_i-1}\left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i\right)^{t_i}(\mathbf{Z}_{i+1}+\mathbf{F}_{i+1}+\mathbf{E}_{i+1}),$$

Therefore

$$\|\mathbf{Z}_i\|_1 \leq \left\|\sum_{t_i=0}^{T_i}\left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i\right)^{t_i}\right\|_1 (\|\mathbf{Z}_{i+1}\|_1 + \|\mathbf{F}_{i+1}\|_1 + \|\mathbf{E}_{i+1}\|_1),$$

We will bound $\left\|\sum_{t_i=0}^{T_i}((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i)^{t_i}\right\|_1$ and $(\|\mathbf{Z}_{i+1}\|_1 + \|\mathbf{F}_{i+1}\|_1 + \|\mathbf{E}_{i+1}\|_1)$ separately. The former is bounded as

$$\begin{aligned}
&\left\|\sum_{t_i=0}^{T_i}((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i)^{t_i}\right\|_1 \leq \sum_{t_i=0}^{T_i}\left\|((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i)^{t_i}\right\|_1 && \text{(by triangle inequality)}\\
\leq &\sum_{t_i=0}^{T_i}\left(\left\|(\mathbf{F}_{i+1}\mathbf{D}_i)^{t_i}\right\|_1 + \left\|((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i)^{t_i} - (\mathbf{F}_{i+1}\mathbf{D}_i)^{t_i}\right\|_1\right) && \text{(by triangle inequality)}\\
\leq &(T_i+1) + \sum_{t_i=0}^{T_i}\left\|((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\widehat{\mathbf{D}}_i)^{t_i} - (\mathbf{F}_{i+1}\mathbf{D}_i)^{t_i}\right\|_1 && \text{(since } \|\mathbf{F}_{i+1}\mathbf{D}_i\|_1 \leq 1 \text{ by Lemma 75)}\\
\leq &(T_i+1)\left(1 + 0.1(10\kappa_{\text{glob}})^{2(m-i)+1}\delta m \cdot \prod_{j=i}^{m} T_j\right) && \text{(by Eq. (G.8))}\\
\leq &1.1(T_i+1) && \text{(by } \delta \leq \tfrac{1}{m(10\kappa_{\text{glob}})^{2m-1}\prod_{i\in[m]}T_i}\text{, see Eq. (G.7))}
\end{aligned}$$

For $\|\mathbf{F}_{i+1}+\mathbf{E}_{i+1}+\mathbf{Z}_{i+1}\|_1$ we have

$$\begin{aligned}&\|\mathbf{F}_{i+1}+\mathbf{E}_{i+1}+\mathbf{Z}_{i+1}\|_1 \le \|\mathbf{F}_{i+1}\|_1+\|\mathbf{E}_{i+1}\|_1+\|\mathbf{Z}_{i+1}\|_1\\ \le& 1+(10\kappa_{\text{glob}})^{2(m-i)-1}\cdot\delta m\prod_{j=i+1}^{m}T_j+\|\mathbf{Z}_{i+1}\|_1 \le 1.01+\|\mathbf{Z}_{i+1}\|_1 \quad \text{(by } \delta \text{ bound Eq. (G.7))}\end{aligned}$$

Consequently $\|\mathbf{Z}_i\|_1 \le 1.1(T_i+1)\,(1.01+\|\mathbf{Z}_{i+1}\|_1)$. By induction we have $\|\mathbf{Z}_i\|_1 \le 2^{m-i+1}\prod_{j=i}^{m}(T_j+1)$. ■

G.3. Finishing the proof of Theorem 70

We are ready to finish the proof of Theorem 70 (general initialization).

Proof [Proof of Theorem 70] By Lemma 74 we have

$$\mathbf{\Delta}(\widehat{\texttt{BSLS}_1}(\mathbf{x}^{(0)})) \le (\mathbf{F}_1+\mathbf{E}_1)\mathbf{\Delta}(\mathbf{x}^{(0)})+2\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{Z}_1\mathbf{1}.$$

Since $f(\mathbf{x})-f^\star=\|\mathbf{\Delta}(\mathbf{x})\|_1$, we obtain

$$f(\widehat{\texttt{BSLS}_1}(\mathbf{x}^{(0)}))-f^\star \le (\|\mathbf{F}_1\|_1+\|\mathbf{E}_1\|_1)(f(\mathbf{x}^{(0)})-f^\star)+2\delta m L_m\|\mathbf{x}^\star\|_2^2\|\mathbf{Z}_1\|_1$$

Plugging in the bound of $\|\mathbf{F}_1\|_1$ (from Lemma 75), $\|\mathbf{E}_1\|_1$ (from Lemma 77), and $\|\mathbf{Z}_1\|_1$ (from Lemma 78), we obtain

$$\begin{aligned}&f(\widehat{\texttt{BSLS}_1}(\mathbf{x}^{(0)}))-f^\star\\ \le&\left(\frac{\epsilon}{f(\mathbf{x}^{(0)})-f^\star}+(10\kappa_{\text{glob}})^{2m-1}\cdot\delta m\prod_{i=1}^{m}T_i\right)(f(\mathbf{x}^{(0)})-f^\star)+2\cdot 4^m\delta m L_m\|\mathbf{x}^\star\|_2^2\cdot\prod_{i=1}^{m}T_i.\end{aligned}$$

By δ bound

$$\delta \le \prod_{i\in[m]}T_i^{-1}\min\left\{\frac{1}{m\cdot(10\kappa_{\text{glob}})^{2m-1}},\frac{\epsilon}{m(10\kappa_{\text{glob}})^{2m-1}\cdot(f(\mathbf{x}^{(0)})-f^\star)},\frac{\epsilon}{4^{m+1}mL_m\|\mathbf{x}^\star\|_2^2}\right\},$$

we immediately obtain $f(\widehat{\texttt{BSLS}_1}(\mathbf{x}^{(0)}))-f^\star\le 3\epsilon$, completing the proof of Theorem 70. ■

Appendix H. Proof of AcBSLS under finite-precision arithmetic

In this section, we will prove AcBSLS under finite-precision arithmetic.

In Theorem 21, we specialized our initialization of $\mathbf{x}^{(0)},\mathbf{v}^{(0)}$ both to $\mathbf{0}$ to simplify the exposition of the theorem. In fact, we can (and will) prove the following general (but less clean) version with arbitrary $\mathbf{x}^{(0)},\mathbf{v}^{(0)}$.

Theorem 79 (AcBSLS under finite-precision arithmetic, general initialization) *Consider multiscale optimization problem defined in Theorem 1, for any initialization $(\mathbf{x}^{(0)},\mathbf{v}^{(0)})$ and $\epsilon>0$, assuming Requirement 2 with*

$$\delta^{-1}\ge\left(\prod_{i\in[m]}T_i\right)\cdot\max\left\{2\cdot(10\kappa_{\text{glob}}^2)^{2m-1},2\cdot(10\kappa_{\text{glob}}^2)^{2m-1}m\cdot\frac{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}{\epsilon},4\cdot 3^{m+1}\cdot mL_m\kappa_{\text{glob}}\frac{\|\mathbf{x}^\star\|_2^2}{\epsilon}\right\}$$

then $\psi(\widehat{\texttt{AcBSLS}}_1(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})) \le 3\epsilon$ *provided that* $T_1, \dots, T_m$ *satisfy* (C.1)*, which we restate here for ease of reference*

$$T_1 \ge \sqrt{\kappa_1} \log\left(\frac{\psi(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})}{\epsilon}\right), \qquad T_i \ge \sqrt{\kappa_i}(\log(4\kappa_{\mathrm{glob}}^4) + 1), \quad \text{for } i = 2, \dots, m.$$

We can also achieve the same asymptotic sample complexity (up to constant factors suppressed in the $\mathcal{O}(\cdot)$*) when* $\{(\mu_i, L_i), i \in [m]\}$ *are unknown and only* m*,* μ_1*,* L_m *and* $\pi_\kappa = \prod_{i=1}^m \kappa_i$ *are known.*

Theorem 21 is clearly a corollary of Theorem 79 since

$$4 \cdot 3^{m+1} \cdot mL_m\kappa_{\mathrm{glob}} \frac{\|\mathbf{x}^\star\|_2^2}{\epsilon} \le 4m(10\kappa_{\mathrm{glob}}^2)^{2m-1} \frac{\psi(\mathbf{0}, \mathbf{0})}{\epsilon}.$$

The proof of Theorem 79 is structured as follows. We first define two vector potentials and establish their relations in Appendix H.1. We then study the progress of one (inexact) $\widehat{\texttt{AGD}}$ step on these vector potentials in Appendix H.2, and inductively estimate the progress of $\widehat{\texttt{AcBSLS}}_i$ by matrix inequalities for all $i \in [m]$ in descent order (see Appendix H.3). The proof of Theorem 79 is then finished in Appendix H.4. As before, the last part regarding the case where $\{(\mu_i, L_i), i \in [m]\}$ are unknown follows from our black-box reduction in Proposition 13 (in the same way as in the proof of Theorem 6).

H.1. Introduction of vector potentials and their relations

We introduce a few more notation to simplify the presentation. For any $(\mathbf{x}, \mathbf{v})$ and $i \in [m]$, define

$$\Delta_i^{\max}(\mathbf{x}, \mathbf{v}) \coloneqq \max\{\Delta_i(\mathbf{x}), \Delta_i(\mathbf{v})\}, \qquad r_i^{\max}(\mathbf{x}, \mathbf{v}) \coloneqq \max\{r_i(\mathbf{x}), r_i(\mathbf{v})\}. \tag{H.1}$$

Define a series of vector-valued potential functions $\boldsymbol{\phi}_1^\Delta, \dots, \boldsymbol{\phi}_m^\Delta$ and $\boldsymbol{\phi}_1^r, \dots, \boldsymbol{\phi}_m^r$:

$$\boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) \coloneqq \begin{bmatrix} \Delta_1^{\max}(\mathbf{x}, \mathbf{v}) \\ \vdots \\ \Delta_{i-1}^{\max}(\mathbf{x}, \mathbf{v}) \\ \frac{1}{2}\psi_i(\mathbf{x}, \mathbf{v}) \\ \vdots \\ \frac{1}{2}\psi_m(\mathbf{x}, \mathbf{v}) \end{bmatrix}, \quad \boldsymbol{\phi}_i^r(\mathbf{x}, \mathbf{v}) \coloneqq \begin{bmatrix} r_1^{\max}(\mathbf{x}, \mathbf{v}) \\ \vdots \\ r_{i-1}^{\max}(\mathbf{x}, \mathbf{v}) \\ \frac{1}{2}\psi_i(\mathbf{x}, \mathbf{v}) \\ \vdots \\ \frac{1}{2}\psi_m(\mathbf{x}, \mathbf{v}) \end{bmatrix}. \tag{H.2}$$

We establish two lemmas on the relations of vector potentials $\boldsymbol{\phi}_i^\Delta$ and $\boldsymbol{\phi}_i^r$ for varying i. Lemma 80 bounds the maximum of two vector potentials; Lemma 81 bounds the sum of two vector potentials.

H.1.1. BOUNDING THE MAXIMUM OF TWO VECTOR POTENTIALS

Lemma 80 *Consider multiscale optimization problem defined in Theorem 1, for any* $\mathbf{x}, \mathbf{v}$*, for any* $i \in [m-1]$*, the following two matrix inequalities hold (recall* $\le$ *denotes entry-wise inequality)*

$$\max\left\{\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}), \boldsymbol{\phi}_{i+1}^\Delta(\mathbf{v}, \mathbf{v})\right\} \le \begin{bmatrix} \mathbf{I}_{i-1} & 0 \\ 0 & 2\kappa_{\mathrm{glob}} \cdot \mathbf{I}_{m-i+1} \end{bmatrix} \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}),$$

$$\max\left\{\boldsymbol{\phi}_{i+1}^r(\mathbf{x}, \mathbf{x}), \boldsymbol{\phi}_{i+1}^r(\mathbf{v}, \mathbf{v})\right\} \le \begin{bmatrix} \mathbf{I}_{i-1} & 0 \\ 0 & 2\kappa_{\mathrm{glob}} \cdot \mathbf{I}_{m-i+1} \end{bmatrix} \boldsymbol{\phi}_i^r(\mathbf{x}, \mathbf{v}).$$

Proof [Proof of Lemma 80] We study $\mathbf{e}_j^\top \max\left\{\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}), \boldsymbol{\phi}_{i+1}^\Delta(\mathbf{v}, \mathbf{v})\right\}$ for three possible cases: $j < i$, $j = i$, or $j > i$.

Case of $j < i$. By definition of $\boldsymbol{\phi}_i^\Delta$ we have $\mathbf{e}_j^\top \boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}) = \Delta_j(\mathbf{x})$ for any $\mathbf{x}$. Thus

$$\mathbf{e}_j^\top \max\left\{\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}), \boldsymbol{\phi}_{i+1}^\Delta(\mathbf{v}, \mathbf{v})\right\} = \max\left\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\right\} = \mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}),$$

where the last equality is by definition of $\mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v})$.

Case of $j = i$. Again by definition of $\boldsymbol{\phi}_i^\Delta$

$$\begin{aligned}
&\mathbf{e}_i^\top \max\left\{\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}), \boldsymbol{\phi}_{i+1}^\Delta(\mathbf{v}, \mathbf{v})\right\} = \max\left\{\Delta_i(\mathbf{x}), \Delta_i(\mathbf{v})\right\} \\
\le& \max\{\Delta_i(\mathbf{x}), \kappa_i r_i(\mathbf{v})\} \le \kappa_i \psi_i(\mathbf{x}, \mathbf{v}) && \text{(by definition of } \psi_i) \\
=& 2\kappa_i \cdot \mathbf{e}_i^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}),
\end{aligned}$$

where the last equality is by definition of $\boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v})$ since $\mathbf{e}_i^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) = \frac{1}{2}\psi_i(\mathbf{x}, \mathbf{v})$.

Case of $j > i$. By definition of $\boldsymbol{\phi}_{i+1}^\Delta$:

$$\begin{aligned}
&\mathbf{e}_j^\top \max\left\{\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}), \boldsymbol{\phi}_{i+1}^\Delta(\mathbf{v}, \mathbf{v})\right\} = \frac{1}{2}\max\left\{\psi_j(\mathbf{x}, \mathbf{x}), \psi_j(\mathbf{v}, \mathbf{v})\right\} && \text{(by definition)} \\
=& \frac{1}{2}\max\left\{\Delta_j(\mathbf{x}) + r_j(\mathbf{x}), \Delta_j(\mathbf{v}) + r_j(\mathbf{v})\right\} \le \frac{1}{2}(1+\kappa_j)\psi_j(\mathbf{x}, \mathbf{v}) \le \kappa_j \psi_j(\mathbf{x}, \mathbf{v}) \\
=& 2\kappa_j \mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) && \text{(by definition)}
\end{aligned}$$

Concatenating the above three inequalities yields the first statement of the lemma. The second statement holds for the same reason. ■

H.1.2. BOUNDING THE SUM OF TWO VECTOR POTENTIALS

Lemma 81 *Consider multiscale optimization problem defined in Theorem 1, for any $\mathbf{x}, \mathbf{v}$, for any $i \in [m-1]$, the following two inequalities hold*

$$\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}) + \boldsymbol{\phi}_{i+1}^r(\mathbf{v}, \mathbf{v}) \le \begin{bmatrix} 2\kappa_{\text{glob}}\mathbf{I}_i & & \\ & 2 & \\ & & 2\kappa_{\text{glob}}\mathbf{I}_{m-i-1} \end{bmatrix} \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}),$$

$$\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}) + \boldsymbol{\phi}_{i+1}^r(\mathbf{v}, \mathbf{v}) \le \begin{bmatrix} 2\kappa_{\text{glob}}\mathbf{I}_i & & \\ & 2 & \\ & & 2\kappa_{\text{glob}}\mathbf{I}_{m-i-1} \end{bmatrix} \boldsymbol{\phi}_i^r(\mathbf{x}, \mathbf{v}).$$

Proof [Proof of Lemma 81] We study $\mathbf{e}_j^\top(\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}) + \boldsymbol{\phi}_{i+1}^r(\mathbf{v}, \mathbf{v}))$ for three possible cases: $j < i$, $j = i$, or $j > i$.

Case of $j < i$. By definition of $\boldsymbol{\phi}_i^\Delta$ and $\boldsymbol{\phi}_i^r$ we have $\mathbf{e}_j^\top \boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}) = \Delta_j(\mathbf{x})$ and $\mathbf{e}_j^\top \boldsymbol{\phi}_{i+1}^r(\mathbf{v}, \mathbf{v}) = r_j(\mathbf{v})$. Thus

$$\begin{aligned}
&\mathbf{e}_j^\top(\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}) + \boldsymbol{\phi}_{i+1}^r(\mathbf{v}, \mathbf{v})) = \Delta_j(\mathbf{x}) + r_j(\mathbf{v}) \\
\le& \Delta_j(\mathbf{x}) + \kappa_j \Delta_j(\mathbf{v}) \le 2\kappa_j \max\{\Delta_j(\mathbf{x}), \Delta_j(\mathbf{v})\} \\
=& 2\kappa_j \mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}). && \text{(by definition of } \boldsymbol{\phi}_i^\Delta)
\end{aligned}$$

Similarly $\mathbf{e}_j^\top(\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x}, \mathbf{x}) + \boldsymbol{\phi}_{i+1}^r(\mathbf{v}, \mathbf{v})) \le 2\kappa_j \mathbf{e}_j^\top \boldsymbol{\phi}_i^r(\mathbf{x}, \mathbf{v})$.

