Proceedings of Machine Learning Research vol 75:1–60, 2018 31st Annual Conference on Learning Theory

Accelerating Stochastic Gradient Descent for Least Squares Regression

Prateek Jain PRAJAIN@MICROSOFT.COM
Praneeth Netrapalli PRANEETH@MICROSOFT.COM
Microsoft Research, Bangalore, India

Sham M. Kakade SHAM@CS.WASHINGTON.EDU
Rahul Kidambi RKIDAMBI@UW.EDU
University of Washington, Seattle, WA, USA

Aaron Sidford SIDFORD@STANFORD.COM
Stanford University, Palo Alto, CA, USA

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

There is widespread sentiment that fast gradient methods (*e.g.* Nesterov's acceleration, conjugate gradient, heavy ball) are not effective for the purposes of stochastic optimization due to their instability and error accumulation. Numerous works have attempted to quantify these instabilities in the face of either statistical or non-statistical errors (Paige, 1971; Proakis, 1974; Polyak, 1987; Greenbaum, 1989; Devolder et al., 2014). This work considers these issues for the special case of stochastic approximation for the least squares regression problem, and our main result refutes this conventional wisdom by showing that acceleration can be made robust to statistical errors. In particular, this work introduces an accelerated stochastic gradient method that provably achieves the minimax optimal statistical risk faster than stochastic gradient descent. Critical to the analysis is a sharp characterization of accelerated stochastic gradient descent as a stochastic process. We hope this characterization gives insights towards the broader question of designing simple and effective accelerated stochastic methods for more general convex and non-convex optimization problems.

Keywords: Stochastic Approximation, Acceleration, Stochastic Gradient Descent, Accelerated Stochastic Gradient Descent, Least Squares Regression.

1. Introduction

Stochastic gradient descent (SGD) is the workhorse algorithm for optimization in machine learning and stochastic approximation problems; improving its runtime dependencies is a central issue in large scale stochastic optimization that often arise in machine learning problems at scale (Bottou and Bousquet, 2007), where one can only resort to streaming algorithms.

This work examines these broader runtime issues for the special case of stochastic approximation in the following least squares regression problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} P(\mathbf{x}), \ \ \text{where, } P(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{2}\cdot\mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[(b-\langle\mathbf{x},\mathbf{a}\rangle)^2\right], \tag{1}$$

where we have access to a *stochastic first order oracle*, which, when provided with $\mathbf{x}$ as an input, returns a noisy unbiased stochastic gradient using a tuple $(\mathbf{a}, b)$ sampled from $\mathcal{D}(\mathbb{R}^d\times\mathbb{R})$, with d being the dimension of the problem. A query to the stochastic first-order oracle at $\mathbf{x}$ produces:

$$\widehat{\nabla} P(\mathbf{x}) = -(b-\langle\mathbf{a},\mathbf{x}\rangle)\cdot\mathbf{a}. \tag{2}$$

Algorithm	Final error	Runtime	Memory
Accelerated SVRG (Allen-Zhu, 2016)	$\mathcal{O}\left(\frac{\sigma^2 d}{n}\right)$	$(n+\sqrt{n\kappa})d\log\left(\frac{P(\mathbf{x}_0)-P(\mathbf{x}_*)}{(\sigma^2 d/n)}\right)$	nd
Streaming SVRG (Frostig et al., 2015b) Iterate Averaged SGD (Jain et al., 2016)	$\mathcal{O}\left(\exp\left(\frac{-n}{\kappa}\right)\cdot(P(\mathbf{x}_0)-P(\mathbf{x}_*))+\frac{\sigma^2 d}{n}\right)$	nd	$\mathcal{O}(d)$
Accelerated Stochastic Gradient Descent (this paper)	$\mathcal{O}^*\left(\exp\left(\frac{-n}{\sqrt{\kappa\widetilde{\kappa}}}\right)(P(\mathbf{x}_0)-P(\mathbf{x}_*))\right)+\mathcal{O}\left(\frac{\sigma^2 d}{n}\right)$	nd	$\mathcal{O}(d)$

Table 1: Comparison of this work to the best known non-asymptotic results (Frostig et al., 2015b; Jain et al., 2016) for the least squares stochastic approximation problem. Here, d, n are the problem dimension, number of samples; κ, $\widetilde{\kappa}$ denote the condition number and statistical condition number of the distribution; σ^2, $P(\mathbf{x}_0)-P(\mathbf{x}_*)$ denote the noise level and initial excess risk, $\mathcal{O}^*$ hides lower order terms in $d, \kappa, \widetilde{\kappa}$ (see section 2 for definitions and a proof for $\widetilde{\kappa}\leq\kappa$). Note that Accelerated SVRG (Allen-Zhu, 2016) is not a streaming algorithm.

Note $\mathbb{E}\left[\widehat{\nabla}P(\mathbf{x})\right]=\nabla P(\mathbf{x})$ (i.e. eq(2) is an unbiased estimate). Note that nearly all practical stochastic algorithms use sampled gradients of the specific form as in equation 2. We discuss differences to the more general stochastic first order oracle (Nemirovsky and Yudin, 1983) in section 1.4.

Let $\mathbf{x}^*\stackrel{\text{def}}{=}\arg\min_{\mathbf{x}}P(\mathbf{x})$ be a population risk minimizer. Given any estimation procedure which returns $\widehat{\mathbf{x}}_n$ using n samples, define the *excess risk* (which we also refer to as the *generalization error* or the *error*) of $\widehat{\mathbf{x}}_n$ as $\mathbb{E}\left[P(\widehat{\mathbf{x}}_n)\right]-P(\mathbf{x}^*)$. Now, equipped a stochastic first-order oracle (equation (2)), our goal is to provide a computationally efficient (and streaming) estimation method whose excess risk is comparable to the optimal statistical minimax rate.

In the limit of large n, this minimax rate is achieved by the *empirical risk minimizer* (ERM), which is defined as follows. Given n i.i.d. samples $\mathcal{S}_n=\{(\mathbf{a}_i,b_i)\}_{i=1}^n$ drawn from $\mathcal{D}$, define

$$\widehat{\mathbf{x}}_n^{\text{ERM}}\stackrel{\text{def}}{=}\arg\min_{\mathbf{x}}P_n(\mathbf{x}),\ \text{where } P_n(\mathbf{x})\stackrel{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(b_i-\mathbf{a}_i^\top\mathbf{x}\right)^2,$$

where $\widehat{\mathbf{x}}_n^{\text{ERM}}$ denotes the ERM over the samples $\mathcal{S}_n$. For the case of additive noise models (i.e. where $b=\mathbf{a}^\top\mathbf{x}^*+\epsilon$, with ϵ being independent of $\mathbf{a}$), the minimax estimation rate is $d\sigma^2/n$ (Kushner and Clark, 1978; Polyak and Juditsky, 1992; Lehmann and Casella, 1998; van der Vaart, 2000), i.e.:

$$\lim_{n\to\infty}\frac{\mathbb{E}_{\mathcal{S}_n}[P(\widehat{\mathbf{x}}_n^{\text{ERM}})]-P(\mathbf{x}^*)}{d\sigma^2/n}=1, \tag{3}$$

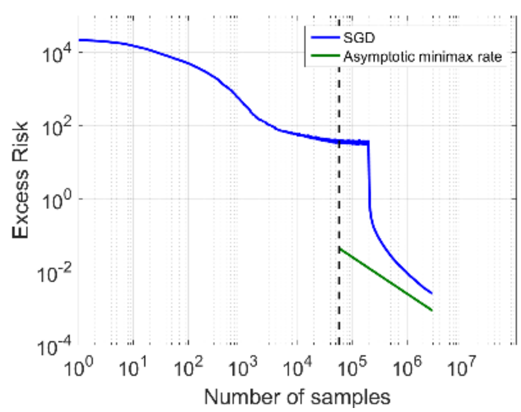
where $\sigma^2=\mathbb{E}\left[\epsilon^2\right]$ is the variance of the additive noise and the expectation is over the samples $\mathcal{S}_n$ drawn from $\mathcal{D}$. The seminal works of Ruppert (1988); Polyak and Juditsky (1992) proved that a certain averaged stochastic gradient method enjoys this minimax rate, in the limit. The question we seek to address is: how fast (in a non-asymptotic sense) can we achieve the minimax rate of $d\sigma^2/n$?