Case of $j = i$. Similarly

$$\begin{aligned}&\mathbf{e}_i^\top(\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x},\mathbf{x}) + \boldsymbol{\phi}_{i+1}^r(\mathbf{v},\mathbf{v})) = \Delta_j(\mathbf{x}) + r_j(\mathbf{v}) && \text{(by definition)}\\ =&\psi_i(\mathbf{x},\mathbf{v}) = 2\mathbf{e}_i^\top\boldsymbol{\phi}_i^\Delta(\mathbf{x},\mathbf{v}) = 2\mathbf{e}_i^\top\boldsymbol{\phi}_i^r(\mathbf{x},\mathbf{v}). && \text{(by definition)}\end{aligned}$$

Case of $j \geq i$. By definition we have $\mathbf{e}_j^\top\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x},\mathbf{x}) = \frac{1}{2}\psi_j(\mathbf{x},\mathbf{x})$ and $\mathbf{e}_j^\top\boldsymbol{\phi}_{i+1}^r(\mathbf{v},\mathbf{v}) = \frac{1}{2}\psi_j(\mathbf{v},\mathbf{v})$. Thus

$$\begin{aligned}&\mathbf{e}_j^\top(\boldsymbol{\phi}_{i+1}^\Delta(\mathbf{x},\mathbf{x}) + \boldsymbol{\phi}_{i+1}^r(\mathbf{v},\mathbf{v})) = \frac{1}{2}\psi_j(\mathbf{x},\mathbf{x}) + \frac{1}{2}\psi_j(\mathbf{v},\mathbf{v}) && \text{(by definition)}\\ =&\frac{1}{2}(\Delta_j(\mathbf{x}) + r_j(\mathbf{v}) + \Delta_j(\mathbf{v}) + r_j(\mathbf{v}))\\ \leq&\frac{1}{2}(1+\kappa_j)\psi_j(\mathbf{x},\mathbf{v}) \leq \kappa_j\psi_j(\mathbf{x},\mathbf{v}) = 2\kappa_j\cdot\mathbf{e}_j^\top\boldsymbol{\phi}_i^\Delta(\mathbf{x},\mathbf{v}) = 2\kappa_j\cdot\mathbf{e}_j^\top\boldsymbol{\phi}_i^r(\mathbf{x},\mathbf{v}).\end{aligned}$$

Concatenating the above inequalities completes the proof. ■

H.2. Progress of AGD step under finite arithmetic

In this subsection, we study the effect of one inexact AGD (also denoted as $\widehat{\text{AGD}}$ step) on the vector potentials $\boldsymbol{\phi}_i^\Delta$ and $\boldsymbol{\phi}_i^r$. The main goal of this subsection is to prove the following Lemma 82.

Lemma 82 (Progress of one $\widehat{\text{AGD}}$ step under finite arithmetic) *Consider multiscale optimization problem defined in Theorem 1, assuming Requirement 2, then for any $\mathbf{x}, \mathbf{v}$ and $i \in [m]$, the following two inequalities hold*

(a) $\boldsymbol{\phi}_i^\Delta(\widehat{\text{AGD}}(\mathbf{x},\mathbf{v};L_i,\mu_i)) \leq (\mathbf{I} + 10\delta\kappa_{\text{glob}}^2\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i\boldsymbol{\phi}_i^\Delta(\mathbf{x},\mathbf{v}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{1}.$

(b) $\boldsymbol{\phi}_i^r(\widehat{\text{AGD}}(\mathbf{x},\mathbf{v};L_i,\mu_i)) \leq (\mathbf{I} + 10\delta\kappa_{\text{glob}}^2\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i\boldsymbol{\phi}_i^r(\mathbf{x},\mathbf{v}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{1}.$

where $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_m$ are $m \times m$ diagonal matrices defined by

$$\mathbf{D}_i = \begin{bmatrix}\mathbf{I}_{i-1} & & \\ & 1-\kappa_i^{-\frac{1}{2}} & \\ & & 2\kappa_{\text{glob}}^2\mathbf{I}_{m-i}\end{bmatrix} \tag{H.3}$$

To simplify the notation we will define (throughout this section)

$$\widehat{\mathbf{D}_i} \coloneqq (\mathbf{I} + 10\delta\kappa_{\text{glob}}^2\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i. \tag{H.4}$$

We will prove Lemma 82 in three steps. First, we first bound the perturbation of residual r_j and functional error Δ_j under multiplicative error in Lemma 83. Then we bound the potential $r_j^{\max}, \Delta_j^{\max}$, and ψ_j, and the vector potential $\boldsymbol{\phi}_j^r$ and $\boldsymbol{\phi}_j^\Delta$ in Lemma 84. The proof of Lemma 82 is finished in Appendix H.2.3.

H.2.1. SENSITIVITY OF RESIDUAL AND FUNCTIONAL ERROR UNDER MULTIPLICATIVE ERROR

In this subsubsection, we establish the first supporting lemma for Lemma 82.

Lemma 83 (Sensitivity of residual and functional error under multiplicative error) *Assuming* $\widehat{\mathbf{x}}, \mathbf{x}$ *satisfies*

$$|\widehat{\mathbf{x}} - \mathbf{x}| \leq \delta |\mathbf{x}| \tag{H.5}$$

for some $\delta < 1$*, then for any* $j \in [m]$*,*

(a) $\Delta_j(\widehat{\mathbf{x}}) \leq \Delta_j(\mathbf{x}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m \Delta_k(\mathbf{x}) + 2\delta L_j \|\mathbf{x}^\star\|_2^2$

(b) $r_j(\widehat{\mathbf{x}}) \leq r_j(\mathbf{x}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m r_k(\mathbf{x}) + 2\delta\mu_j \|\mathbf{x}^\star\|_2^2$

Proof [Proof of Lemma 83]

(a) Same as Lemma 73.

(b) Let $\boldsymbol{\epsilon} \coloneqq \widehat{\mathbf{x}} - \mathbf{x}$, then by Cauchy-Schwartz inequality,

$$\|\mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x}^\star)\|_2^2 = \|\mathbf{P}_j(\mathbf{x} - \mathbf{x}^\star + \boldsymbol{\epsilon})\|_2^2 \leq (1+\delta)\|\mathbf{P}_j(\mathbf{x} - \mathbf{x}^\star)\|_2^2 + 2\delta^{-1}\|\mathbf{P}_j\boldsymbol{\epsilon}\|_2^2 \tag{H.6}$$

By assumption (H.5) we have

$$\|\mathbf{P}_j\boldsymbol{\epsilon}\|_2^2 \leq \|\boldsymbol{\epsilon}\|_2^2 \leq \delta^2\|\mathbf{x}\|_2^2 \leq 2\delta^2\|\mathbf{x} - \mathbf{x}^\star\|_2^2 + 2\delta^2\|\mathbf{x}^\star\|_2^2 \leq 2\delta^2 \sum_{k=1}^m \|\mathbf{P}_k(\mathbf{x} - \mathbf{x}^\star)\|_2^2 + 2\delta^2\|\mathbf{x}^\star\|_2^2 \tag{H.7}$$

Combining (H.6) and (H.7) yields

$$\|\mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x}^\star)\|_2^2 \leq (1+\delta)\|\mathbf{P}_j(\mathbf{x} - \mathbf{x}^\star)\|_2^2 + 4\delta \sum_{k=1}^m \|\mathbf{P}_k(\mathbf{x} - \mathbf{x}^\star)\|_2^2 + 4\delta\|\mathbf{x}^\star\|_2^2 \tag{H.8}$$

It follows that

$$\begin{aligned}
r_j(\widehat{\mathbf{x}}) &\coloneqq \frac{1}{2}\mu_j\|\mathbf{P}_j(\widehat{\mathbf{x}} - \mathbf{x}^\star)\|_2^2 && \text{(by definition of } r_j\text{)} \\
&\leq (1+\delta)\cdot\frac{1}{2}\mu_j\|\mathbf{P}_j(\mathbf{x} - \mathbf{x}^\star)\|_2^2 + 2\mu_j\delta\sum_{k=1}^m \|\mathbf{P}_k(\mathbf{x} - \mathbf{x}^\star)\|_2^2 + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2 && \text{(by (H.8))} \\
&= (1+\delta)r_j(\mathbf{x}) + 4\delta\sum_{k=1}^m \frac{\mu_j}{\mu_k} r_k(\mathbf{x}) + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2 && \text{(by definition of } r_j\text{'s)} \\
&\leq (1+\delta)r_j(\mathbf{x}) + 4\delta\kappa_{\text{glob}}\sum_{k=1}^m r_k(\mathbf{x}) + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2 && \text{(since } \tfrac{\mu_j}{\mu_k} \leq \kappa_{\text{glob}} \text{ for any } j,k \in [m]\text{)} \\
&\leq r_j(\mathbf{x}) + 5\delta\kappa_{\text{glob}}\sum_{k=1}^m r_k(\mathbf{x}) + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2.
\end{aligned}$$

■

H.2.2. SENSITIVITY OF POTENTIALS UNDER MULTIPLICATIVE ERROR

In this subsubsection, we establish the second supporting lemma for Lemma 82.

Lemma 84 (Sensitivity of potentials under multiplicative error) *Assuming* $\widehat{\mathbf{x}}, \mathbf{x}, \widehat{\mathbf{v}}, \mathbf{v}$ *satisfies*

$$|\widehat{\mathbf{x}} - \mathbf{x}| \leq \delta|\mathbf{x}|, \quad |\widehat{\mathbf{v}} - \mathbf{v}| \leq \delta|\mathbf{v}|$$

for some $\delta < 1$. *Then for any* $j \in [m]$,

(a) $r_j^{\max}(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) \leq r_j^{\max}(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m r_k^{\max}(\mathbf{x}, \mathbf{v}) + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2$.

(b) $\Delta_j^{\max}(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) \leq \Delta_j^{\max}(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m \Delta_k^{\max}(\mathbf{x}, \mathbf{v}) + 2\delta L_j\|\mathbf{x}^\star\|_2^2$.

(c) $\psi_j(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) \leq \psi_j(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m \psi_k(\mathbf{x}, \mathbf{v}) + 4\delta L_j\|\mathbf{x}^\star\|_2^2$.

(d) $\boldsymbol{\phi}_i^\Delta(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) \leq (\mathbf{I} + 10\delta\kappa_{\text{glob}}^2 \mathbf{1}\mathbf{1}^\top)\boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{1}$.

(e) $\boldsymbol{\phi}_i^r(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) \leq (\mathbf{I} + 10\delta\kappa_{\text{glob}}^2 \mathbf{1}\mathbf{1}^\top)\boldsymbol{\phi}_i^r(\mathbf{x}, \mathbf{v}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{1}$.

Proof [Proof of Lemma 84]

(a) By Lemma 83, for any $j \in [m]$,

$$\begin{aligned}
&r_j^{\max}(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) = \max\{r_j(\widehat{\mathbf{x}}), r_j(\widehat{\mathbf{x}})\} && \text{(by definition of } r_j^{\max} \text{ (H.1))} \\
\leq &\max\left\{r_j(\mathbf{x}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m r_k(\mathbf{x}) + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2, r_j(\mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m r_k(\mathbf{v}) + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2\right\} \\
\leq &r_j^{\max}(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m r_k^{\max}(\mathbf{x}, \mathbf{v}) + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2 && \text{(by definition of } r_j^{\max})
\end{aligned}$$

(b) Holds for the same reason as (a).

(c) For any $j \in [m]$, by Lemma 83,

$$\begin{aligned}
&\psi_j(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) = \Delta_j(\widehat{\mathbf{x}}) + r_j(\widehat{\mathbf{v}}) && \text{(by definition of } \psi_j) \\
\leq &\Delta_j(\mathbf{x}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m \Delta_k(\mathbf{x}) + 2\delta L_j\|\mathbf{x}^\star\|_2^2 + r_j(\mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m r_k(\mathbf{v}) + 2\delta\mu_j\|\mathbf{x}^\star\|_2^2 \\
& && \text{(by Lemma 83)} \\
\leq &\psi_j(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^m \psi_k(\mathbf{x}, \mathbf{v}) + 4\delta L_j\|\mathbf{x}^\star\|_2^2. && \text{(by definition of } \psi_j \text{ and } \mu_j \leq L_j)
\end{aligned}$$

(d) We will prove (d) by considering $\mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta$ for two different cases: $j < i$ or $j \geq i$ (recall $\mathbf{e}_j$ is defined as the j-th unit vector). For $j < i$, by definition of $\boldsymbol{\phi}_i^\Delta$ (H.2), we have

$$\begin{aligned}
&\mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) = \Delta_j^{\max}(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) && \text{(definition of } \boldsymbol{\phi}_i^\Delta\text{)} \\
\leq &\Delta_j^{\max}(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^{m} \Delta_k^{\max}(\mathbf{x}, \mathbf{v}) + 2\delta L_j \|\mathbf{x}^\star\|_2^2 && \text{(by (b))} \\
\leq &\Delta_j^{\max}(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \left(\sum_{k<i} \Delta_k^{\max}(\mathbf{x}, \mathbf{v}) + \kappa_{\text{glob}} \sum_{k \geq i} \psi_k(\mathbf{x}, \mathbf{v}) \right) + 2\delta L_j \|\mathbf{x}^\star\|_2^2 \\
&\text{(since } \Delta_k^{\max}(\mathbf{x}, \mathbf{v}) = \max\{\Delta_k(\mathbf{x}), \Delta_k(\mathbf{v})\} \leq \Delta_k(\mathbf{x}) + \Delta_k(\mathbf{v}) \leq \Delta_k(\mathbf{x}) + \kappa_k r_k(\mathbf{v}) \leq \kappa_k \psi_k(\mathbf{x}, \mathbf{v})\text{)} \\
\leq &\Delta_j^{\max}(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}}^2 \left(\sum_{k<i} \Delta_k^{\max}(\mathbf{x}, \mathbf{v}) + \sum_{k \geq i} \psi_k(\mathbf{x}, \mathbf{v}) \right) + 2\delta L_j \|\mathbf{x}^\star\|_2^2 \\
= &\mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}}^2 \sum_{k=1}^{m} \mathbf{e}_k^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) + 2\delta L_j \|\mathbf{x}^\star\|_2^2 && \text{(definition of } \boldsymbol{\phi}_i^\Delta\text{)}
\end{aligned}$$

For $j \geq i$, by definition,

$$\begin{aligned}
&\mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) = \psi_j(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) && \text{(definition of } \boldsymbol{\phi}_i^\Delta\text{)} \\
\leq &\psi_j(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \sum_{k=1}^{m} \psi_k(\mathbf{x}, \mathbf{v}) + 4\delta L_j \|\mathbf{x}^\star\|_2^2 && \text{(by (c))} \\
\leq &\psi_j(\mathbf{x}, \mathbf{v}) + 5\delta\kappa_{\text{glob}} \left(2 \sum_{k<i} \Delta_k^{\max}(\mathbf{x}, \mathbf{v}) + \sum_{k \geq i} \psi_k(\mathbf{x}, \mathbf{v}) \right) + 4\delta L_j \|\mathbf{x}^\star\|_2^2 \\
&\text{(since } \psi_k(\mathbf{x}, \mathbf{v}) = \Delta_k(\mathbf{x}) + r_k(\mathbf{v}) \leq \Delta_k(\mathbf{x}) + \Delta_k(\mathbf{v}) \leq 2\Delta_k^{\max}(\mathbf{x}, \mathbf{v})\text{)} \\
\leq &\psi_j(\mathbf{x}, \mathbf{v}) + 10\delta\kappa_{\text{glob}} \left(\sum_{k<i} \Delta_k^{\max}(\mathbf{x}, \mathbf{v}) + \sum_{k \geq i} \psi_k(\mathbf{x}, \mathbf{v}) \right) + 4\delta L_j \|\mathbf{x}^\star\|_2^2 \\
= &\mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) + 10\delta\kappa_{\text{glob}} \sum_{k=1}^{m} \mathbf{e}_k^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) + 4\delta L_j \|\mathbf{x}^\star\|_2^2 && \text{(definition of } \boldsymbol{\phi}_i^\Delta\text{)}
\end{aligned}$$

In matrix form we arrive at

$$\boldsymbol{\phi}_i^\Delta(\widehat{\mathbf{x}}, \widehat{\mathbf{v}}) \leq (\mathbf{I} + 10\delta\kappa_{\text{glob}}^2 \mathbf{1}\mathbf{1}^\top) \boldsymbol{\phi}_i^\Delta(\mathbf{x}, \mathbf{v}) + 4\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{1}.$$

(e) Holds for the same reason as (d).

■

H.2.3. FINISHING THE PROOF OF LEMMA 82

We are ready to finish the proof of Lemma 82.

Proof [Proof of Lemma 82] By Lemma 16 from exact AGD analysis we have

$$\boldsymbol{\phi}_i^{\Delta}(\texttt{AGD}(\mathbf{x}, \mathbf{v}; L_i, \mu_i)) \leq \mathbf{D}_i \boldsymbol{\phi}_i^{\Delta}(\mathbf{x}, \mathbf{v}),$$

then applying Lemma 84 shows (a). (b) holds for the same reason. ■

H.3. Inductively bound the progress of inexact AcBSLS by matrix inequalities

In the following lemma, we iteratively construct the bound of vector potentials $\boldsymbol{\phi}^{\Delta}$ and $\boldsymbol{\phi}^{r}$ after executing $\widehat{\texttt{AcBSLS}}_i$.

Lemma 85 (Estimate the progress of $\widehat{\texttt{AcBSLS}}_i$ by matrix inequalities) *Consider multiscale optimization problem defined in Theorem 1, and assuming Requirement 2, define the following three sequences of $m \times m$ matrices $\{\mathbf{F}_i\}_{i=1}^{m+1}$, $\{\mathbf{E}_i\}_{i=1}^{m+1}$, $\{\mathbf{Z}_i\}_{i=1}^{m+1}$:*

$$\mathbf{F}_{m+1} \coloneqq \mathbf{I}, \qquad \mathbf{E}_{m+1} \coloneqq \mathbf{0}, \qquad \mathbf{Z}_{m+1} \coloneqq \mathbf{0}$$

and for $i = m, m-1$ down to 1, define

$$\widehat{\mathbf{F}_{i+1}} \coloneqq \begin{bmatrix} [\mathbf{I}_{i-1} | \mathbf{0}_{(i-1)\times(m-i+1)}](\mathbf{F}_{i+1} + \mathbf{E}_{i+1}) \begin{bmatrix} \mathbf{I}_{i-1} & \mathbf{0} \\ \mathbf{0} & 2\kappa_{\text{glob}} \cdot \mathbf{I}_{m-i+1} \end{bmatrix} \\ \mathbf{e}_i^{\top}(\mathbf{F}_{i+1} + \mathbf{E}_{i+1}) \begin{bmatrix} \kappa_{\text{glob}}\mathbf{I}_i & & \\ & 1 & \\ & & \kappa_{\text{glob}}\mathbf{I}_{m-i-1} \end{bmatrix} \\ [\mathbf{0}_{(m-i)\times i} | \mathbf{I}_{m-i}](\mathbf{F}_{i+1} + \mathbf{E}_{i+1}) \end{bmatrix}$$

$$\mathbf{F}_i \coloneqq (\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i)^{T_i}, \quad \mathbf{E}_i \coloneqq (\mathbf{K}_i \widehat{\mathbf{F}_{i+1}} \widehat{\mathbf{D}_i})^{T_i} - (\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i)^{T_i},$$

$$\mathbf{Z}_i \coloneqq \sum_{t_i=0}^{T_i-1} (\mathbf{K}_i \widehat{\mathbf{F}_{i+1}} \widehat{\mathbf{D}_i})^{t_i} \left(\mathbf{K}_i \widehat{\mathbf{F}_{i+1}} + 2\mathbf{Z}_{i+1} \right).$$

where $\mathbf{D}_i, \widehat{\mathbf{D}_i}$ were defined in Eqs. (H.3) and (H.4), and $\mathbf{K}_i$ is defined by

$$\mathbf{K}_i \coloneqq \begin{bmatrix} \mathbf{I}_i & \\ & 2\kappa_{\text{glob}}\mathbf{I}_{m-i}. \end{bmatrix}$$

Then,

(a) For any $i \in [m+1]$, $\mathbf{F}_i$ are non-negative diagonal matrices, $\mathbf{E}_i$ and $\mathbf{Z}_i$ are non-negative matrices.