1.1. Review: Acceleration with Exact Gradients

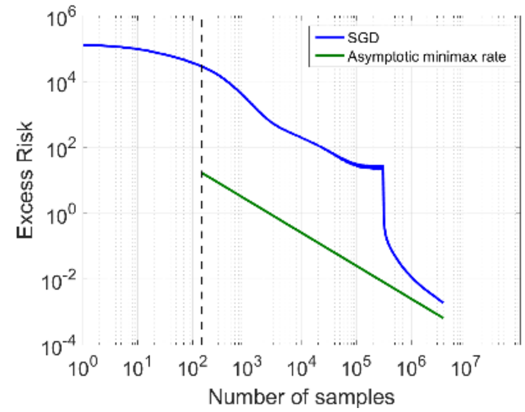
Let us review results in convex optimization in the exact first-order oracle model. Running $t-$steps of gradient descent (Cauchy, 1847) with an exact first-order oracle yields the following guarantee:

$$P(\mathbf{x}_t)-P(\mathbf{x}^*)\leq\exp\left(-t/\kappa_o\right)\cdot\left(P(\mathbf{x}_0)-P(\mathbf{x}^*)\right),$$

where $\mathbf{x}_0$ is the starting iterate, $\kappa_o=\lambda_{\max}(\mathbf{H})/\lambda_{\min}(\mathbf{H})$ is the condition number of $P(.)$, where, $\lambda_{\max}(\mathbf{H})$, $\lambda_{\min}(\mathbf{H})$ are the largest and smallest eigenvalue of the hessian $\mathbf{H}=\nabla^2P(\mathbf{x})=\mathbb{E}\left[\mathbf{a}\mathbf{a}^\top\right]$.



(*a*) Discrete Distribution



(*b*) Gaussian Distribution

Figure 1: Plot of total error vs number of samples for averaged SGD and the minimax risk (green) of $d\sigma^2/n$ for the discrete and Gaussian distributions with $d = 50$, $\kappa \approx 10^5$ (see section 1.2 for details on the distribution). The kink in the SGD curve represents when the tail-averaging phase begins (Jain et al., 2016); this point is chosen appropriately. The vertical dashed line shows the sample size at which the empirical covariance, $\frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i\mathbf{a}_i^\top$, becomes full rank, which is shown at $\frac{1}{\min_i p_i}$ in the discrete case and d in the Gaussian case. With fewer samples than this (i.e. before the dashed line), it is information theoretically not possible to guarantee non-trivial risk (without further assumptions). For the Gaussian case, note how the behavior of SGD is far from the minimax risk; it is this behavior that one might hope to improve upon. See the text for more discussion.

Thus gradient descent requires $\mathcal{O}(\kappa_o)$ oracle calls to solve the problem to a given target accuracy, which is sub-optimal amongst the class of methods with access to an exact first-order oracle (Nesterov, 2004). This sub-optimality can be addressed through Nesterov's Accelerated Gradient Descent (Nesterov, 1983), which when run for t-steps, yields the following guarantee:

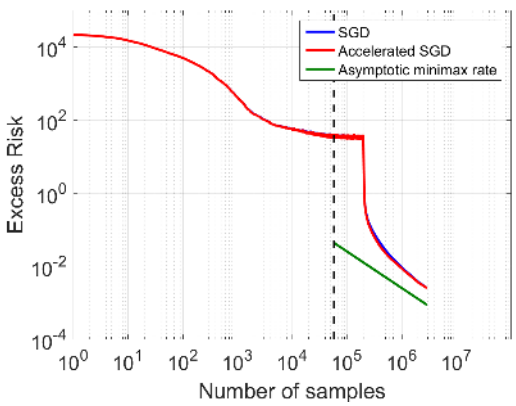
$$P(\mathbf{x}_t) - P(\mathbf{x}^*) \leq \exp\big(-t/\sqrt{\kappa_o}\big)\cdot\big(P(\mathbf{x}_0) - P(\mathbf{x}^*)\big),$$