(b) For any $i \in [m]$, the following bound holds

$$\boldsymbol{\phi}_i^{\Delta}(\widehat{\texttt{AcBSLS}}_i(\mathbf{x}, \mathbf{v})) \leq (\mathbf{F}_i + \mathbf{E}_i)\boldsymbol{\phi}_i^{\Delta}(\mathbf{x}, \mathbf{v}) + 4\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{Z}_i \mathbf{1},$$
$$\boldsymbol{\phi}_i^{r}(\widehat{\texttt{AcBSLS}}_i(\mathbf{x}, \mathbf{v})) \leq (\mathbf{F}_i + \mathbf{E}_i)\boldsymbol{\phi}_i^{r}(\mathbf{x}, \mathbf{v}) + 4\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{Z}_i \mathbf{1}.$$

Proof [Proof of Lemma 85] The proof of (a) is the same as the proof for Lemma 74(a). We will prove the first inequalities in (b) by induction in reverse order (from m back to 1). The second inequality holds for the same reason.

The induction base apparently holds. Now assume for any $\mathbf{x}, \mathbf{v}$,

$$\boldsymbol{\phi}^{\Delta}_{i+1}(\widehat{\texttt{AcBSLS}}_{i+1}(\mathbf{x},\mathbf{v})) \le (\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\boldsymbol{\phi}^{\Delta}_{i+1}(\mathbf{x},\mathbf{v}) + 4\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{Z}_{i+1}\mathbf{1},$$

then we will show that

$$\boldsymbol{\phi}^{\Delta}_{i}(\widehat{\texttt{AcBSLS}}_{i}(\mathbf{x},\mathbf{v})) \le (\mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i})^{T_i}\boldsymbol{\phi}^{\Delta}_{i}(\mathbf{x},\mathbf{v}), +4\delta L_m\|\mathbf{x}^\star\|_2^2 \cdot \sum_{t_i=0}^{T_i-1}(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i})^{t_i}(\widehat{\mathbf{F}_{i+1}}+2\mathbf{Z}_{i+1})\mathbf{1},$$

To this end, let $\begin{bmatrix}\mathbf{x}^{(0)}\\ \mathbf{v}^{(0)}\end{bmatrix}, \begin{bmatrix}\tilde{\mathbf{x}}^{(0)}\\ \tilde{\mathbf{v}}^{(0)}\end{bmatrix}, \cdots \begin{bmatrix}\mathbf{x}^{(T_i)}\\ \mathbf{v}^{(T_i)}\end{bmatrix}$ be the trajectory generated by running $\widehat{\texttt{AcBSLS}}_i$. Since $(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) = \widehat{\texttt{AGD}}(\mathbf{x}^{(t)}, \mathbf{v}^{(t)}; L_i, \mu_i)$, we have by Lemma 82,

$$\boldsymbol{\phi}^{\Delta}_{i}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) \le (\mathbf{I} + 10\delta\kappa_{\text{glob}}^2\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i\boldsymbol{\phi}^{\Delta}_{i}(\mathbf{x}^{(t)}, \mathbf{v}^{(t)}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{1},$$

We study the $\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}$ by considering 3 cases. For $j < i$,

$$\begin{aligned}
&\mathbf{e}_j^\top\boldsymbol{\phi}^{\Delta}_{i}(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) = \Delta_j^{\max}(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) = \max\{\Delta_j(\mathbf{x}^{(t+1)}), \Delta_j(\mathbf{v}^{(t+1)})\}\\
\le&\max\{\Delta_j^{\max}(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})), \Delta_j^{\max}(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)}))\}\\
=&\mathbf{e}_j^\top\max\{\boldsymbol{\phi}^{\Delta}_{i+1}(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})), \boldsymbol{\phi}^{\Delta}_{i+1}(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)}))\} \text{ (by definition of } \boldsymbol{\phi}^{\Delta}_{i+1})\\
\le&\mathbf{e}_j^\top\max\{(\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\boldsymbol{\phi}^{\Delta}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)}), (\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\boldsymbol{\phi}^{\Delta}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)})\} + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{e}_j^\top\mathbf{Z}_{i+1}\mathbf{1}\\
\le&\mathbf{e}_j^\top(\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\max\{\boldsymbol{\phi}^{\Delta}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)}), \boldsymbol{\phi}^{\Delta}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)})\} + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{e}_j^\top\mathbf{Z}_{i+1}\mathbf{1}\\
\le&\mathbf{e}_j^\top(\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\begin{bmatrix}\mathbf{I}_{i-1} & 0\\ 0 & 2\kappa_{\text{glob}}\cdot\mathbf{I}_{m-i+1}\end{bmatrix}\boldsymbol{\phi}^{\Delta}_{i}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{e}_j^\top\mathbf{Z}_{i+1}\mathbf{1}.
\end{aligned}$$

(by Lemma 80)

For $j = i$,

$$\begin{aligned}
&\mathbf{e}_i^\top\boldsymbol{\phi}^{\Delta}_{i}(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) = \frac{1}{2}\psi_i(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) = \frac{1}{2}\Delta_i(\mathbf{x}^{(t+1)}) + \frac{1}{2}r_i(\mathbf{v}^{(t+1)})\\
\le&\frac{1}{2}\Delta_i^{\max}(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})) + \frac{1}{2}r_i^{\max}(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)}))\\
=&\frac{1}{2}\mathbf{e}_i^\top\left(\boldsymbol{\phi}^{\Delta}_{i+1}(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})) + \boldsymbol{\phi}^{r}_{i+1}(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)}))\right)\\
\le&\frac{1}{2}\mathbf{e}_i^\top\left((\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\boldsymbol{\phi}^{\Delta}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)}) + (\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\boldsymbol{\phi}^{r}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)})\right) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{e}_i^\top\mathbf{Z}_{i+1}\mathbf{1}\\
\le&\mathbf{e}_i^\top(\mathbf{F}_{i+1}+\mathbf{E}_{i+1})\begin{bmatrix}\kappa_{\text{glob}}\mathbf{I}_i & & \\ & 1 & \\ & & \kappa_{\text{glob}}\mathbf{I}_{m-i-1}\end{bmatrix}\boldsymbol{\phi}^{\Delta}_{i}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{e}_i^\top\mathbf{Z}_{i+1}\mathbf{1}.
\end{aligned}$$

(by Lemma 81)

For $j > i$

$$\begin{aligned}
&\mathbf{e}_j^\top \boldsymbol{\phi}_i^\Delta(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) = \frac{1}{2}\psi_j(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) = \frac{1}{2}\Delta_j(\mathbf{x}^{(t+1)}) + \frac{1}{2}r_j(\mathbf{v}^{(t+1)}) \\
=&\frac{1}{2}\psi_j(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})) + \frac{1}{2}\psi_j(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)})) \\
=&\mathbf{e}_j^\top \left(\boldsymbol{\phi}_{i+1}^\Delta(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)})) + \boldsymbol{\phi}_{i+1}^r(\widehat{\texttt{AcBSLS}}_{i+1}(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)}))\right) \\
\leq&\mathbf{e}_j^\top \left((\mathbf{F}_{i+1} + \mathbf{E}_{i+1})\boldsymbol{\phi}_{i+1}^\Delta(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{x}}^{(t)}) + (\mathbf{F}_{i+1} + \mathbf{E}_{i+1})\boldsymbol{\phi}_{i+1}^r(\tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)})\right) + 8\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{e}_j^\top \mathbf{Z}_{i+1}\mathbf{1} \\
\leq&\mathbf{e}_j^\top (\mathbf{F}_{i+1} + \mathbf{E}_{i+1}) \begin{bmatrix} 2\kappa_{\text{glob}}\mathbf{I}_i & & \\ & 2 & \\ & & 2\kappa_{\text{glob}}\mathbf{I}_{m-i-1} \end{bmatrix} \boldsymbol{\phi}_i^\Delta(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) + 8\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{e}_j^\top \mathbf{Z}_{i+1}\mathbf{1} \\
\leq&2\kappa_{\text{glob}}\mathbf{e}_j^\top (\mathbf{F}_{i+1} + \mathbf{E}_{i+1})\boldsymbol{\phi}_i^\Delta(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) + 8\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{e}_j^\top \mathbf{Z}_{i+1}\mathbf{1}.
\end{aligned}$$

In matrix form we obtain

$$\boldsymbol{\phi}_i^\Delta(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) \leq \mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\boldsymbol{\phi}_i^\Delta(\tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{v}}^{(t)}) + 8\delta L_m \|\mathbf{x}^\star\|_2^2 \mathbf{Z}_{i+1}\mathbf{1}.$$

Hence

$$\boldsymbol{\phi}_i^\Delta(\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}) \leq \mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i}\boldsymbol{\phi}_i^\Delta(\mathbf{x}^{(t)}, \mathbf{v}^{(t)}) + 4\delta L_m \|\mathbf{x}^\star\|_2^2 (\mathbf{K}_i\widehat{\mathbf{F}_{i+1}} + 2\mathbf{Z}_{i+1})\mathbf{1}.$$

Telescoping

$$\boldsymbol{\phi}_i^\Delta(\mathbf{x}^{(T_i)}, \mathbf{v}^{(T_i)}) \leq (\mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i})^{T_i}\boldsymbol{\phi}_i^\Delta(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}) + 4\delta L_m \|\mathbf{x}^\star\|_2^2 \cdot \sum_{t_i=0}^{T_i-1} (\mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i})^{t_i}(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}} + 2\mathbf{Z}_{i+1})\mathbf{1}.$$

■

Next, we estimate the upper bounds of $\mathbf{F}_i$ (in Lemma 86), $\mathbf{E}_i$ (in Lemma 88) and $\mathbf{Z}_i$'s (in Lemma 89).

H.3.1. UPPER BOUND OF $\mathbf{F}$

We first bound $\|\mathbf{F}_i\|_1$ with the following lemma.

Lemma 86 (Upper bound of $\|\mathbf{F}_i\|_1$) *Using the same notation as in Lemma 85, and in addition assume $T_1, \ldots, T_m$ satisfies (C.1), then the following statements hold,*

(a) For any $i \in [m+1]$, $\mathbf{F}_i$ is a diagonal matrix of the form $\prod_{j=i}^m \left((\mathbf{K}_j\mathbf{D}_j)^{\prod_{k=i}^j T_k}\right)$.

(b) For any $i \in [m]$, $\|\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i\|_1 \leq 1$.

(c) For any $i \in [m+1]$, $\|\mathbf{F}_i\|_1 \leq 1$.

(d) $\|\mathbf{F}_1\|_1 \leq \frac{\epsilon}{\psi(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})}$.

Proof [Proof of Lemma 86]

(a) The first statement (a) follows immediately by definition of $\mathbf{F}_i$'s. We prove by induction in reverse order from $m+1$ doewn to 1. For $i = m+1$ we have $\mathbf{F}_{m+1} = \mathbf{I}$ by definition, which is consistent. Now assume the statement holds for the case of $i+1$, then the case of i also holds in that

$$\mathbf{F}_i = (\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i)^{T_i} = \left(\mathbf{K}_i \prod_{j=i+1}^{m} \left((\mathbf{K}_j \mathbf{D}_j)^{\prod_{k=i+1}^{j} T_k} \right) \mathbf{D}_i \right)^{T_i} = \prod_{j=i}^{m} \left((\mathbf{K}_j \mathbf{D}_j)^{\prod_{k=i}^{j} T_k} \right),$$

where the last equality holds because of the commutability among diagonal matrices.

(b) By (a) we have

$$\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i = \mathbf{K}_i \prod_{j=i+1}^{m} \left((\mathbf{K}_j \mathbf{D}_j)^{\prod_{k=i+1}^{j} T_k} \right) \mathbf{D}_i = \prod_{j=i}^{m} \left((\mathbf{K}_j \mathbf{D}_j)^{\prod_{k=i+1}^{j} T_k} \right),$$

By definition of $\mathbf{D}_i$ we can write the l-th diagonal element of $\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i$ as

$$(\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i)_{ll} = \begin{cases} 1 & l < i, \\ (1 - \kappa_l^{-\frac{1}{2}})^{\prod_{k=i+1}^{l} T_k} \cdot (4\kappa_{\mathrm{glob}}^4)^{\sum_{j=i}^{l-1} \prod_{k=i+1}^{j} T_k} & l \geq i. \end{cases}$$

Note that

$$(1 - \kappa_l^{-\frac{1}{2}})^{\prod_{k=i+1}^{l} T_k} \cdot (4\kappa_{\mathrm{glob}}^4)^{\sum_{j=i}^{l-1} \prod_{k=i+1}^{j} T_k} \leq \exp\left(\underbrace{-\kappa_l^{-\frac{1}{2}} \prod_{k=i+1}^{l} T_k + \log(4\kappa_{\mathrm{glob}}^4) \cdot \sum_{j=i}^{l-1} \prod_{k=i+1}^{j} T_k}_{\text{denoted as } \gamma_l} \right).$$

Observe that $\gamma_i = -\kappa_i^{-1} < 0$. For $l \geq i$, it is the case that

$$\begin{aligned}
&\gamma_{l+1} - \gamma_l \\
&= -\kappa_{l+1}^{-\frac{1}{2}} \prod_{k=i+1}^{l+1} T_k + \log(4\kappa_{\mathrm{glob}}^4) \cdot \sum_{j=i}^{l} \prod_{k=i+1}^{j} T_k + \kappa_l^{-\frac{1}{2}} \prod_{k=i+1}^{l} T_k - \log(4\kappa_{\mathrm{glob}}^4) \cdot \sum_{j=i}^{l-1} \prod_{k=i+1}^{j} T_k \\
&= -\kappa_{l+1}^{-\frac{1}{2}} \prod_{k=i+1}^{l+1} T_k + \kappa_l^{-\frac{1}{2}} \prod_{k=i+1}^{l} T_k + \log(4\kappa_{\mathrm{glob}}^4) \cdot \prod_{k=i+1}^{l} T_k \\
&= \prod_{k=i+1}^{l} T_k \left(-\kappa_{l+1}^{-\frac{1}{2}} T_{l+1} + \kappa_l^{-\frac{1}{2}} + \log(4\kappa_{\mathrm{glob}}^4) \right) \leq 0 \\
&\qquad\qquad \text{(since } T_{l+1} \geq \kappa_{l+1}^{\frac{1}{2}}(1 + \log(4\kappa_{\mathrm{glob}}^4)) \text{ by (C.1))}
\end{aligned}$$

Hence $\gamma_m \leq \gamma_{m-1} \leq \cdots \leq \gamma_i \leq 0$. Consequently we have $(\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i)_{ll} \leq 1$ for all l, and thus $\|\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i\|_1 \leq 1$.

(c) By (b), $\|\mathbf{F}_i\|_1 = \|(\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i)^{T_i}\|_1 \leq \|(\mathbf{K}_i \mathbf{F}_{i+1} \mathbf{D}_i)\|_1^{T_i} \leq 1$.

(d) Follows by the same argument as in the exact arithmetic proof Theorem 15, which we sketch here for completeness. By (a),

$$(\mathbf{F}_1)_{ll} = \left[\prod_{j=1}^{m}\left((\mathbf{K}_j\mathbf{D}_j)^{\prod_{k=1}^{j}T_k}\right)\right]_{ll} = (1-\kappa_l^{-\frac{1}{2}})^{\prod_{k=1}^{l}T_k}\cdot(4\kappa_{\text{glob}}^4)^{\sum_{j=i}^{l-1}\prod_{k=1}^{j}T_k}$$

$$\leq \exp\left(\underbrace{-\kappa_l^{-\frac{1}{2}}\prod_{k=1}^{l}T_k + \log(4\kappa_{\text{glob}}^4)\cdot\sum_{j=1}^{l-1}\prod_{k=1}^{j}T_k}_{\text{denoted as }\gamma_l}\right).$$

Since $T_1 \geq \kappa_1^{\frac{1}{2}}\log(\frac{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}{\epsilon})$, we have $\gamma_1 = -\kappa_1^{-\frac{1}{2}}T_1 \leq -\log(\frac{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}{\epsilon})$. Following the same argument as in (b), we have $\gamma_m \leq \gamma_{m-1} \leq \cdots \leq \gamma_1$. Consequently, for any $l \in [m]$

$$(\mathbf{F}_1)_{ll} \leq \exp\left(-\log(\frac{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}{\epsilon})\right) \leq \frac{\epsilon}{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})},$$

which implies $\|\mathbf{F}_1\|_1 \leq \frac{\epsilon}{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}$ since $\mathbf{F}_1$ is diagonal (by (a)).

■

H.3.2. UPPER BOUND OF $\mathbf{E}$

Before we state the upper bound of $\mathbf{E}_i$'s, we first establish the following Lemma 87, which is essential towards the bound for $\mathbf{E}_i$'s and $\mathbf{Z}_i$'s.

Lemma 87 *Using the same notation of Lemma 85, and assume $T_1,\ldots,T_m$ satisfies (C.1), and assume $\delta \leq \frac{1}{20m\kappa_{\text{glob}}^2}$, then for any $t \geq 0$, for any $i \in [m]$, the following inequality holds*

$$\left\|(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i})^t - (\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i)^t\right\|_1 \leq \begin{cases}\varphi(10t\delta m\kappa_{\text{glob}}^2) & i=m\\ \varphi(40t\delta m\kappa_{\text{glob}}^6 + 12t\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1) & i<m,\end{cases}$$

where $\varphi(x) \coloneqq xe^x$.

Proof [Proof of Lemma 87] Let $\mathbf{\Xi}_i \coloneqq \mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i} - \mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i$ (which is non-negative), then

$$\left\|(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i})^t - (\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i)^t\right\|_1 = \left\|(\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i + \mathbf{\Xi}_i)^t - (\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i)^t\right\|_1$$

$$\leq \sum_{s=1}^{t}\binom{t}{s}\|\mathbf{\Xi}_i\|^s\,\|\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i\|_1^{t-s} \leq \sum_{s=1}^{t}\binom{t}{s}\|\mathbf{\Xi}_i\|_1^s \qquad \text{(since } \|\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i\|_1 \leq 1\text{)}$$

$$\leq \|\mathbf{\Xi}_i\|_1 t\sum_{s=0}^{t-1}\binom{t-1}{s}\|\mathbf{\Xi}_i\|_1^s = \|\mathbf{\Xi}_i\|_1 t\,(1+\|\mathbf{\Xi}_i\|_1)^{t-1} \qquad \text{(by Theorem 100)}$$

$$\leq \|\mathbf{\Xi}_i\|_1 t\exp(\|\mathbf{\Xi}_i\|_1 t). \qquad \text{(H.9)}$$

It remains to bound $\|\mathbf{\Xi}_i\|_1$. For $i = m$ we have $\mathbf{E}_{m+1} = 0$, $\mathbf{F}_{m+1} = \mathbf{I}$, $\widehat{\mathbf{F}_{m+1}} = \mathbf{I}$, $\mathbf{K}_m = \mathbf{I}$, which implies $\|\mathbf{\Xi}_m\|_1 = \|\widehat{\mathbf{D}_m} - \mathbf{I}\|_1 \le 10\delta m \kappa_{\text{glob}}^2$.

For other $i < m$, first note that (since $\mathbf{F}_i$ is diagonal)

$$\widehat{\mathbf{F}_{i+1}} = \mathbf{F}_{i+1} + \underbrace{\begin{bmatrix} [\mathbf{I}_{i-1}|\mathbf{0}_{(i-1)\times(m-i+1)}]\mathbf{E}_{i+1}\begin{bmatrix}\mathbf{I}_{i-1} & 0 \\ 0 & 2\kappa_{\text{glob}}\cdot\mathbf{I}_{m-i+1}\end{bmatrix} \\ \mathbf{e}_i^\top\mathbf{E}_{i+1}\begin{bmatrix}\kappa_{\text{glob}}\mathbf{I}_i & & \\ & 1 & \\ & & \kappa_{\text{glob}}\mathbf{I}_{m-i-1}\end{bmatrix} \\ [\mathbf{0}_{(m-i)\times i}|\mathbf{I}_{m-i}]\mathbf{E}_{i+1}\end{bmatrix}}_{\text{denoted as } \clubsuit} \tag{H.10}$$

thus

$$\begin{aligned} &\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i} - \mathbf{F}_{i+1}\mathbf{D}_i = (\mathbf{F}_{i+1} + \clubsuit)(\mathbf{I} + 10\delta\kappa_{\text{glob}}^2\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i - \mathbf{F}_{i+1}\mathbf{D}_i \\ =& 10\delta\kappa_{\text{glob}}^2\mathbf{F}_{i+1}\mathbf{1}\mathbf{1}^\top\mathbf{D}_i + \clubsuit(\mathbf{I} + 10\delta\kappa_{\text{glob}}^2\mathbf{1}\mathbf{1}^\top)\mathbf{D}_i \end{aligned}$$

Since $\|\clubsuit\|_1 \le 2\kappa_{\text{glob}}\|\mathbf{E}_{i+1}\|_1$ and $\|\mathbf{D}_i\| \le 2\kappa_{\text{glob}}^3$ we have bound

$$\left\|\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i} - \mathbf{F}_{i+1}\mathbf{D}_i\right\|_1 \le 20\delta m\kappa_{\text{glob}}^5 + 4\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1(1 + 10\delta m\kappa_{\text{glob}}^2).$$

Thus (for $i < m$)

$$\begin{aligned} &\|\mathbf{\Xi}_i\|_1 \le 2\kappa_{\text{glob}}\left(20\delta m\kappa_{\text{glob}}^5 + 4\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1(1 + 10\delta m\kappa_{\text{glob}}^2)\right) \\ \le& 40\delta m\kappa_{\text{glob}}^6 + 8\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1(1 + 10\delta m\kappa_{\text{glob}}^2) \le 40\delta m\kappa_{\text{glob}}^6 + 12\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1. \\ & \text{(since } \delta \le \tfrac{1}{20m\kappa_{\text{glob}}^2}\text{)} \end{aligned}$$

Substituting back to Eq. (H.9) completes the proof. ■

Now we state the bound for $\mathbf{E}_i$'s.

Lemma 88 (Upper bound of $\|\mathbf{E}_i\|_1$) *Using the same notation of Lemma 85, and assume $T_1, \ldots, T_m$ satisfies (C.1), and assume*

$$\delta \le \frac{1}{2\cdot(10\kappa_{\text{glob}}^2)^{2m-1}m\prod_{j=1}^m T_j}, \tag{H.11}$$

then for any $t \ge 0$, for any $i \in [m]$, the following inequality holds then the following inequality holds

$$\|\mathbf{E}_i\|_1 \le 2\cdot(10\kappa_{\text{glob}}^2)^{2(m-i)+1}\delta m\prod_{j=i}^m T_j. \tag{H.12}$$

Proof [Proof of Lemma 88] We prove by induction in reverse order from m down to 1.

For $i = m$, by definition of $\mathbf{E}_m$,

$$\begin{aligned} &\|\mathbf{E}_m\|_1 = \left\|(\mathbf{K}_m\widehat{\mathbf{F}_{m+1}}\widehat{\mathbf{D}_m})^{T_m} - (\mathbf{K}_m\mathbf{F}_{m+1}\mathbf{D}_m)^{T_m}\right\|_1 \\ \le& 10\delta m\kappa_{\text{glob}}^2 T_m\exp(10\delta m\kappa_{\text{glob}}^2 T_m) && \text{(by Lemma 87)} \\ \le& 10\sqrt{e}\delta m\kappa_{\text{glob}}^2 T_m \le 17\delta m\kappa_{\text{glob}}^2 T_m, && \text{(since } \delta \le \tfrac{1}{20m\kappa_{\text{glob}}^2}\text{)} \end{aligned}$$

which satisfies the bound (H.12).

Now suppose the statement holds for the case of $i+1$, we then study the case of i. By definition of $\mathbf{E}_i$,

$$\begin{aligned}\|\mathbf{E}_i\|_1 &= \left\|(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}}\widehat{\mathbf{D}_i})^{T_i} - (\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i)^{T_i}\right\|_1 \\ &\le \left(40\delta m\kappa_{\text{glob}}^6 + 12\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1\right)T_i \cdot \exp\left[\left(40\delta m\kappa_{\text{glob}}^6 + 12\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1\right)T_i\right] \\ &\qquad\qquad \text{(by Lemma 87)}\end{aligned}$$

Observe that

$$\begin{aligned}&\left(40\delta m\kappa_{\text{glob}}^6 + 12\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1\right)T_i \\ \le& 40\delta m\kappa_{\text{glob}}^6 T_i + 24\kappa_{\text{glob}}^4(10\kappa_{\text{glob}}^2)^{2(m-i-1)+1}\delta m\prod_{j=i}^m T_j \qquad \text{(by induction hypothesis (H.12))} \\ \le& 28\kappa_{\text{glob}}^4(10\kappa_{\text{glob}}^2)^{2(m-i-1)+1}\delta m\prod_{j=i}^m T_j \qquad \text{(H.13)}\end{aligned}$$

and thus

$$\begin{aligned}&\exp\left[\left(40\delta m\kappa_{\text{glob}}^6 + 12\kappa_{\text{glob}}^4\|\mathbf{E}_{i+1}\|_1\right)T_i\right] \le \exp\left[28\kappa_{\text{glob}}^4(10\kappa_{\text{glob}}^2)^{2(m-i-1)+1}\delta m\prod_{j=i}^m T_j\right] \\ &\qquad\qquad \text{(by (H.13))} \\ =& \exp\left[0.28(10\kappa_{\text{glob}}^2)^{2m-1}\delta m\prod_{j=1}^m T_j\right] \le \exp(0.14) < 1.16 \qquad \text{(by } \delta \text{ bound (H.11))}\end{aligned}$$

Consequently

$$\begin{aligned}&\|\mathbf{E}_i\|_1 \le 28\kappa_{\text{glob}}^4(10\kappa_{\text{glob}}^2)^{2(m-i)+1}\delta m\prod_{j=i}^m T_j \times 1.16 \\ \le& 45\kappa_{\text{glob}}^4(10\kappa_{\text{glob}}^2)^{2(m-i-1)+1}\delta m\prod_{j=i}^m T_j = 0.45\cdot(10\kappa_{\text{glob}}^2)^{2(m-i)+1}\delta m\prod_{j=i}^m T_j \\ \le& 2\cdot(10\kappa_{\text{glob}}^2)^{2(m-i)+1}\delta m\prod_{j=i}^m T_j.\end{aligned}$$

■

H.3.3. UPPER BOUND OF $\mathbf{Z}$

Lemma 89 (Upper bound of $\|\mathbf{Z}_i\|_1$) *Using the same notation of Lemma 85 and assuming the same assumptions of Lemma 88, then the following inequality holds for any $i \in [m]$,*

$$\|\mathbf{Z}_i\|_1 \le 3^{m-i+2}\kappa_{\text{glob}}\prod_{j=i}^m T_j.$$

Proof [Proof of Lemma 89] Recall the definition of $\mathbf{Z}_i$ from Lemma 85,

$$\mathbf{Z}_i \coloneqq \sum_{t_i=0}^{T_i-1} (\mathbf{K}_i\widehat{\mathbf{F}_{i+1}\mathbf{D}_i})^{t_i} \left(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}} + 2\mathbf{Z}_{i+1}\right)$$

We will bound $\| \sum_{t_i=0}^{T_i-1} (\mathbf{K}_i\widehat{\mathbf{F}_{i+1}\mathbf{D}_i})^{t_i}\|_1$ and $\|\mathbf{K}_i\widehat{\mathbf{F}_{i+1}} + 2\mathbf{Z}_{i+1}\|_1$ separately. The former is bounded as

$$\begin{aligned}
&\left\|\sum_{t_i=0}^{T_i-1} (\mathbf{K}_i\widehat{\mathbf{F}_{i+1}\mathbf{D}_i})^{t_i}\right\|_1 \le \sum_{t_i=0}^{T_i-1} \left\|(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}\mathbf{D}_i})^{t_i}\right\|_1 && \text{(by triangle inequality)}\\
\le &\sum_{t_i=0}^{T_i-1} \left(\|(\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i)^{t_i}\|_1 + \left\|(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}\mathbf{D}_i})^{t_i} - (\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i)^{t_i}\right\|\right) && \text{(by triangle inequality)}\\
\le &T_i + \sum_{t_i=0}^{T_i-1} \left\|(\mathbf{K}_i\widehat{\mathbf{F}_{i+1}\mathbf{D}_i})^{t_i} - (\mathbf{K}_i\mathbf{F}_{i+1}\mathbf{D}_i)^{t_i}\right\| && \text{(since } \|\mathbf{K}_i\widehat{\mathbf{F}_{i+1}\mathbf{D}_i}\|_1 \le 1 \text{ by Lemma 86)}\\
\le &T_i \left(1 + 0.45\cdot(10\kappa_{\text{glob}}^2)^{2(m-i)+1}\delta m \prod_{j=i}^{m} T_j\right) && \text{(by the proof of Lemmas 87 and 88)}\\
\le &1.23T_i && \text{(by } \delta \text{ bound (H.11))}
\end{aligned}$$

The second term is bounded as

$$\begin{aligned}
&\left\|\mathbf{K}_i\widehat{\mathbf{F}_{i+1}} + 2\mathbf{Z}_{i+1}\right\|_1 \le \|\mathbf{K}_i\|_1\cdot\|\widehat{\mathbf{F}_{i+1}}\|_1 + 2\|\mathbf{Z}_{i+1}\|_1\\
\le &2\kappa_{\text{glob}}\cdot(1 + 2\kappa_{\text{glob}}\|\mathbf{E}_{i+1}\|_1) + 2\|\mathbf{Z}_{i+1}\|_1 && \text{(by (H.10))}\\
\le &2.04\kappa_{\text{glob}} + 2\|\mathbf{Z}_{i+1}\|_1 && \text{(since } \|\mathbf{E}_{i+1}\|_1 \le \tfrac{1}{100m\kappa_{\text{glob}}^2} \text{ by Lemma 88)}
\end{aligned}$$

In summary $\|\mathbf{Z}_i\|_1 \le 1.23T_i(2.04\kappa_{\text{glob}} + 2\|\mathbf{Z}_{i+1}\|_1)$. By induction we can show that $\|\mathbf{Z}_i\|_1 \le 3^{m-i+2}\kappa_{\text{glob}}\prod_{j=1}^{m} T_j$. ■

H.4. Finishing the proof of Theorem 79

We are ready to finish the proof of Theorem 79.

Proof [Proof of Theorem 79] By Lemma 85 we have

$$\boldsymbol{\phi}_1^{\Delta}(\widehat{\texttt{AcBSLS}_1}(\mathbf{x}^{(0)}, \mathbf{v}^{(0)})) \le (\mathbf{F}_1 + \mathbf{E}_1)\boldsymbol{\phi}_1^{\Delta}(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{Z}_1\mathbf{1}.$$

By definition of vector potential $\boldsymbol{\phi}_1^{\Delta}$ we have $\|\boldsymbol{\phi}_1^{\Delta}(\mathbf{x}, \mathbf{v})\|_1 = \psi(\mathbf{x}, \mathbf{v})$ for any $\mathbf{x}, \mathbf{v}$. Consequently

$$\begin{aligned}
&\psi(\widehat{\texttt{AcBSLS}_1}(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}))\\
\le &\left\|(\mathbf{F}_1 + \mathbf{E}_1)\boldsymbol{\phi}_1^{\Delta}(\mathbf{x}, \mathbf{v}) + 4\delta L_m\|\mathbf{x}^\star\|_2^2\mathbf{Z}_1\mathbf{1}\right\|_1 && \text{(by Lemma 85)}\\
\le &(\|\mathbf{F}_1\|_1 + \|\mathbf{E}_1\|_1)\psi(\mathbf{x}, \mathbf{v}) + 4\delta m L_m\|\mathbf{x}^\star\|_2^2\|\mathbf{Z}_1\|_1 && \text{(triangle inequality)}
\end{aligned}$$

Plugging in the bound of $\|\mathbf{F}_1\|$ (from Lemma 86), $\|\mathbf{E}_1\|$ (from Lemma 88), and $\|\mathbf{Z}_1\|_1$ (from Lemma 89), we arrive at

$$\begin{aligned}&\psi(\widehat{\texttt{AcBSLS}}_1(\mathbf{x}^{(0)},\mathbf{v}^{(0)}))\\ &\leq\left(\frac{\epsilon}{\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})}+2\cdot(10\kappa_{\mathrm{glob}}^2)^{2(m-i)+1}\delta m\prod_{j=i}^{m}T_j\right)\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})+4\cdot 3^{m+1}\delta m\kappa_{\mathrm{glob}}L_m\|\mathbf{x}^\star\|_2^2\prod_{j=1}^{m}T_j.\end{aligned}$$

By δ bound

$$\delta\leq\left(\prod_{i\in[m]}T_i^{-1}\right)\min\left\{\frac{1}{2\cdot(10\kappa_{\mathrm{glob}}^2)^{2m-1}m},\frac{\epsilon}{2\cdot(10\kappa_{\mathrm{glob}}^2)^{2m-1}\cdot\psi(\mathbf{x}^{(0)},\mathbf{v}^{(0)})},\frac{\epsilon}{4\cdot 3^{m+1}\cdot mL_m\kappa_{\mathrm{glob}}\|\mathbf{x}^\star\|_2^2}\right\},$$

we immediately obtain $\psi(\widehat{\texttt{AcBSLS}}_1(\mathbf{x}^{(0)},\mathbf{v}^{(0)}))\leq 3\epsilon$. ■

Appendix I. Deferred proof of supporting lemmas of Lemma 26

In this appendix section we provide the proof of several supporting lemmas toward Lemma 26.

I.0.1. DEFERRED PROOF OF LEMMA 41

Proof [Proof of Lemma 41] Apply Lemma 39, since for any $k\in[m-1]$,

$$\int_{L_k}^{\mu_{k+1}}\frac{h(\zeta)}{\sqrt{q(\zeta)}}=0,$$

we conclude that $h(z)$ has $m-1$ real roots $r_1,r_2,\ldots,r_{m-1}$ such that $r_k\in[L_k,\mu_{k+1}]$. Therefore

$$g_S(0)=\int_0^{\mu_1}\frac{\prod_{k\in[m-1]}(r_k-\zeta)}{\sqrt{\prod_{k\in[m]}(\mu_k-\zeta)(L_k-\zeta)}}\mathrm{d}\zeta$$

By monotonocity,

$$\begin{aligned}&\int_0^{\mu_1}\frac{\prod_{k\in[m-1]}(r_k-\zeta)}{\sqrt{\prod_{k\in[m]}(\mu_k-\zeta)(L_k-\zeta)}}\mathrm{d}\zeta\leq\frac{\prod_{k\in[m-1]}r_k}{\sqrt{\prod_{k\in[m]}(L_k-\mu_1)\cdot\prod_{k=2}^{m}(\mu_k-\mu_1)}}\cdot\int_0^{\mu_1}\frac{\mathrm{d}\zeta}{\sqrt{\mu_1-\zeta}}\mathrm{d}\zeta\\ =&\frac{\prod_{k\in[m-1]}r_k}{\sqrt{\prod_{k\in[m]}(L_k-\mu_1)\cdot\prod_{k=2}^{m}(\mu_k-\mu_1)}}\cdot 2\sqrt{\mu_1}=\frac{2\prod_{k\in[m-1]}\frac{r_k}{\mu_k}}{\sqrt{\prod_{k\in[m]}(\frac{L_k}{\mu_k}-\frac{\mu_1}{\mu_k})\cdot\prod_{k=2}^{m}(1-\frac{\mu_1}{\mu_k})}}.\end{aligned}$$

By assumption $\frac{L_k}{\mu_k}\geq 2$ we have $\frac{\mu_1}{\mu_k}\leq\frac{1}{2^{k-1}}$, and thus $\frac{L_k}{\mu_k}-\frac{\mu_1}{\mu_k}\geq(1-\frac{1}{2^k})\frac{L_k}{\mu_k}$. Hence

$$\prod_{k\in[m]}\left(\frac{L_k}{\mu_k}-\frac{\mu_1}{\mu_k}\right)\cdot\prod_{k=2}^{m}\left(1-\frac{\mu_1}{\mu_k}\right)\geq\left(\prod_{k\in[m]}\frac{L_k}{\mu_k}\right)\cdot\left(\prod_{k\in[m]}(1-2^{-k})\right)\left(\prod_{k\in[m-1]}(1-2^{-k})\right)$$