which implies that $\mathcal{O}(\sqrt{\kappa_o})$ oracle calls are sufficient to achieve a given target accuracy. This matches the oracle lower bounds (Nesterov, 2004) that state that $\Theta(\sqrt{\kappa_o})$ calls to the exact first order oracle are necessary to achieve a given target accuracy. The conjugate gradient method (Hestenes and Stiefel, 1952) and heavy ball method (Polyak, 1964) are also known to obtain this convergence rate for solving a system of linear equations and for quadratic functions. These methods are termed fast gradient methods owing to the improvements offered by these methods over Gradient Descent. This paper seeks to address the question: "Can we accelerate stochastic approximation in a manner similar to what has been achieved with the exact first order oracle model?"

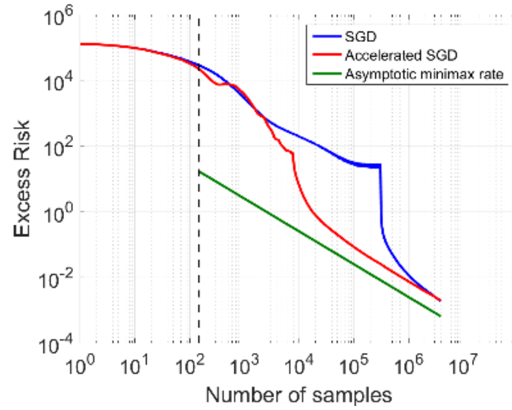
1.2. A thought experiment: Is Accelerating Stochastic Approximation possible?

Let us recollect known results in stochastic approximation for the least squares regression problem (in equation 1). Running n-steps of tail-averaged SGD (Jain et al., 2016) (or, streaming SVRG (Frostig et al., 2015b)[1]) provides an output $\widehat{\mathbf{x}}_n$ that satisfies the following excess risk bound:

1. Streaming SVRG does not function in the stochastic first order oracle model (Frostig et al., 2015b)



(*a*) Discrete Distribution



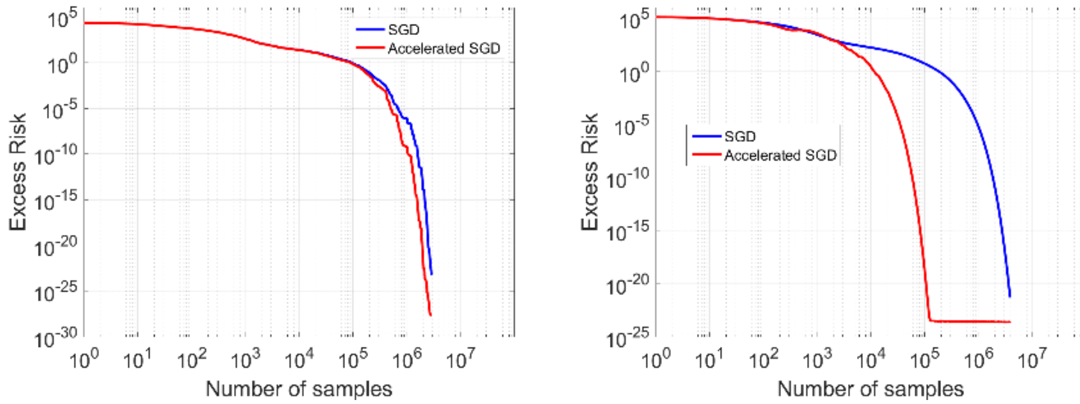
(*b*) Gaussian Distribution

Figure 2: Plot of total error vs number of samples for averaged SGD, (this paper's) accelerated SGD method and the minimax risk for the discrete and Gaussian distributions with $d = 50, \kappa \approx 10^5$ (see section 1.2 for details on the distribution). For the discrete case, accelerated SGD mimics SGD, which nearly matches the minimax risk (when it becomes well defined). For the Gaussian case, accelerated SGD significantly improves upon SGD.

$$\mathbb{E}\left[P(\widehat{\mathbf{x}}_n)\right] - P(\mathbf{x}^*) \le \exp(-n/\kappa) \cdot \big(P(\mathbf{x}_0) - P(\mathbf{x}^*)\big) + 2\sigma^2 d/n, \tag{4}$$

where κ is the condition number of the distribution, which can be upper bounded as $L/\lambda_{\min}(\mathbf{H})$, assuming that $\|\mathbf{a}\| \le L$ with probability one (refer to section 2 for a precise definition of κ). Under appropriate assumptions, these are the best known rates under the stochastic first order oracle model (see section 1.4 for further discussion). A natural implication of the bound implied by averaged SGD is that with $\widetilde{\mathcal{O}}(\kappa)$ oracle calls (Jain et al., 2016) (where, $\widetilde{\mathcal{O}}(\cdot)$ hides $\log$ factors in d, κ), the excess risk attains (up to constants) the (asymptotic) minimax statistical rate. Note that the excess risk bounds in stochastic approximation consist of two terms: (a) *bias*: which represents the dependence of the generalization error on the initial excess risk $P(\mathbf{x}_0) - P(\mathbf{x}^*)$, and (b) the *variance:* which represents the dependence of the generalization error on the noise level σ^2 in the problem.

A precise question regarding accelerating stochastic approximation is: "is it possible to improve the rate of decay of the bias term, while retaining (up to constants) the statistical minimax rate?" The key technical challenge in answering this question is in sharply characterizing the error accumulation of fast gradient methods in the stochastic approximation setting. Common folklore and prior work suggest otherwise: several efforts have attempted to quantify instabilities in the face of statistical or non-statistical errors (Paige, 1971; Proakis, 1974; Polyak, 1987; Greenbaum, 1989; Roy and Shynk, 1990; Sharma et al., 1998; d'Aspremont, 2008; Devolder et al., 2013, 2014; Yuan et al., 2016). Refer to section 1.4 for a discussion on robustness of acceleration to error accumulation. Optimistically, as suggested by the gains enjoyed by accelerated methods in the exact first order oracle model, we may hope to replace the $\widetilde{\mathcal{O}}(\kappa)$ oracle calls achieved by averaged SGD to $\widetilde{\mathcal{O}}(\sqrt{\kappa})$. We now provide a counter example, showing that such an improvement is not possible. Consider a (discrete) distribution $\mathcal{D}$ where the input $\mathbf{a}$ is the i^{th} standard basis vector with probability p_i, $\forall\ i = 1, 2, ..., d$. The covariance of $\mathbf{a}$ in this case is a diagonal matrix with diagonal entries p_i. The condition number of this distribution is $\kappa = \frac{1}{\min_i p_i}$. In this case, it is impossible to make non-trivial reduction in error by observing fewer than κ samples, since with constant probability, we would not have seen the vector corresponding to the smallest probability.



(a) Discrete Distribution (b) Gaussian Distribution

Figure 3: Comparison of averaged SGD with this paper's accelerated SGD in the absence of noise ($\sigma^2 = 0$) for the Discrete and Gaussian distributions with $d = 50, \kappa \approx 10^5$. Acceleration yields substantial gains over averaged SGD for the Gaussian case, while degenerating to SGD's behavior for the discrete case. See section 1.2 for discussion.

On the other hand, consider a case where the distribution $\mathcal{D}$ is a Gaussian with a large condition number κ. Matrix concentration informs us that (with high probability and irrespective of how large κ is) after observing $n = \mathcal{O}(d)$ samples, the empirical covariance matrix will be a spectral approximation to the true covariance matrix, i.e. for some constant $c > 1$, $\mathbf{H}/c \preceq \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}_i\mathbf{a}_i^\top \preceq c\mathbf{H}$. Here, we may hope to achieve a faster convergence rate, as information theoretically $\mathcal{O}(d)$ samples suffice to obtain a non-trivial statistical estimate (see Hsu et al. (2014) for further discussion).