Since $\prod_{k=1}^{\infty}(1-2^{-k}) > 0.288$ (see Theorem 101) we arrive at

$$g_S(0) \le \frac{2\prod_{k\in[m-1]}\frac{r_k}{\mu_{k+1}}}{\sqrt{0.288^2\prod_{k\in[m]}\frac{L_k}{\mu_k}}} \le \frac{7\prod_{k\in[m-1]}\frac{r_k}{\mu_{k+1}}}{\sqrt{\prod_{k\in[m]}\frac{L_k}{\mu_k}}}.$$

completing the proof. ■

I.0.2. DEFERRED PROOF OF LEMMA 43

Proof [Proof of Lemma 43] Since $h(z) = \prod_{j=1}^{m-1}(z-r_j)$ satisfies

$$0 = \int_{L_k}^{\mu_{k+1}} \frac{h(\zeta)}{\sqrt{q(\zeta)}}\mathrm{d}\zeta = \int_{L_k}^{\mu_{k+1}} \frac{\prod_{j\in[m]}(\zeta-r_j)}{\sqrt{q(\zeta)}}\mathrm{d}\zeta,$$

we have for any $k \in [m-1]$

$$\int_{L_k}^{\mu_{k+1}} \frac{r_k\prod_{j\neq k}(\zeta-r_j)}{\sqrt{q(\zeta)}}\mathrm{d}\zeta = \int_{L_k}^{\mu_{k+1}} \frac{\zeta\prod_{j\neq k}(\zeta-r_j)}{\sqrt{q(\zeta)}}\mathrm{d}\zeta.$$

Thus

$$r_k = \frac{\int_{L_k}^{\mu_{k+1}} \frac{\zeta\prod_{j\in[k-1]}(r_j-\zeta)\cdot\prod_{j=k+1}^{m-1}(\zeta-r_j)}{\sqrt{\prod_{j\in[m]}(\zeta-\mu_j)\prod_{j\in[m]}(\zeta-L_j)}}\mathrm{d}\zeta}{\int_{L_k}^{\mu_{k+1}} \frac{\prod_{j\in[k-1]}(r_j-\zeta)\cdot\prod_{j=k+1}^{m-1}(\zeta-r_j)}{\sqrt{\prod_{j\in[m]}(\zeta-\mu_j)\prod_{j\in[m]}(\zeta-L_j)}}\mathrm{d}\zeta}.$$

Rearranging

$$r_k = \frac{\int_{L_k}^{\mu_{k+1}} \frac{\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}\left(\prod_{j\in[k-1]}\frac{\zeta-r_j}{\sqrt{(\zeta-\mu_j)(\zeta-L_j)}}\right)\left(\prod_{j=k+1}^{m-1}\frac{r_j-\zeta}{\sqrt{(\mu_{j+1}-\zeta)(L_{j+1}-\zeta)}}\right)\mathrm{d}\zeta}{\int_{L_k}^{\mu_{k+1}} \frac{1}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}\left(\prod_{j\in[k-1]}\frac{\zeta-r_j}{\sqrt{(\zeta-\mu_j)(\zeta-L_j)}}\right)\left(\prod_{j=k+1}^{m-1}\frac{r_j-\zeta}{\sqrt{(\mu_{j+1}-\zeta)(L_{j+1}-\zeta)}}\right)\mathrm{d}\zeta}.$$

Observe that

- For any $j < k$, $\frac{\zeta-r_j}{\sqrt{(\zeta-\mu_j)(\zeta-L_j)}}$ is non-negative and monotonically **increasing** in $\zeta \in [L_k, \mu_{k+1}]$ since $\mu_j < L_j < r_j$.
- For any $j > k$, $\frac{r_j-\zeta}{\sqrt{(\zeta-\mu_{j+1})(\zeta-L_{j+1})}}$ is non-negative and monotonically **decreasing** in $\zeta \in [L_k, \mu_{k+1}]$ since $r_j < \mu_{j+1} < L_{j+1}$.

Consequently

$$r_k \le \frac{\int_{L_k}^{\mu_{k+1}} \frac{\zeta\mathrm{d}\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}}{\int_{L_k}^{\mu_{k+1}} \frac{\mathrm{d}\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}} \cdot \prod_{j\in[k-1]}\left(\underbrace{\frac{\frac{\mu_{k+1}-r_j}{\sqrt{(\mu_{k+1}-\mu_j)(\mu_{k+1}-L_j)}}}{\frac{L_k-r_j}{\sqrt{(L_k-\mu_j)(L_k-L_j)}}}}_{\text{denoted as } \gamma_j}\right) \tag{I.1}$$

Note that

$$\gamma_j := \frac{\frac{\mu_{k+1}-r_j}{\sqrt{(\mu_{k+1}-\mu_j)(\mu_{k+1}-L_j)}}}{\frac{L_k-r_j}{\sqrt{(L_k-\mu_j)(L_k-L_j)}}} = \frac{1-\frac{r_j}{\mu_{k+1}}}{1-\frac{r_j}{L_k}} \cdot \frac{\sqrt{(1-\frac{\mu_j}{L_k})(1-\frac{L_j}{L_k})}}{\sqrt{(1-\frac{\mu_j}{\mu_{k+1}})(1-\frac{L_j}{\mu_{k+1}})}} \le \frac{1}{1-\frac{r_j}{L_k}}.$$

and by $\min_{j\in[m]} \frac{L_j}{\mu_j} \ge 2$ one has

$$\frac{r_j}{L_k} \le \frac{\mu_{j+1}}{L_k} \le 2^{-(k-j)}.$$

We arrive at (by Theorem 101, $\prod_{j=1}^{\infty} 1-2^{-i} \ge 0.288$)

$$\prod_{j\in[k-1]} \gamma_j \le \prod_{j\in[k-1]} \frac{1}{1-2^{-(k-j)}} \le \frac{1}{0.288} \le 4.$$

Substitute back to Eq. (I.1),

$$r_k \le 4 \cdot \frac{\int_{L_k}^{\mu_{k+1}} \frac{\zeta \mathrm{d}\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}}{\int_{L_k}^{\mu_{k+1}} \frac{\mathrm{d}\zeta}{\sqrt{(\zeta-\mu_k)(\zeta-L_k)(\mu_{k+1}-\zeta)(L_{k+1}-\zeta)}}}.$$

■

I.0.3. DEFERRED PROOF OF LEMMA 44

We will prove Lemma 44 by analyzing the integrals in the numerator and denominator. In particular, we will apply the tools from elliptic integral theory. We adopt the following Legendre forms of elliptic integrals, defined as follows:

- $\mathcal{F}(\phi|p) := \int_0^\phi \frac{\mathrm{d}\theta}{\sqrt{1-p\sin^2\theta}}$ denotes the (incomplete) elliptic integral of the **first** kind.
- $\mathcal{K}(p) := \mathcal{F}(\frac{\pi}{2}|p)$ denotes the complete elliptic integral of the **first** kind.
- $\mathcal{E}(\phi|p) := \int_0^\phi \sqrt{1-p\sin^2\theta}\mathrm{d}\theta$ denotes the (incomplete) elliptic integral of the **second** kind.
- $\mathcal{E}(p) := \mathcal{E}(\frac{\pi}{2}|p)$ denotes the complete elliptic integral of the **second** kind.
- $\Pi(n;\phi|p) := \int_0^\phi \frac{\mathrm{d}\theta}{(1-n\sin^2\theta)\sqrt{(1-p\sin^2\theta)}}$ denotes the incomplete elliptic integral of the **third** kind.
- $\Pi(n|p) := \Pi(n;\frac{\pi}{2}|p)$ denotes the complete elliptic integral of the **third** kind.

To simplify the notation, throughout this subsubsection we assume without loss of generality that $k=1$. The result apparently holds for any $k \in [m-1]$.

Denote (throughout this subsubsection) that

$$\varphi(z) := (z-\mu_1)(z-L_1)(\mu_2-z)(L_2-z).$$

The following Lemma 90 analyzes $\int_{L_1}^{\mu_2} \frac{\mathrm{d}\zeta}{\sqrt{\varphi(\zeta)}}$.

Lemma 90 *For any $0 < \mu_1 < L_1 < \mu_2 < L_2$, the following equality holds*

$$\int_{L_1}^{\mu_2} \frac{\mathrm{d}\zeta}{\sqrt{\varphi(\zeta)}} = \frac{2\mathcal{K}\left(\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_2-L_1)(\mu_2-\mu_1)}\right)}{\sqrt{(L_2-L_1)(\mu_2-\mu_1)}}.$$

Proof [Proof of Lemma 90] The antiderivative of $\frac{1}{\sqrt{\varphi(\zeta)}}$ in $[L_1, \mu_2]$ is

$$Q(\zeta) := -2\sqrt{\frac{-1}{(L_1-\mu_1)(L_2-\mu_2)}}\mathcal{F}\left(\mathrm{Arcsin}\left(\sqrt{-\frac{(\zeta-L_1)(L_2-\mu_2)}{(L_2-L_1)(\mu_2-\zeta)}}\right)\middle|\frac{(L_2-L_1)(\mu_2-\mu_1)}{(L_1-\mu_1)(L_2-\mu_2)}\right).$$

The lemma then follows by the fact that $\lim_{\zeta\to L_1} Q(\zeta) = 0$ and

$$\lim_{\zeta\to\mu_2} Q(\zeta) = 2\sqrt{\frac{1}{(L_1-\mu_1)(L_2-\mu_2)}}\mathcal{K}\left(-\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_1-\mu_1)(L_2-\mu_2)}\right) = \frac{2\mathcal{K}\left(\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_2-L_1)(\mu_2-\mu_1)}\right)}{\sqrt{(L_2-L_1)(\mu_2-\mu_1)}}.$$

■

The following Lemma 91 analyzes $\int_{L_1}^{\mu_2} \frac{\zeta\mathrm{d}\zeta}{\sqrt{\varphi(\zeta)}}$.

Lemma 91 *For any $0 < \mu_1 < L_1 < \mu_2 < L_2$, the following equality holds*

$$\int_{L_1}^{\mu_2} \frac{\zeta\mathrm{d}\zeta}{\sqrt{\varphi(\zeta)}} = 2\sqrt{\frac{1}{(\mu_2-\mu_1)(L_2-L_1)}}\cdot \left(L_2\mathcal{K}\left(\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_2-L_1)(\mu_2-\mu_1)}\right) - (L_2-\mu_2)\Pi\left(\frac{\mu_2-L_1}{L_2-L_1}\middle|\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_2-L_1)(\mu_2-\mu_1)}\right)\right).$$

Proof [Proof of Lemma 91] The antiderivative of $\frac{\zeta}{\sqrt{\varphi(\zeta)}}$ in $[L_1, \mu_2]$ is

$$\begin{aligned} P(\zeta) :=& 2\sqrt{\frac{1}{(\mu_2-\mu_1)(L_2-L_1)}}\cdot \\ & \left(iL_1\mathcal{F}\left(\mathrm{Arcsin}\left(\sqrt{-\frac{(L_2-L_1)(\mu_2-\zeta)}{(\zeta-L_1)(L_2-\mu_2)}}\right)\middle|\frac{(L_1-\mu_1)(L_2-\mu_2)}{(L_2-L_1)(\mu_2-\mu_1)}\right)\right. \\ & \left.+i(\mu_2-L_1)\Pi\left(\frac{L_2-\mu_2}{L_2-L_1};\mathrm{Arcsin}\left(\sqrt{-\frac{(L_2-L_1)(\mu_2-\zeta)}{(\zeta-L_1)(L_2-\mu_2)}}\right)\middle|\frac{(L_1-\mu_1)(L_2-\mu_2)}{(L_2-L_1)(\mu_2-\mu_1)}\right)\right). \end{aligned}$$

The lemma then follows by the fact that $\lim_{\zeta\to\mu_2} P(\zeta) = 0$ and

$$\begin{aligned} \lim_{\zeta\to L_1} P(\zeta) =& 2\sqrt{\frac{1}{(\mu_2-\mu_1)(L_2-L_1)}}\cdot \\ & \left(-L_2\mathcal{K}\left(\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_2-L_1)(\mu_2-\mu_1)}\right) + (L_2-\mu_2)\Pi\left(\frac{\mu_2-L_1}{L_2-L_1}\middle|\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_2-L_1)(\mu_2-\mu_1)}\right)\right). \end{aligned}$$

■

Next, we establish the following inequality regarding elliptic integrals.

Lemma 92 *For any $x \in [0, \frac{1}{2}]$ and $y \in (0, 1]$, the following inequality holds*

$$\frac{\Pi(x|1-y)}{\mathcal{K}(1-y)} \geq \frac{1}{1-x}\left(1 - \frac{4x}{\log(\frac{16}{y})}\right).$$

Proof [Proof of Theorem 92] We first prove two claims:

Claim 93 *The function $(1-x)\Pi(x|1-y)$ is concave in x for $x \in [0, 1)$ and $y \in (0, 1]$.*

Proof [Proof of Claim 93] By standard convex analysis we can show that $\frac{1-x}{1-x\sin^2\theta}$ is concave in x for any $x \in [0, 1)$ and $\theta \in [0, \frac{\pi}{2}]$. Thus by linearity of integrals we have that $(1-x)\Pi(x|1-y) = \int_0^{\frac{\pi}{2}} \frac{(1-x)\mathrm{d}\theta}{(1-x\sin^2\theta)\sqrt{1-(1-y)\sin^2\theta}}$ is also concave in $x \in [0, 1)$ provided that $y \in (0, 1]$. ■

Claim 94 *Let $\varphi(x, y) := (1-x) \cdot \frac{\Pi(x|1-y)}{\mathcal{K}(1-y)}$. Then for any $y \in (0, 1]$, the following inequality holds*

$$\frac{\partial\varphi(\frac{1}{2}, y)}{\partial x} \geq \frac{-4}{\log\frac{16}{y}}.$$

Proof [Proof of Claim 94] By standard elliptic integral analysis

$$\frac{\partial\varphi(\frac{1}{2}, y)}{\partial x} = -1 + \frac{2\mathcal{E}(1-y) - \Pi(\frac{1}{2}, 1-y)}{(2y-1)\mathcal{K}(1-y)}.$$

Expanding the above quantity around $y = 0$ shows the lower bound $\frac{-4}{\log\frac{16}{y}}$. ■

The proof of Theorem 92 then follows by the above two claims. Since $(1-x)\Pi(x|1-y)$ is concave in $x \in [0, \frac{1}{2}]$, so is $\varphi(x, y)$ defined in Claim 94. Thus for any $x \in [0, \frac{1}{2}]$,

$$\varphi(x, y) \geq \varphi(0, y) + x\frac{\partial\varphi(\frac{1}{2}, y)}{\partial x} \geq 1 - \frac{4x}{\log\frac{16}{y}}.$$

Hence

$$\frac{\Pi(x|1-y)}{\mathcal{K}(1-y)} = \frac{\varphi(x, y)}{1-x} \geq \frac{1}{1-x}\left(1 - \frac{4x}{\log(\frac{16}{y})}\right).$$

■

Finally, the proof of Lemma 44 then follows by applying Lemmas 90, 91 and 92.

Proof [Proof of Lemma 44] By Lemmas 90, 91 and 92,

$$\left(\int_{L_1}^{\mu_2} \frac{\mathrm{d}\zeta}{\sqrt{\varphi(\zeta)}}\right)^{-1}\left(\int_{L_1}^{\mu_2} \frac{\zeta\mathrm{d}\zeta}{\sqrt{\varphi(\zeta)}}\right) = L_2 - (L_2-\mu_2)\frac{\Pi\left(\frac{\mu_2-L_1}{L_2-L_1}\middle|\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_2-L_1)(\mu_2-\mu_1)}\right)}{\mathcal{K}\left(\frac{(L_2-\mu_1)(\mu_2-L_1)}{(L_2-L_1)(\mu_2-\mu_1)}\right)}$$

$$\leq L_2 - (L_2-\mu_2)\frac{1}{1-\frac{\mu_2-L_1}{L_2-L_1}}\left(1 - \frac{4\frac{\mu_2-L_1}{L_2-L_1}}{\log\left(16\frac{(\mu_2-\mu_1)(L_2-L_1)}{(L_1-\mu_1)(L_2-\mu_2)}\right)}\right)$$

(by Theorem 92 and assumption $\frac{\mu_2}{L_2} \leq \frac{1}{2}$)

$$= L_1 + \frac{4(\mu_2-L_1)}{\log\left(16\frac{(\mu_2-\mu_1)(L_2-L_1)}{(L_1-\mu_1)(L_2-\mu_2)}\right)} \leq L_1 + \frac{4\mu_2}{\log\left(16\frac{(\mu_2-\mu_1)(L_2-L_1)}{(L_1-\mu_1)(L_2-\mu_2)}\right)}.$$

Since

$$\frac{(\mu_2-\mu_1)(L_2-L_1)}{(L_1-\mu_1)(L_2-\mu_2)} = \left(1+\frac{\frac{\mu_2}{L_1}-1}{1-\frac{\mu_1}{L_1}}\right)\left(1+\frac{\mu_2-L_1}{L_2-\mu_2}\right) \geq 1+\frac{\mu_2}{L_1}-1 = \frac{\mu_2}{L_1},$$

and

$$\frac{L_1}{\mu_2} \leq \frac{3}{\log(16\frac{\mu_2}{L_1})},$$

we obtain

$$\left(\int_{L_1}^{\mu_2}\frac{\mathrm{d}\zeta}{\sqrt{\varphi(\zeta)}}\right)^{-1}\left(\int_{L_1}^{\mu_2}\frac{\zeta\mathrm{d}\zeta}{\sqrt{\varphi(\zeta)}}\right) \leq L_1+\frac{4\mu_2}{\log(16\frac{\mu_2}{L_1})} \leq \frac{7\mu_2}{\log(16\frac{\mu_2}{L_1})}.$$

■

Appendix J. Technical details for Appendix E

Remark 95 *We can obtain an example of a non-product, non-Gaussian distribution that satisfies the second-order independence condition by considering any collection of pairwise independent random variables that take values in $\{\pm 1\}^d$ and have expected value 0. As an example: let $\mathbf{P}$ be the $d \times d$ identity matrix. Let $\mathbf{a} \in \{\pm 1\}^d$ be uniform on the subset of the hypercube $\{\pm 1\}^d$ with even parity: $\Pi_{i=1}^d \mathbf{a}_i = 1$. Now, note that the distribution of $\mathbf{a}$ does not have a product structure but the second-order independence condition is satisfied: $\mathbb{E}[(\mathbf{Pa})_i(\mathbf{Pa})_k^2(\mathbf{Pa})_j] = 0$ since each pair of coordinates of $\mathbf{a}$ is pairwise independent (for $d > 2$). Note that we could also obtain a candidate distribution which satisfies the property if $\mathbf{P}$ is not the identity: We can choose the distribution over $\mathbf{a}$ such that $\mathbf{Pa}$ is still uniform on the subset of the hypercube with even parity as before. Another different example is any distribution over $\left\{\mathbf{a} \in \mathbb{R}^d \text{ s.t. } \|\mathbf{Pa}\|_0 \leq 1\right\}$.*