Figure 1 shows the behavior of SGD in these cases; both are synthetic examples in $50-$dimensions, with a condition number $\kappa \approx 10^5$ and noise level $\sigma^2 = 100$. See the figure caption for more details.

These examples suggest that if acceleration is indeed possible, then the degree of improvement (say, over averaged SGD) must depend on distributional quantities that go beyond the condition number κ. A natural conjecture is that this improvement must depend on the number of samples required to spectrally approximate the covariance matrix of the distribution; below this sample size it is not possible to obtain any non-trivial statistical estimate due to information theoretic reasons. This sample size is quantified by a notion which we refer to as the *statistical condition number* $\widetilde{\kappa}$ (see section 2 for a precise definition and for further discussion about $\widetilde{\kappa}$). As we will see in section 2, we have $\widetilde{\kappa} \le \kappa$, $\widetilde{\kappa}$ is affine invariant, unlike κ (i.e. $\widetilde{\kappa}$ is invariant to linear transformations over $\mathbf{a}$).

1.3. Contributions

This paper introduces an accelerated stochastic gradient descent scheme, which can be viewed as a stochastic variant of Nesterov's accelerated gradient method (Nesterov, 2012). As pointed out in Section 1.2, the excess risk of this algorithm can be decomposed into two parts namely, *bias* and *variance*. For the stochastic approximation problem of least squares regression, this paper establishes bias contraction at a geometric rate of $\mathcal{O}(1/\sqrt{\kappa\widetilde{\kappa}})$, improving over prior results (Frostig et al., 2015b; Jain et al., 2016),which prove a geometric rate of $\mathcal{O}(1/\kappa)$, while retaining statistical minimax rates (up to constants) for the variance. Here κ is the condition number and $\widetilde{\kappa}$ is the statistical condition number of the distribution, and a rate of $\mathcal{O}(1/\sqrt{\kappa\widetilde{\kappa}})$ is an improvement over $\mathcal{O}(1/\kappa)$ since $\widetilde{\kappa} \le \kappa$ (see Section 2 for definitions and a short proof of $\widetilde{\kappa} \le \kappa$).

See Table 1 for a theoretical comparison. Figure 2 provides an empirical comparison of the proposed (tail-averaged) accelerated algorithm to (tail-averaged) SGD (Jain et al., 2016) on our two running examples. Our result gives improvement over SGD even in the noiseless (i.e. realizable) case where $\sigma = 0$; this case is equivalent to the setting where we have a distribution over a (possibly infinite) set of consistent linear equations. See Figure 3 for a comparison on the case where $\sigma = 0$.

On a more technical note, this paper introduces two new techniques in order to analyze the proposed accelerated stochastic gradient method: (a) the paper introduces a new potential function in order to show faster rates of decaying the bias, and (b) the paper provides a sharp understanding of the behavior of the proposed accelerated stochastic gradient descent updates as a stochastic process and utilizes this in providing a near-exact estimate of the covariance of its iterates. This viewpoint is critical in order to prove that the algorithm achieves the statistical minimax rate.

We use the operator viewpoint for analyzing stochastic gradient methods, introduced in Défossez and Bach (2015). This viewpoint was also used in Dieuleveut and Bach (2015); Jain et al. (2016).

1.4. Related Work

Non-asymptotic Stochastic Approximation: Stochastic gradient descent (SGD) and its variants are by far the most widely studied algorithms for the stochastic approximation problem. While initial works (Robbins and Monro, 1951) considered the final iterate of SGD, later works (Ruppert, 1988; Polyak and Juditsky, 1992) demonstrated that averaged SGD obtains statistically optimal estimation rates. Several works provide non-asymptotic analyses for averaged SGD and variants (Bach and Moulines, 2011; Bach, 2014; Frostig et al., 2015b) for various stochastic approximation problems. For stochastic approximation with least squares regression Bach and Moulines (2013); Défossez and Bach (2015); Needell et al. (2016); Frostig et al. (2015b); Jain et al. (2016) provide non-asymptotic analysis of the behavior of SGD and its variants. Défossez and Bach (2015); Dieuleveut and Bach (2015) provide non-asymptotic results which achieve the minimax rate on the variance (where the bias is lower order, not geometric). Needell et al. (2016) achieves a geometric rate on the bias (and where the variance is not minimax). Frostig et al. (2015b); Jain et al. (2016) obtain both the minimax rate on the variance and a geometric rate on the bias, as seen in equation 4.