J.1. Missing details for proof of Lemma 53 Eq. (J.1)

We want to show that if

$$\delta \leq \min\left\{\frac{1}{3m}\frac{1}{T_{\max}(N_1+2)^2+1}, \right.$$
$$\frac{1}{6m(4^m)}\left(48mT_{\max}^2(N_1+2)^3\right)^{-1/2},$$
$$\left.\frac{1}{6m(4^m)}\left(\max_{\ell\in[m]}\{\kappa_\ell\}\cdot mT_{\max}^2(N_1+2)^3\right)^{-1/2}\right\},$$

then

$$4^m(1+3\delta m)^{m\left(T_{\max}(N_1+2)^2+1\right)} \leq \min\left\{\frac{\kappa_{\mathrm{glob}}}{(1+3\delta m)^{N_1}}, \frac{1}{144N_1T_{\max}\delta m}, \frac{1}{2\max_{\ell\in[m]}\{\kappa_\ell\}}\frac{1}{6(N_1+1)\delta m}\right\}. \tag{J.1}$$

Using that for any $x \geq 0$, $(1+\frac{1}{x})^x \leq e$ and that $\kappa_{\mathrm{glob}} \geq (4e)^m$ we have,

$$4^m(1+3\delta m)^{m\left(T_{\max}(N_1+2)^2+1\right)} \leq (4e)^m \leq \kappa_{\mathrm{glob}},$$

therefore

$$4^m(1+3\delta m)^{m\left(T_{\max}(N_1+2)^2+1\right)} \leq \frac{\kappa_{\text{glob}}}{(1+3\delta m)^{N_1}}.$$

Next suppose for some $x, p, C \in \mathbb{R}$ we have $px \leq 1/4$, $x \leq \sqrt{C/4p}$, and $C, p \geq 1$. Then

$$(1+x)^p x \leq (1+2px)x \leq \sqrt{\frac{C}{4p}} + \frac{C}{4} \leq C.$$

Then letting $x = 3\delta m$, $p = m(T_{\max}(N_1+2)^2+1)$, and $C = (144 N_1 T_{\max} 4^m)^{-1}$ we have that when $\delta \leq \frac{1}{6m(4^m)} \left(48 m T_{\max}^2 (N_1+2)^3\right)^{-1/2}$ then

$$4^m(1+3\delta m)^{m\left(T_{\max}(N_1+2)^2+1\right)} \leq \frac{1}{144 N_1 T_{\max} \delta m}.$$

Similarly, letting $x = 3\delta m$, $p = m(T_{\max}(N_1+2)^2+1)$, and $C = \left(2\max_{\ell\in[m]}\{\kappa_\ell\}\, 6(N_1+1)4^m\right)^{-1}$ we have that when $\delta \leq \frac{1}{6m(4^m)} \left(\max_{\ell\in[m]}\{\kappa_\ell\} \cdot m T_{\max}^2 (N_1+2)^3\right)^{-1/2}$ then

$$4^m(1+3\delta m)^{m\left(T_{\max}(N_1+2)^2+1\right)} \leq \frac{1}{2\max_{\ell\in[m]}\{\kappa_\ell\}} \frac{1}{6(N_1+1)\delta m}.$$

Therefore Eq. J.1 holds.

J.2. Proof of Lemma 52

Proof [Proof of Lemma 52] We prove Lemma 52 by induction on i. We start with the base case of $\texttt{StochBSLSRes}_m$ and then we will show that if Lemma 52 is true for $\texttt{StochBSLSRes}_{i+1}$ then it is true for $\texttt{StochBSLSRes}_i$. Let $\mathbf{u}^{(0)}$ denote our input vector $\mathbf{u}$ to $\texttt{StochBSLSRes}_i$ and $\mathbf{u}^{(t)}$ denotes the t^{th} iteration of $\texttt{StochBSLSRes}_i$. Before beginning the proof by induction we establish several useful claims that will be used throughout the proof. First note that if

$$n_{\text{avg}} \geq \mathsf{Kurt}(\mathcal{D}) m^2 n_{\max} \left(\prod_{i=1}^m \kappa_i\right) \log^m(\kappa_{\text{glob}}),$$

then since $\delta \coloneqq \mathsf{Kurt}(\mathcal{D}) n_{\max}/n_{\text{avg}}$ we have

$$\delta \leq \left(m^2 \left(\prod_{i=1}^m \kappa_i\right) \log^m(\kappa_{\text{glob}})\right)^{-1}. \tag{J.2}$$

Note that if $\mathbf{u}_i^{(t+1)} = \mathbf{U}_i \mathbf{u}_i^{(t)}$ we have

$$\mathbf{u}_i^{(t+1)} \leq \gamma_i \mathbf{u}_i^{(t)} + \left(\delta \sum_{k=i+1}^m \frac{L_k}{L_i} \mathbf{u}_k^{(t)}\right) + \left(\delta \sum_{k=1}^{i-1} \frac{L_k}{L_i} \mathbf{u}_k^{(t)}\right), \tag{J.3}$$

$$\mathbf{u}_j^{(t+1)} \leq \left(\frac{L_j}{L_i} - 1\right)^2 \mathbf{u}_j^{(t)} + \delta \sum_{k=1}^m \frac{L_k L_j}{L_i^2} \mathbf{u}_k^{(t)}, \qquad \text{(For all } j \geq i+1\text{)}$$

$$\mathbf{u}_j^{(t+1)} \leq \left(1 - \frac{\mu_j}{C_i L_i}\right)^2 \mathbf{u}_j^{(t)} + \delta \sum_{k=1}^m \frac{L_k L_j}{L_i^2} \mathbf{u}_k^{(t)}. \qquad \text{(For all } j \leq i-1\text{)}$$

Finally,

$$T_i \geq \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right). \tag{J.4}$$

To see this note,

$$\begin{aligned}
\log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right) &= \frac{\log\left(\frac{1}{\rho^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)}{\log(\tilde{\gamma}_i)} \\
&= \frac{\log\left(\rho^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)\frac{L_i}{L_{i-1}}\frac{\mathbf{u}_i^{(0)}}{\mathbf{u}_{i-1}^{(0)}}\right)}{\log(1/\tilde{\gamma}_i)} \\
&\leq 2\kappa_i \log\left(\rho^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)\frac{L_i}{L_{i-1}}\frac{\mathbf{u}_i^{(0)}}{\mathbf{u}_{i-1}^{(0)}}\right) \\
&\qquad (\tilde{\gamma}_i = 1 - \tfrac{1}{2\kappa_i} \text{ and } \log(1/(1-x)) \geq x) \\
&\leq 2\kappa_i \log\left(\rho^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)\beta_0\frac{L_i^2}{L_{i-1}^2}\right) \\
&\qquad (\mathbf{u}_i^{(0)}/\mathbf{u}_{i-1}^{(0)} \leq \beta_0 L_i/L_{i-1}) \\
&\leq 6\kappa_i \log\left(\kappa_{\mathrm{glob}}\right) \\
&\qquad (\beta_0\rho^{N_{i+1}}\left(\mathbf{u}^{(0)}\right) \leq \kappa_{\mathrm{glob}} \text{ and } L_i/L_{i-1} \leq \kappa_{\mathrm{glob}}) \\
&\leq T_i.
\end{aligned}$$

Base Case. Consider $\texttt{StochBSLSRes}_m$. Suppose that $\beta_0\left(\mathbf{u}^{(0)}\right)$ satisfies Eq. E.7. Since $\mathbf{u}^{(0)}$ is not ambiguous we will shorten $\beta_0\left(\mathbf{u}^{(0)}\right), \rho\left(\mathbf{u}^{(0)}\right)$, and $\beta_m\left(\mathbf{u}^{(0)}\right)$ all to β_0, ρ, and β_m respectively. First we will prove the following claim:

Claim 96 (Fast Convergence Phase) *Recall $\tilde{\gamma}_m$. For any $t \leq \log_{\tilde{\gamma}_m}\left(\frac{1}{\beta_m^{N_{m+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{m-1}}{L_m}\frac{\mathbf{u}_{m-1}^{(0)}}{\mathbf{u}_m^{(0)}}\right) + 1$ we have*

$$\mathbf{u}^{(t)} = \mathbf{V}_m^t \mathbf{u}^{(0)}.$$

Moreover, defining

$$\beta_m\left(\mathbf{u}^{(0)}\right) := \rho\left(\mathbf{u}^{(0)}\right)^{N_m+2},$$

we have for any $j, k \in [m]$,

$$\mathbf{u}_k^{(t)} \leq \beta_0\beta_m^t\frac{1}{\tilde{\gamma}_m}\max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\}\mathbf{u}_j^{(t)}. \tag{J.5}$$

We will prove Claim 96 by induction. The case $t = 0$ holds immediately. Now suppose Claim 96 is true for t. First we prove that for any $t \leq T_m$,

$$\beta_m^t \leq 3. \tag{J.6}$$

We first bound β_m by $1 + (1/T_m)$,

$$\begin{aligned}\beta_m &= (1 + 3\delta m\beta_0)^{N_m+2} \\ &\leq 1 + 12(N_m + 2)\delta m\beta_0 \\ &\leq 1 + \frac{1}{T_m}. && (\beta_0 < 1/(24N_m T_m \delta m) \text{ by Eq. E.7})\end{aligned}$$

Therefore since $\sup_{x\geq 0}(1 + (1/x))^x \leq 3$ we have that for any $t \leq T_m$

$$\beta_m^t \leq \beta_m^{T_m} \leq \left(1 + \frac{1}{T_m}\right)^{T_m} \leq 3.$$

This concludes the proof of Eq. J.6. Now we can continue with our proof by induction of Claim 96. Suppose the inductive hypothesis (I.H.) that Claim 96 holds for t. Let $\overline{\mathbf{u}}^{(t+1)} := \mathbf{U}_m \mathbf{u}^{(t)}$. By Eq. J.3 we have,

$$\begin{aligned}\overline{\mathbf{u}}_m^{(t+1)} &\leq \gamma_m \mathbf{u}_m^{(t)} + \delta \sum_{k=1}^{m-1} \frac{L_k}{L_m} \mathbf{u}_k^{(t)} \\ &\leq \gamma_m \mathbf{u}_m^{(t)} + \delta\beta_0\beta_m^t \sum_{k=1}^{m-1} \frac{L_k}{L_m}\frac{L_m}{L_k}\mathbf{u}_m^{(t)} && \text{(I.H. of Claim 96 Eq. J.5)} \\ &\leq \left(\gamma_m + \beta_0\beta_m^t\delta(m-1)\right)\mathbf{u}_m^{(t)} \\ &\leq \tilde{\gamma}_m \mathbf{u}_m^{(t)}. && \text{(I.H. of Claim 96, Eq. J.6, and Eq. J.2)}\end{aligned}$$

Therefore

$$\mathbf{u}_m^{(t+1)} = \max\left\{\overline{\mathbf{u}}_m^{(t+1)}, \left(\mathbf{V}_i \mathbf{u}^{(t)}\right)_m\right\} = \max\left\{\overline{\mathbf{u}}^{(t+1)}, \tilde{\gamma}_m \mathbf{u}_m^{(t)}\right\} = \tilde{\gamma}_m \mathbf{u}_m^{(t)}. \tag{J.7}$$

Next for $j \leq m-1$,

$$\begin{aligned}\overline{\mathbf{u}}_j^{(t+1)} &\leq \mathbf{u}_j^{(t)} + \left(\delta \sum_{k=1}^{j} \frac{L_k L_j}{L_m^2}\mathbf{u}_k^{(t)}\right) + \left(\delta \sum_{k=j+1}^{m} \frac{L_k L_j}{L_m^2}\mathbf{u}_k^{(t)}\right) \\ &\leq \mathbf{u}_j^{(t)} + \left(\delta\beta_0\beta_m^t \sum_{k=1}^{j} \frac{L_k L_j}{L_m^2}\frac{L_j}{L_k}\mathbf{u}_j^{(t)}\right) + \left(\delta\beta_0\beta_m^t \sum_{k=j+1}^{m} \frac{L_k L_j}{L_m^2}\frac{L_k}{L_j}\mathbf{u}_j^{(t)}\right) \\ &\qquad \text{(I.H. of Claim 96 Eq. J.5)} \\ &= \left(1 + \beta_0\beta_m^t\delta m\right)\mathbf{u}_j^{(t)} \\ &\leq \rho \mathbf{u}_j^{(t)}. && \text{(Eq. J.6)}\end{aligned}$$

Therefore for $j \leq m-1$,

$$\mathbf{u}_j^{(t+1)} = \max\left\{\overline{\mathbf{u}}_j^{(t+1)}, \left(\mathbf{V}_i \mathbf{u}^{(t)}\right)_j\right\} = \max\left\{\overline{\mathbf{u}}^{(t+1)}, \rho \mathbf{u}_j^{(t)}\right\} = \rho \mathbf{u}_j^{(t)}. \tag{J.8}$$

Combining Eq. J.7 and Eq. J.8 with the I.H. of Claim 96 for t shows that $\mathbf{u}^{(t+1)} = \mathbf{V}_m^{t+1}\mathbf{u}^{(0)}$. Finally to establish Claim 96 we first show that if $t \le \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right) - 1$ then for all j, k,

$$\mathbf{u}_k^{(t)} \le \beta_0\beta_m^t \max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\}\mathbf{u}_j^{(t)}.$$

Then we show that this implies that for $t = \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$, for all j, k,

$$\mathbf{u}_k^{(t)} \le \beta_0\beta_m^t \frac{1}{\tilde{\gamma}_m}\max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\}\mathbf{u}_j^{(t)}.$$

If both $j, k \le m-1$ then

$$\begin{aligned}
\frac{\mathbf{u}_k^{(t+1)}}{\mathbf{u}_j^{(t+1)}} &= \frac{\rho\mathbf{u}_k^{(t)}}{\rho\mathbf{u}_j^{(t)}} && \text{(Eq. J.8)}\\
&\le \beta_0\beta_m^t \max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\} && \text{(I.H. of Claim 96)}\\
&\le \beta_0\beta_m^{t+1} \max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\}. && (\beta_m \ge 1)
\end{aligned}$$

Next suppose $k = m$. By Eq. J.7, $\mathbf{u}_m^{(t+1)} \le \mathbf{u}_m^{(t)}$ and by Eq. J.8 for $j \le m-1$, $\mathbf{u}_j^{(t+1)} \ge \mathbf{u}_j^{(t)}$. Therefore

$$\frac{\mathbf{u}_m^{(t+1)}}{\mathbf{u}_j^{(t+1)}} \le \frac{\mathbf{u}_m^{(t)}}{\mathbf{u}_j^{(t)}}.$$

Then combining this with the I.H. of Claim 96,

$$\mathbf{u}_m^{(t)} \le \beta_0\beta_m^t \max\left\{\frac{L_j}{L_m}, \frac{L_m}{L_j}\right\}\mathbf{u}_j^{(t)} \qquad \Longrightarrow \qquad \mathbf{u}_m^{(t+1)} \le \beta_0\beta_m^{t+1} \max\left\{\frac{L_j}{L_m}, \frac{L_m}{L_j}\right\}\mathbf{u}_j^{(t+1)}.$$

Finally we consider the case where $j = m$. First suppose $t+1 \le \log_{\tilde{\gamma}_m}\left(\frac{1}{\beta_m^{N_{m+1}}(\mathbf{u}^{(0)})}\frac{L_{m-1}}{L_m}\frac{\mathbf{u}_{m-1}^{(0)}}{\mathbf{u}_m^{(0)}}\right)$. By I.H. of Claim 96,

$$\mathbf{u}_m^{(t+1)} = \tilde{\gamma}_m^{t+1}\mathbf{u}_m^{(0)} \ge \frac{1}{\rho}\frac{L_{m-1}}{L_m}\mathbf{u}_{m-1}^{(0)} \ge \frac{L_{m-1}}{L_m}\frac{1}{\rho^{t+1}}\mathbf{u}_{m-1}^{(t)}. \tag{J.9}$$

Therefore for any $k \le m-1$,

$$\begin{aligned}
\mathbf{u}_k^{(t+1)} &= \rho\mathbf{u}_k^{(t)}\\
&\le \rho\beta_0\beta_m^t \max\left\{\frac{L_{m-1}}{L_k}, \frac{L_k}{L_{m-1}}\right\}\mathbf{u}_{m-1}^{(t)}\\
&\le \rho\beta_0\beta_m^t \max\left\{\frac{L_{m-1}}{L_k}, \frac{L_k}{L_{m-1}}\right\}\rho^{t+1}\frac{L_m}{L_{m-1}}\mathbf{u}_m^{(t+1)} && \text{(Eq. J.9)}\\
&= \rho^{t+2}\beta_0\beta_m^t\frac{L_m}{L_k}\mathbf{u}_m^{(t+1)}\\
&\le \beta_0\beta_m^{t+1}\frac{L_m}{L_k}\mathbf{u}_m^{(t+1)}. && (\rho^{t+2} \le \beta_m)
\end{aligned}$$

Next if $t+1 = \log_{\tilde{\gamma}_m}\left(\frac{1}{\beta_m^{N_{m+1}}(\mathbf{u}^{(0)})}\frac{L_{m-1}}{L_m}\frac{\mathbf{u}_{m-1}^{(0)}}{\mathbf{u}_m^{(0)}}\right)+1$ then we have

$$\mathbf{u}_m^{(t+1)} \geq \tilde{\gamma}_m \frac{L_{m-1}}{L_m}\frac{1}{\rho^{t+1}}\mathbf{u}_{m-1}^{(t)},$$

and so using the same logic as before,

$$\mathbf{u}_k^{(t+1)} \leq \frac{1}{\tilde{\gamma}_m}\beta_0\beta_m^{t+1}\frac{L_m}{L_k}\mathbf{u}_m^{(t+1)}.$$

This concludes the proof of Claim 96. Now we consider the steps for which $t > \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$. We have the following claim.

Claim 97 (Extraneous Steps Phase) *Suppose* $t > \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$. *Then*

$$\mathbf{u}^{(t+1)} = \mathbf{W}_m\mathbf{u}^{(t)}.$$

Moreover Eq. J.5 holds from Claim 96.