Acceleration and Noise Stability: While there have been several attempts at understanding if it is possible to accelerate SGD , the results have been largely negative. With regards to acceleration with adversarial (non-statistical) errors in the exact first order oracle model, d'Aspremont (2008) provide negative results and Devolder et al. (2013, 2014) provide lower bounds showing that fast gradient methods do not improve upon standard gradient methods. There is also a series of works considering statistical errors. Polyak (1987) suggests that the relative merits of heavy ball (HB) method (Polyak, 1964) in the noiseless case vanish with noise unless strong assumptions on the noise model are considered; an instance of this is when the noise variance decays as the iterates approach the minimizer. The Conjugate Gradient (CG) method (Hestenes and Stiefel, 1952) is suggested to face similar robustness issues in the face of statistical errors (Polyak, 1987); this is in addition to the issues that CG is known to suffer from owing to roundoff errors (due to finite precision arithmetic) (Paige, 1971; Greenbaum, 1989). In the signal processing literature, where SGD goes by Least Mean Squares (LMS) (Widrow and Stearns, 1985), there have been efforts that date to several decades (Proakis, 1974; Roy and Shynk, 1990; Sharma et al., 1998) which study accelerated LMS methods (stochastic variants of CG/HB) in the same oracle model as the one considered by this paper (equation 2). These efforts consider the final iterate (i.e. no iterate averaging) of accelerated LMS methods with a fixed step-size and conclude that while it allows for a faster decay of the initial error (bias) (which is unquantified), their steady state behavior (i.e. variance) is worse compared to that of LMS. Yuan et al. (2016) considered a constant step size accelerated scheme with no iterate averaging in the same oracle model as this paper, and conclude that these do not offer any improvement over standard SGD. More concretely, Yuan et al. (2016) show that the variance of their accelerated SGD method with a sufficiently small constant step size is the same as that of SGD with a significantly larger step size. Note that none of the these efforts (Proakis, 1974; Roy and Shynk, 1990; Sharma et al., 1998; Yuan et al., 2016) achieve minimax error rates or quantify (any improvement whatsoever on the) rate of bias decay.

Oracle models and optimality: With regards to notions of optimality, there are (at least) two lines of thought: one is a statistical objective where the goal is (on every problem instance) to match the rate of the statistically optimal estimator (Anbar, 1971; Fabian, 1973; Kushner and Clark, 1978; Polyak and Juditsky, 1992); another is on obtaining algorithms whose worst case upper bounds (under various assumptions such as bounded noise) match the lower bounds provided in Nemirovsky and Yudin (1983). The work of Polyak and Juditsky (1992) are in the former model, where they show that the distribution of the averaged SGD estimator matches, on *every* problem, that of the statistically optimal estimator, in the limit (under appropriate regularization conditions standard in the statistics literature, where the optimal estimator is often referred to as the maximum likelihood estimator/the empirical risk minimizer/an M-estimator (Lehmann and Casella, 1998; van der Vaart, 2000)). Along these lines, non-asymptotic rates towards statistically optimal estimators are given by Bach and Moulines (2013); Bach (2014); Défossez and Bach (2015); Dieuleveut and Bach (2015); Needell et al. (2016); Frostig et al. (2015b); Jain et al. (2016). This work can be seen as improving this non-asymptotic rate (to the statistically optimal estimation rate) using an accelerated method. As to the latter (i.e. matching the worst-case lower bounds in Nemirovsky and Yudin (1983)), there are a number of positive results on using accelerated stochastic optimization procedures; the works of Lan (2008); Hu et al. (2009); Ghadimi and Lan (2012, 2013); Dieuleveut et al. (2016) match the lower bounds provided in Nemirovsky and Yudin (1983). We compare these assumptions and works in more detail.