To prove Claim 97 we recall that for $\overline{\mathbf{u}}^{(t+1)} \coloneqq \mathbf{U}_m\mathbf{u}^{(t)}$ we have $\overline{\mathbf{u}}^{(t+1)} \leq \mathbf{V}_m\mathbf{u}^{(t)}$. Next since $\mathbf{W}_m$ is entry-wise larger than $\mathbf{V}_m$ we conclude

$$\mathbf{u}^{(t+1)} = \max\left\{\overline{\mathbf{u}}^{(t+1)}, \mathbf{W}_m\mathbf{u}^{(t)}\right\} = \mathbf{W}_m\mathbf{u}^{(t)}.$$

Finally we need to show that Eq. J.5 holds. The same reasoning that Eq. J.5 holds in the case $t \leq \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$ can be used to show that Eq. J.5 holds in the case where $t > \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$ and $j,k \leq m-1$. We simply need to consider the case where $j = m$ or $k = m$. First suppose $j = m$. We have for any $k \leq m-1$,

$$\begin{aligned}
\mathbf{u}_k^{(t+1)} &= \rho\mathbf{u}_k^{(t)} && \text{(I.H. of Claim 97)}\\
&\leq \rho\beta_0\beta_m^t\frac{L_m}{L_k}\mathbf{u}_m^{(t)} && \text{(I.H. of Claim 97)}\\
&= \rho\beta_0\beta_m^t\frac{L_m}{L_k}\mathbf{u}_m^{(t+1)} && ((\mathbf{W}_m)_{mm} = 1)\\
&\leq \beta_0\beta_m^{t+1}\frac{L_m}{L_k}\mathbf{u}_m^{(t+1)}. && (\rho \leq \beta_m)
\end{aligned}$$

Next suppose $k = m$ and $j \leq m-1$. This case is trivial since $\mathbf{u}_j^{(t+1)} \geq \mathbf{u}_j^{(t)}$ and $\mathbf{u}_i^{(t+1)} = \mathbf{u}_i^{(t)}$. Thus since $\beta_i \geq 1$,

$$\mathbf{u}_i^{(t)} \leq \beta_0\beta_m^t\mathbf{u}_j^{(t)} \implies \mathbf{u}_i^{(t+1)} \leq \beta_0\beta_m^{t+1}\mathbf{u}_j^{(t+1)}.$$

This concludes the proof of Claim 97.

Conclusion of Base Case To conclude we use Claim 96, Claim 97, and Eq. J.4 which guarantee that

$$\begin{aligned}\tilde{\mathbf{u}}_m &= \mathbf{u}_m^{(T_m)} \\ &= \tilde{\gamma}_m^{\left\lceil \log_{\tilde{\gamma}_m}\left(\frac{1}{\beta_m^{N_{m+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{m-1}}{L_m}\frac{\mathbf{u}_{m-1}^{(0)}}{\mathbf{u}_m^{(0)}}\right)\right\rceil}\mathbf{u}_m^{(0)} \\ &\leq \tilde{\gamma}_m^{\log_{\tilde{\gamma}_m}\left(\frac{1}{\beta_m^{N_{m+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{m-1}}{L_m}\frac{\mathbf{u}_{m-1}^{(0)}}{\mathbf{u}_m^{(0)}}\right)}\mathbf{u}_m^{(0)} \\ &= \frac{L_{m-1}}{L_m}\mathbf{u}_{m-1}^{(0)}.\end{aligned}$$

Next since $N_m = T_m$ we have for any $j \leq m-1$,

$$\tilde{\mathbf{u}}_j = \mathbf{u}_j^{(N_m)} = \rho^{N_m}\mathbf{u}_j^{(0)}.$$

This concludes the base case.

Inductive Case Suppose $\beta_0\left(\mathbf{u}^{(0)}\right)$ satisfies Eq. E.7.

Claim 98 (Fast Convergence Phase) *Suppose we are in the t^{th} iteration of* `StochBSLSRes`$_i$ *and assume*

$$t \leq \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil.$$

Then

$$\mathbf{u}^{(t)} = \mathbf{V}_i^t\mathbf{u}^{(0)},$$

and letting

$$\beta_i\left(\mathbf{u}\right) \coloneqq \rho\left(\mathbf{u}\right)^{(N_i+2)(N_{i+1}+1)+1}$$

we have for any $j, k \in [m]$,

$$\mathbf{u}_k^{(t)} \leq \beta_0\left(\mathbf{u}^{(0)}\right)\max\left\{\frac{1}{\tilde{\gamma}_i^2}\beta_i^t\left(\mathbf{u}^{(0)}\right), \frac{\rho^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}{\tilde{\gamma}_i}\right\}\max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\}\mathbf{u}_j^{(t)}. \tag{J.10}$$

We will prove Claim 98 by induction (be careful not to confuse this with the overarching proof by induction of Lemma 52). Again since $\mathbf{u}^{(0)}$ unambiguously denotes the initial vector passed to `StochBSLSRes`$_i$ for the majority of the proof we will shorten $\beta_0\left(\mathbf{u}^{(0)}\right), \rho\left(\mathbf{u}^{(0)}\right)$, and $\beta_i\left(\mathbf{u}^{(0)}\right)$ all to β_0, ρ, and β_i respectively. Before proceeding with our proof of Claim 98 we show that for any $t \leq T_i$,

$$\max\left\{\frac{1}{\tilde{\gamma}_i^2}\beta_i^t, \frac{\rho^{N_{i+1}}}{\tilde{\gamma}_i}\right\} \leq 3, \tag{J.11}$$

so that

$$\mathbf{u}_k^{(t)} \leq 3\beta_0\max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\}\mathbf{u}_j^{(t)}.$$

First we bound $\beta_i\left(\mathbf{u}^{(0)}\right)$ by $1+1/(3T_i)$. Using that $(1+x)^p \le 1+2px$ if $px \le 1/4$ and that $(3\beta_0\delta m)(2N_{i+1}) \le 1/4$ we have,

$$\begin{aligned}\beta_i &= (1+3\beta_0\delta m)^{2N_{i+1}} \\ &\le (1+12N_{i+1}\beta_0\delta m) \\ &\le 1+\frac{1}{3T_i}. && (\beta_0 \le 1/(144N_{i+1}\delta m T_i) \text{ by Eq. E.7})\end{aligned}$$

Therefore for any $t \le T_i$,

$$\frac{1}{\tilde{\gamma}_i^2}\beta_i^t \le 2\left(1+\frac{1}{3T_i}\right)^t \le 2\left(\frac{3}{2}\right) \le 3. \tag{J.12}$$

Next we have

$$\begin{aligned}\frac{\rho^{N_{i+1}}}{\tilde{\gamma}_i} &\le 2\rho^{N_{i+1}} && (\tilde{\gamma}_i \ge 1/2 \text{ since } \kappa_i \ge 1) \\ &\le 2(1+6N_{i+1}\beta_0\delta m) && ((1+x)^p \le 1+2px \text{ if } px \le 1/4 \text{ and } (3\beta_0\delta m)(N_{i+1}) \le 1/4) \\ &\le 3. && (12N_{i+1}\beta_0\delta m \le 1)\end{aligned}$$

Therefore Eq. J.11 holds. Next note that

$$(\gamma_i + 3\beta_0\delta m)\,\rho^{N_{i+1}} \le 1-\frac{1}{2\kappa_i}. \tag{J.13}$$

Indeed,

$$\begin{aligned}(\gamma_i + 3\beta_0\delta m)\,\rho^{N_{i+1}} &= \left(1-\frac{2}{\kappa_i}+\frac{1}{\kappa_i^2}+3\beta_0\delta m\right)\rho^{N_{i+1}} \\ &\le \left(1-\frac{2}{\kappa_i}+\frac{1}{\kappa_i^2}+3\beta_0\delta m\right)(1+6N_{i+1}\beta_0\delta m) \\ &\qquad ((1+x)^p \le 1+2px \text{ if } px \le 1/4 \text{ and } (3\beta_0\delta m)(N_{i+1}) \le 1/4) \\ &\le 1-\frac{2}{\kappa_i}+\frac{2}{\kappa_i^2}+6\beta_0\delta m+6N_{i+1}\beta_0\delta m && (6N_{i+1}\beta_0\delta m \le 1) \\ &\le 1-\frac{1}{2\kappa_i}. && (6(N_{i+1}+1)\beta_0\delta m \le 1/(2\kappa_i))\end{aligned}$$

With Eq. J.11 and Eq. J.13 in hand we are ready to prove Claim 98 by induction. The base case $t=0$ holds immediately by assumption. Now using the I.H. of Claim 98 at t we will show Claim 98 holds at $t+1$. Let $\overline{\mathbf{u}}^{(t+1)} := \mathbf{U}_{(1/L_i)}\mathbf{u}^{(t)}$. If Claim 98 holds for $\mathbf{u}^{(t)}$ then the necessary conditions to apply Lemma 52 to $\texttt{StochBSLSRes}_{i+1}$ hold. Indeed, when we call $\texttt{StochBSLSRes}_{i+1}$ to $\mathbf{u}^{(t)}$ we now have initial vector $\mathbf{u}^{(0)}_{\text{recursive}} = \mathbf{u}^{(t)}$ and so we simply need to show that Eq. E.7 holds for $\mathbf{u}^{(0)}_{\text{recursive}}$. To this end we see by Claim 98

$$\beta_0\left(\mathbf{u}^{(t)}\right) \le \beta_0\left(\mathbf{u}^{(0)}\right)\frac{1}{\tilde{\gamma}_i^2}\beta_i^t\left(\mathbf{u}^{(0)}\right) \le \beta_0\left(\mathbf{u}^{(0)}\right)\beta_{\text{total}}\left(\mathbf{u}^{(0)}\right).$$

Therefore

$$\begin{aligned}
\beta_0\left(\mathbf{u}^{(0)}_{\text{recursive}}\right)\beta_{\text{total}}^{m-(i+1)+1}\left(\mathbf{u}^{(0)}_{\text{recursive}}\right) &= \beta_0\left(\mathbf{u}^{(t)}\right)\beta_{\text{total}}^{m-(i+1)+1}\left(\mathbf{u}^{(t)}\right)\\
&\le \left(\beta_0\left(\mathbf{u}^{(0)}\right)\beta_{\text{total}}\left(\mathbf{u}^{(0)}\right)\right)\beta_{\text{total}}^{m-(i+1)+1}\left(\mathbf{u}^{(t)}\right)\\
&= \left(\beta_0\left(\mathbf{u}^{(0)}\right)\beta_{\text{total}}\left(\mathbf{u}^{(0)}\right)\right)\left(1+3\delta m\beta_0\left(\mathbf{u}^{(t)}\right)\right)^{T_{\max}(N_1+1)+1}\cdot\max_{\ell\in[m]}\left\{\frac{1}{\tilde{\gamma}_\ell^2}\right\}\\
&\le \left(\beta_0\left(\mathbf{u}^{(0)}\right)\beta_{\text{total}}\left(\mathbf{u}^{(0)}\right)\right)\left(1+3\delta m\beta_0\left(\mathbf{u}^{(0)}\right)\frac{1}{\tilde{\gamma}_i^2}\beta_i^t\left(\mathbf{u}^{(0)}\right)\right)^{T_{\max}(N_1+1)+1}\\
&\qquad\cdot\max_{\ell\in[m]}\left\{\frac{1}{\tilde{\gamma}_\ell^2}\right\}.
\end{aligned}$$

Now that we have expressed the above only in terms of $\mathbf{u}^{(0)}$ we drop its notation and use our typical shorthand. Thus we have,

$$\begin{aligned}
\beta_0\left(\mathbf{u}^{(0)}_{\text{recursive}}\right)\beta_{\text{total}}^{m-(i+1)+1}\left(\mathbf{u}^{(0)}_{\text{recursive}}\right) &\le \left(\beta_0\beta_{\text{total}}^{m-(i+1)+1}\right)\left(1+3\delta m\beta_0\frac{1}{\tilde{\gamma}_i^2}\beta_i^t\right)^{T_{\max}(N_1+1)+1}\cdot\max_{\ell\in[m]}\left\{\frac{1}{\tilde{\gamma}_\ell^2}\right\}\\
&\le \left(\beta_0\beta_{\text{total}}^{m-(i+1)+1}\right)\left(1+9\delta m\beta_0\frac{1}{\tilde{\gamma}_i^2}\right)^{T_{\max}(N_1+1)+1}\cdot\max_{\ell\in[m]}\left\{\frac{1}{\tilde{\gamma}_\ell^2}\right\}\\
&\le \left(\beta_0\beta_{\text{total}}^{m-(i+1)+1}\right)\beta_{\text{total}}\\
&= \beta_0\beta_{\text{total}}^{m-i+1}
\end{aligned}$$

Then since Eq. E.7 holds for $\texttt{StochBSLSRes}_i$ we can conclude that it holds for $\texttt{StochBSLSRes}_{i+1}$ with initialization $\mathbf{u}^{(t)}$. Recalling the notation that $\tilde{\mathbf{u}}^{(t)} = \texttt{StochBSLSRes}_{i+1}(\mathbf{u}^{(t)})$, by the (original) inductive hypothesis that Lemma 52 is true for $\texttt{StochBSLSRes}_{i+1}$ we have,

$$\begin{aligned}
\overline{\mathbf{u}}_i^{(t+1)} &= \gamma_i\tilde{\mathbf{u}}_i^{(t)} + \left(\delta\sum_{k=i+1}^{m}\frac{L_k}{L_i}\tilde{\mathbf{u}}_k^{(t)}\right) + \left(\delta\sum_{k=1}^{i-1}\frac{L_k}{L_i}\tilde{\mathbf{u}}_k^{(t)}\right)\\
&\le \gamma_i\rho^{N_{i+1}}\mathbf{u}_i^{(t)} + \left(\delta\sum_{k=i+1}^{m}\mathbf{u}_i^{(t)}\right) + \rho^{N_{i+1}}\left(\delta\sum_{k=1}^{i-1}\frac{L_k}{L_i}\mathbf{u}_k^{(t)}\right)\\
&\le \gamma_i\rho^{N_{i+1}}\mathbf{u}_i^{(t)} + \delta(m-i-1)\mathbf{u}_i^{(t)} + \rho^{N_{i+1}}\left(\delta\beta_0\beta_i^t\sum_{k=1}^{i-1}\frac{L_k}{L_i}\frac{L_i}{L_k}\mathbf{u}_i^{(t)}\right) \quad \text{(I.H. of Claim 98)}\\
&\le \left(\gamma_i\rho^{N_{i+1}} + \delta(m-i-1) + \rho^{N_{i+1}}\delta i\beta_0\beta_i^t\right)\mathbf{u}_i^{(t)}\\
&\le \left(\gamma_i + 3\beta_0\delta m\right)\rho^{N_{i+1}}\mathbf{u}_i^{(t)} \quad \text{(Eq. J.12)}\\
&\le \tilde{\gamma}_i\mathbf{u}_i^{(t)}. \quad \text{(Eq. J.13)}
\end{aligned}$$

Then

$$\mathbf{u}_i^{(t+1)} = \max\left\{\overline{\mathbf{u}}_i^{(t+1)}, \left(\mathbf{V}_i\mathbf{u}^{(t)}\right)_i\right\} = \tilde{\gamma}_i\mathbf{u}_i^{(t)}. \tag{J.14}$$

Now we turn our attention to $\mathbf{u}_j^{(t+1)}$ for $j \le i-1$. We have

$$\begin{aligned}
\overline{\mathbf{u}}_j^{(t+1)} &= \left(1-\frac{\mu_j}{L_i}\right)^2 \tilde{\mathbf{u}}_j^{(t)} + \left(\delta \sum_{k=1}^{j} \frac{L_k L_j}{L_i^2} \tilde{\mathbf{u}}_k^{(t)}\right) + \left(\delta \sum_{k=j+1}^{i-1} \frac{L_k L_j}{L_i^2} \tilde{\mathbf{u}}_k^{(t)}\right) + \left(2\delta \sum_{k=i}^{m} \frac{L_k L_j}{L_i^2} \tilde{\mathbf{u}}_k^{(t)}\right) \\
&\le \rho^{N_{i+1}} \left(\mathbf{u}_j^{(t)} + \delta \sum_{k=1}^{j} \frac{L_k L_j}{L_i^2} \mathbf{u}_k^{(t)} + \delta \sum_{k=j+1}^{i-1} \frac{L_k L_j}{L_i^2} \mathbf{u}_k^{(t)}\right) + \left(\delta \sum_{k=i}^{m} \frac{L_k L_j}{L_i^2} \frac{L_i}{L_k} \mathbf{u}_i^{(t)}\right) \\
&\qquad \text{(I.H. of Lemma 52)} \\
&\le \rho^{N_{i+1}} \left(\mathbf{u}_j^{(t)} + \delta\beta_i^t \sum_{k=1}^{j} \frac{L_k L_j}{L_i^2} \frac{L_j}{L_k} \mathbf{u}_j^{(t)} + \delta\beta_i^t \sum_{k=j+1}^{i-1} \frac{L_k L_j}{L_i^2} \frac{L_k}{L_j} \mathbf{u}_j^{(t)}\right) \\
&\quad + \left(2\delta\beta_i^t \sum_{k=i}^{m} \frac{L_k L_j}{L_i^2} \frac{L_i}{L_k} \frac{L_i}{L_j} \mathbf{u}_j^{(t)}\right) \qquad \text{(I.H. of Claim 98)} \\
&= \rho^{N_{i+1}} \left(1 + \delta\beta_i^t \sum_{k=1}^{j} \frac{L_j^2}{L_i^2} + \delta\beta_i^t \sum_{k=j+1}^{i-1} \frac{L_k^2}{L_i^2}\right) \mathbf{u}_j^{(t)} + \left(2\delta\beta_i^t \sum_{k=i}^{m} \mathbf{u}_j^{(t)}\right) \\
&\le \rho^{N_{i+1}+1} \mathbf{u}_j^{(t)}. \qquad \text{(Eq. J.12)}
\end{aligned}$$

Therefore,

$$\mathbf{u}_j^{(t+1)} = \max\left\{\overline{\mathbf{u}}_j^{(t+1)}, (\mathbf{V}_i \mathbf{u})_j\right\} = \rho^{N_{i+1}+1} \mathbf{u}_j^{(t)}. \tag{J.15}$$

Next we consider $\mathbf{u}_\ell^{(t+1)}$ for $\ell \ge i+1$. We have,

$$\begin{aligned}
\overline{\mathbf{u}}_\ell^{(t+1)} &\le \left(\frac{L_\ell}{L_i}\right)^2 \tilde{\mathbf{u}}_\ell^{(t)} + \left(\delta \sum_{k=i+1}^{m} \frac{L_k L_\ell}{L_i^2} \tilde{\mathbf{u}}_k^{(t)}\right) + \left(\delta \sum_{k=1}^{i} \frac{L_k L_\ell}{L_i^2} \tilde{\mathbf{u}}_k^{(t)}\right) \\
&\le \left(\frac{L_\ell}{L_i}\right)^2 \frac{L_i}{L_\ell} \mathbf{u}_i^{(t)} + \left(2\delta \sum_{k=i+1}^{m} \frac{L_k L_\ell}{L_i^2} \frac{L_i}{L_k} \mathbf{u}_i^{(t)}\right) + \left(\delta \sum_{k=1}^{i} \frac{L_k L_\ell}{L_i^2} \rho^{N_{i+1}} \mathbf{u}_k^{(t)}\right) \\
&\qquad \text{(I.H. of Lemma 52)} \\
&\le \frac{L_\ell}{L_i} \left(\left(1 + 2\delta(m-i-1)\right) \mathbf{u}_i^{(t)} + \left(\delta \sum_{k=1}^{i} \frac{L_k}{L_i} \rho^{N_{i+1}} \beta_0 \beta_i^t \frac{L_i}{L_k} \mathbf{u}_i^{(t)}\right)\right) \qquad \text{(I.H. of Claim 98)} \\
&\le \frac{L_\ell}{L_i} \left(1 + 2\delta(m-i-1) + 3\beta_0 \delta i \rho^{N_{i+1}}\right) \mathbf{u}_i^{(t)} \qquad (\beta_i^t \le 3 \text{ by Eq. J.12}) \\
&\le \frac{L_\ell}{L_i} \rho^{N_{i+1}+1} \mathbf{u}_i^{(t)}.
\end{aligned}$$

Therefore,

$$\mathbf{u}_\ell^{(t+1)} = \max\left\{\overline{\mathbf{u}}_\ell^{(t+1)}, \left(\mathbf{V}_i \mathbf{u}^{(t)}\right)_\ell\right\} = \frac{L_\ell}{L_i} \rho^{N_{i+1}+1} \mathbf{u}_i^{(t)}. \tag{J.16}$$

Combining Eq. J.14, Eq. J.15, and Eq. J.16 we conclude

$$\mathbf{u}^{(t+1)} = \mathbf{V}_i \mathbf{u}^{(t)}.$$

Finally we prove Eq. J.10 of Claim 98. First we will prove by induction that if the pair (j,k) is such that Case One, Case Two, or Case Three holds in Table 1 *and* $t \le \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)-1$, then we have

$$\mathbf{u}_k^{(t)} \le \beta_0\beta_i^t \max\left\{\frac{L_k}{L_j},\frac{L_j}{L_k}\right\}\mathbf{u}_j^{(t)}. \tag{J.17}$$

Instead if $t \in \left[\log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)-1, \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)+1\right]$

$$\mathbf{u}_k^{(t)} \le \beta_0\beta_i^t\left(\frac{1}{\tilde{\gamma}_i^2}\right)\max\left\{\frac{L_k}{L_j},\frac{L_j}{L_k}\right\}\mathbf{u}_j^{(t)}.$$

After proving this we will show that if the pair (j,k) is such that $j \le i$ and $k \ge i+1$ then

$$\mathbf{u}_k^{(t)} \le \beta_0\frac{\rho^{N_{i+1}}}{\tilde{\gamma}_i}\frac{L_k}{L_j}\mathbf{u}_j^{(t)}.$$

Table 1: Cases for which Eq. J.17 holds

Case One	$j \le i$	$k \le i$
Case Two	$j \ge i+1$	$k \le i$
Case Three	$j \ge i+1$	$k \ge i+1$

To that end, we begin by considering Case One where $j \le i$ and $k \le i$. First assume $j \le i-1$. Then we have

$$\frac{\mathbf{u}_k^{(t+1)}}{\mathbf{u}_j^{(t+1)}} \le \frac{\rho^{N_{i+1}+1}\mathbf{u}_k^{(t)}}{\rho^{N_{i+1}+1}\mathbf{u}_j^{(t)}} = \frac{\mathbf{u}_k^{(t)}}{\mathbf{u}_j^{(t)}}.$$

Therefore since $\beta_i \ge 1$,

$$\mathbf{u}_k^{(t)} \le \beta_0\beta_i^t \max\left\{\frac{L_k}{L_j},\frac{L_j}{L_k}\right\}\mathbf{u}_j^{(t)} \implies \mathbf{u}_k^{(t+1)} \le \beta_0\beta_i^{t+1}\max\left\{\frac{L_k}{L_j},\frac{L_j}{L_k}\right\}\mathbf{u}_j^{(t+1)}.$$

Next we set $j=i$. Suppose $t+1 \le \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)$. Then we have

$$\mathbf{u}_i^{(t+1)} = \tilde{\gamma}_i^{t+1}\mathbf{u}_i^{(0)} \ge \tilde{\gamma}_i^{\log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}\left(\mathbf{u}^{(0)}\right)}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)}\mathbf{u}_i^{(0)} = \frac{1}{\rho^{N_{i+1}}}\frac{L_{i-1}}{L_i}\mathbf{u}_{i-1}^{(0)} \ge \frac{L_{i-1}}{L_i}\frac{1}{\rho^{(t+1)(N_{i+1}+1)}}\mathbf{u}_{i-1}^{(t)}.$$

Therefore if $j = i$ and $k \leq i$,

$$\begin{aligned}
\mathbf{u}_k^{(t+1)} &\leq \rho^{N_{i+1}+1}\mathbf{u}_k^{(t)} \\
&\leq \rho^{N_{i+1}+1}\beta_0\beta_i^t \max\left\{\frac{L_{i-1}}{L_k}, \frac{L_k}{L_{i-1}}\right\}\mathbf{u}_{i-1}^{(t)} \\
&\leq \rho^{N_{i+1}+1}\beta_0\beta_i^t \max\left\{\frac{L_{i-1}}{L_k}, \frac{L_k}{L_{i-1}}\right\}\frac{L_i}{L_{i-1}}\rho^{(t+1)(N_{i+1}+1)}\mathbf{u}_i^{(t+1)} \\
&= \rho^{(t+2)(N_{i+1}+1)+1}\beta_0\beta_i^t\frac{L_i}{L_k}\mathbf{u}_i^{(t+1)} \\
&\leq \beta_0\beta_i^{t+1}\frac{L_i}{L_k}\mathbf{u}_i^{(t+1)}. && (\rho^{(t+2)(N_{i+1}+1)+1} \leq \beta_i)
\end{aligned}$$

Next we consider Case Two where $j \geq i+1$ and $k \leq i$. Using similar reasoning we find,

$$\begin{aligned}
\mathbf{u}_k^{(t+1)} &\leq \rho^{N_{i+1}}\mathbf{u}_k^{(t)} && \text{(I.H. of Claim 98)} \\
&\leq \rho^{N_{i+1}}\beta_0\beta_i^t\frac{L_i}{L_k}\mathbf{u}_i^{(t)} && \text{(I.H. of Claim 98)} \\
&= \rho^{N_{i+1}}\beta_0\beta_i^t\frac{L_i}{L_k}\left(\frac{L_i}{L_j}\frac{1}{\rho^{N_{i+1}}}\mathbf{u}_j^{(t+1)}\right) && \text{(Eq. J.16)} \\
&\leq \beta_0\beta_i^t\frac{L_i}{L_k}\frac{L_j}{L_i}\mathbf{u}_j^{(t+1)} && (L_i < L_j) \\
&\leq \beta_0\beta_i^{t+1}\frac{L_j}{L_k}\mathbf{u}_j^{(t+1)}. && (\beta_i \geq 1)
\end{aligned}$$

Finally we consider Case Three where $j \geq i+1$ and $k \geq i+1$. We have

$$\frac{\mathbf{u}_k^{(t+1)}}{\mathbf{u}_j^{(t+1)}} = \frac{\frac{L_k}{L_i}\rho^{N_{i+1}+1}\mathbf{u}_i^{(t)}}{\frac{L_j}{L_i}\rho^{N_{i+1}+1}\mathbf{u}_i^{(t)}} = \frac{L_k}{L_j}.$$

Now we address the case where $t \in \left[\log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right) - 1, \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right) + 1\right]$.
For any t in this range

$$\mathbf{u}_i^{(t)} = \tilde{\gamma}_i^t\mathbf{u}_i^{(0)} \geq \tilde{\gamma}_i^{\log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)+1}\mathbf{u}_i^{(0)} \geq \tilde{\gamma}_i\frac{1}{\rho}\frac{L_{i-1}}{L_i}\mathbf{u}_{i-1}^{(0)} = \tilde{\gamma}_i\frac{L_{i-1}}{L_i}\frac{1}{\rho^{(t+1)(N_{i+1}+1)}}\mathbf{u}_{i-1}^{(t)}.$$

Therefore, using the same logic as before we have that if $j = i$ and $k \leq i$,

$$\mathbf{u}_k^{(t)} \leq \beta_0\beta_i^t\frac{1}{\tilde{\gamma}_i}^2\frac{L_i}{L_k}\mathbf{u}_i^{(t)}.$$

All other cases remain the same.
As promised, we now consider the case where $j \leq i$ and $k \geq i+1$ separately. By I.H. of Claim 98 for t we have

$$\mathbf{u}_k^{(t+1)} = \frac{L_k}{L_i}\rho^{N_{i+1}}\mathbf{u}_i^{(t)} = \frac{L_k}{L_i}\rho^{N_{i+1}}\mathbf{u}_i^{(t)} = \left(\frac{\rho^{N_{i+1}}}{\tilde{\gamma}_i}\right)\frac{L_k}{L_i}\mathbf{u}_i^{(t+1)}.$$

Also note that for any $j \le i$, $\mathbf{u}_i^{(t+1)} \le \mathbf{u}_i^{(t)}$ and $\mathbf{u}_j^{(t+1)} \ge \mathbf{u}_j^{(t)}$ and so for any $t \le \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right) - 1$,

$$\mathbf{u}_i^{(0)} \le \beta_0 \mathbf{u}_j^{(0)} \implies \mathbf{u}_i^{(t)} \le \beta_0 \mathbf{u}_j^{(t)}.$$

Therefore for any $j \le i$,

$$\begin{aligned}
\mathbf{u}_k^{(t+1)} &= \left(\frac{\rho^{N_{i+1}}}{\tilde{\gamma}_i}\right)\frac{L_k}{L_i}\mathbf{u}_i^{(t+1)} \\
&\le \left(\frac{\rho^{N_{i+1}}}{\tilde{\gamma}_i}\right)\frac{L_k}{L_i}\left(\beta_0 \mathbf{u}_j^{(t+1)}\right) \\
&\le \left(\frac{\beta_0\rho^{N_{i+1}}}{\tilde{\gamma}_i}\right)\frac{L_k}{L_j}\mathbf{u}_j^{(t+1)}. && (L_j \le L_i)
\end{aligned}$$

We conclude that for $t \le \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right) + 1$ and for any j, k,

$$\mathbf{u}_k^{(t)} \le \beta_0 \max\left\{\beta_i^t, \frac{\rho^{N_{i+1}}}{\tilde{\gamma}_i}\right\}\max\left\{\frac{L_j}{L_k}, \frac{L_k}{L_j}\right\}\mathbf{u}_j^{(t)}.$$

This concludes the proof of Claim 98. Now we consider the steps for which $t > \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$. We have the following claim.

Claim 99 (Extraneous Steps Phase.) *Suppose* $t > \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$. *Then*

$$\mathbf{u}^{(t+1)} = \mathbf{W}_i \mathbf{u}^{(t)}.$$

Moreover, Eq. J.10 holds from Claim 98.

To prove Claim 99 we first need to show that

$$\max\left\{\mathbf{U}_i\tilde{\mathbf{u}}_i^{(t)}, \mathbf{W}_i\tilde{\mathbf{u}}_i^{(t)}\right\} = \mathbf{W}_i\tilde{\mathbf{u}}_i^{(t)}.$$

However we have already shown this while proving earlier that

$$\max\left\{\mathbf{U}_i\tilde{\mathbf{u}}_i^{(t)}, \mathbf{V}_i\tilde{\mathbf{u}}_i^{(t)}\right\} = \mathbf{V}_i\tilde{\mathbf{u}}_i^{(t)}.$$

All that is left is to note that $\mathbf{V}_i$ and $\mathbf{W}_i$ are identical except $(\mathbf{V}_i)_{ii} < (\mathbf{W}_i)_{ii}$. Next we show that Eq. J.10 holds when $t > \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$. Note that for any $j \le i-1$

$$\mathbf{u}_j^{(t+1)} = \left(\mathbf{W}_i\mathbf{u}^{(t)}\right)_j = \left(\mathbf{V}_i\mathbf{u}^{(t)}\right)_j$$

and for any $j \ge i$,

$$\mathbf{u}_j^{(t+1)} = \left(\mathbf{W}_i\mathbf{u}^{(t)}\right)_j \ge \left(\mathbf{V}_i\mathbf{u}^{(t)}\right)_j.$$

Therefore whenever the pair (j,k) is such that $k \leq i-1$ we can recycle the proof showing Eq. J.10 holds in Claim 98 (i.e. $t \leq \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$). We simply need to consider the case where the pair (j,k) is such that $k \geq i$. Examining the proof of Eq. J.10 for Claim 98 we see that it suffices to simply consider $k=i$. First suppose that $j \leq i-1$. Then we are done since $\mathbf{u}_i^{(t+1)} = \mathbf{u}_i^{(t)}$ and $\mathbf{u}_j^{(t+1)} \geq \mathbf{u}_j^{(j)}$. Therefore since $\beta_i \geq 1$,

$$\mathbf{u}_i^{(t)} \leq \beta_0\beta_i^t \max\left\{\frac{L_i}{L_j}, \frac{L_j}{L_i}\right\}\mathbf{u}_j^{(t)} \implies \mathbf{u}_i^{(t+1)} \leq \beta_0\beta_i^{t+1}\max\left\{\frac{L_i}{L_j}, \frac{L_j}{L_i}\right\}\mathbf{u}_j^{(t+1)}.$$

Next suppose that $j \geq i+1$. Then we have for any t,

$$\mathbf{u}_i^{(t)} = \frac{1}{\rho^{N_{i+1}+1}}\frac{L_i}{L_j}\mathbf{u}_j^{(t)} \leq \frac{L_i}{L_j}\mathbf{u}_j^{(t)}.$$

Therefore we conclude that Eq. J.10 holds when $t > \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil$.

Conclusion of Inductive Case To conclude we recall Eq. J.4,

$$T_i \geq \left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil.$$

Using this we see,

$$\begin{aligned}
\mathbf{u}_i^{(T_i)} &= \tilde{\gamma}_i^{\left\lceil \log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)\right\rceil}\mathbf{u}_i^{(0)} \\
&\leq \tilde{\gamma}_i^{\log_{\tilde{\gamma}_i}\left(\frac{1}{\beta_i^{N_{i+1}}(\mathbf{u}^{(0)})}\frac{L_{i-1}}{L_i}\frac{\mathbf{u}_{i-1}^{(0)}}{\mathbf{u}_i^{(0)}}\right)}\mathbf{u}_i^{(0)} \\
&= \frac{1}{\rho^{N_{i+1}}}\frac{L_{i-1}}{L_i}\mathbf{u}_{i-1}^{(0)}
\end{aligned}$$

Next, since we return $\texttt{StochBSLSRes}_{i+1}(\mathbf{u}^{(T_i)})$ we have by the Inductive Hypothesis that Lemma 52 holds for $\texttt{StochBSLSRes}_{i+1}$ that if $\tilde{\mathbf{u}} = \texttt{StochBSLSRes}_{i+1}(\mathbf{u}^{(T_i)})$ then

$$\tilde{\mathbf{u}}_i \leq \rho^{N_{i+1}}\mathbf{u}_i^{(T_i)} \leq \frac{L_{i-1}}{L_i}\mathbf{u}_{i-1}^{(0)}.$$

Then for all $j \geq i+1$ we have

$$\tilde{\mathbf{u}}_j \leq \frac{L_i}{L_j}\mathbf{u}_i^{(T_i)} \leq \frac{L_i}{L_j}\left(\frac{L_{i-1}}{L_i}\mathbf{u}_{i-1}^{(0)}\right) \leq \frac{L_{i-1}}{L_j}\mathbf{u}_{i-1}^{(0)}.$$

Finally since $N_i = N_{i+1}(2T_i+1)$, for all $j \leq i-1$

$$\mathbf{u}_j^{(T_m)} \leq \rho^{T_iN_{i+1}}\mathbf{u}_j^{(0)} \leq \rho^{N_i}\mathbf{u}_j^{(0)}.$$

■

Appendix K. Miscellaneous helper lemmas

In this section we provide some helper lemmas and that are used throughout the other sections.

Lemma 100 *For any integer pairs (p, q) such that $p \geq q \geq 1$, we have $\binom{p}{q} \leq p \cdot \binom{p-1}{q-1}$.*

Proof [Proof of Lemma 100] By definition

$$\binom{p}{q} = \frac{p!}{q!(p-q)!} = \frac{p}{q}\frac{(p-1)!}{(q-1)!(p-q)!} = \frac{p}{q} \cdot \binom{p-1}{q-1} \leq p \cdot \binom{p-1}{q-1}.$$

■

Lemma 101

$$\prod_{j=1}^{\infty} \left(1 - 2^{-j}\right) > 0.288.$$

The lemma is standard from combinatoric analysis. We provide an elementary proof here for completeness.

Proof [Proof of Lemma 101] Decompose the infinity product into two parts

$$\prod_{j=1}^{\infty}(1 - 2^{-j}) = \left(\prod_{j=1}^{10}(1 - 2^{-j})\right) \cdot \exp\left(\sum_{j=11}^{\infty} \log(1 - 2^{-j})\right).$$

Since $\log(1 - 2^{-j}) \geq -2^{-j+1}$ for $j \geq 1$, we obtain

$$\exp\left(\sum_{j=11}^{\infty} \log(1 - 2^{-j})\right) \geq \exp\left(-\sum_{j=11}^{\infty} 2^{-j+1}\right) = \exp\left(-2^{-9}\right) > 0.998.$$

On the other hand we know that $\prod_{j=1}^{10}(1 - 2^{-j}) > 0.289$. Therefore $\prod_{j=1}^{\infty}(1 - 2^{-j}) > 0.289 \times 0.998 > 0.288$. ■

Lemma 102 *Suppose for any i, $a_i, b_i, x_i > 0$. Then for any n,*

$$\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n b_i x_i} \leq \max_{i \in [n]} \left\{\frac{a_i}{b_i}\right\}.$$

Proof [Proof of Lemma 102] This follows immediately from the fact that we can write the expression on left-hand side of the inequality as a convex combination of the fractions $\frac{a_i}{b_i}$ (with coefficients $c_i = \frac{b_i x_i}{\sum_{i=1}^n b_i x_i}$):

$$\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n b_i x_i} = \sum_{i=1}^{n} \frac{a_i}{b_i}\frac{b_i x_i}{\sum_{i=1}^n b_i x_i} \leq \max_{i \in [n]} \left\{\frac{a_i}{b_i}\right\} \sum_{i=1}^{n} \frac{b_i x_i}{\sum_{i=1}^n b_i x_i} = \max_{i \in [n]} \left\{\frac{a_i}{b_i}\right\}.$$

■