In stochastic first order oracle models (see Kushner and Clark (1978); Kushner and Yin (2003)), one typically has access to sampled gradients of the form:

$$\widehat{\nabla} P(\mathbf{x}) = \nabla P(\mathbf{x}) + \boldsymbol{\eta}, \tag{5}$$

where varying assumptions are made on the noise $\boldsymbol{\eta}$. The worst-case lower bounds in Nemirovsky and Yudin (1983) are based on that $\boldsymbol{\eta}$ is bounded; the accelerated methods in Lan (2008); Hu et al. (2009); Ghadimi and Lan (2012, 2013); Dieuleveut et al. (2016) which match these lower bounds in various cases, all assume either bounded noise or, at least $\mathbb{E}\left[\|\boldsymbol{\eta}\|^2\right]$ is finite. In the least squares setting (such as the one often considered in practice and also considered in Polyak and Juditsky (1992); Bach and Moulines (2013); Défossez and Bach (2015); Dieuleveut and Bach (2015); Frostig et al. (2015b); Jain et al. (2016)), this assumption does not hold, since $\mathbb{E}\left[\|\boldsymbol{\eta}\|^2\right]$ is not bounded. To see this, $\boldsymbol{\eta}$ in our oracle model (equation 2) is:

$$\boldsymbol{\eta} = \widehat{\nabla} P(\mathbf{x}) - \nabla P(\mathbf{x}) = (\mathbf{a}\mathbf{a}^\top - \mathbf{H})(\mathbf{x} - \mathbf{x}^*) - \epsilon \cdot \mathbf{a} \tag{6}$$

which implies that $\mathbb{E}\left[\|\boldsymbol{\eta}\|^2\right]$ is not uniformly bounded (unless additional assumptions are enforced to ensure that the algorithm's iterates $\mathbf{x}$ lie within a compact set). Hence, the assumptions made in Hu et al. (2009); Ghadimi and Lan (2012, 2013); Dieuleveut et al. (2016) do not permit one to obtain finite n-sample bounds on the excess risk. Suppose we consider the case of $\epsilon = 0$, i.e. where the additive noise is zero and $b = \mathbf{a}^\top \mathbf{x}^*$. For this case, this paper provides a geometric rate of convergence to the minimizer $\mathbf{x}^*$, while the results of Ghadimi and Lan (2012, 2013); Dieuleveut et al. (2016) at best indicate a $\mathcal{O}(1/n)$ rate. Finally, in contrast to all other existing work, our result is the first to provide finer distribution dependent characteristics of the improvements offered by accelerating SGD (e.g. refer to the Gaussian and discrete examples in section 1.2).

Acceleration and Finite Sums: As a final remark, there have been results (Shalev-Shwartz and Zhang, 2014; Frostig et al., 2015a; Lin et al., 2015; Lan and Zhou, 2015; Allen-Zhu, 2016) that provide accelerated rates for *offline* stochastic optimization which deal with minimizing sums of convex functions; these results are almost tight due to matching lower bounds (Lan and Zhou, 2015; Woodworth and Srebro, 2016). These results do not immediately translate into rates on the generalization error. Furthermore, these algorithms are not streaming, as they require making multiple passes over a dataset stored in memory. Refer to Frostig et al. (2015b) for more details.

2. Main Results

We now provide our assumptions and main result, before which, we have some notation. For a vector $\mathbf{x} \in \mathbb{R}^d$ and a positive semi-definite matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$ (i.e. $\mathbf{S} \succeq 0$), denote $\|\mathbf{x}\|_{\mathbf{S}}^2 \stackrel{\text{def}}{=} \mathbf{x}^\top \mathbf{S}\mathbf{x}$.

2.1. Assumptions and Definitions

Let $\mathbf{H}$ denote the second moment matrix of the input, which is also the hessian $\nabla^2 P(\mathbf{x})$ of (1):

$$\mathbf{H} \stackrel{\text{def}}{=} \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[\mathbf{a}\otimes\mathbf{a}\right] = \nabla^2 P(\mathbf{x}).$$

($\mathcal{A}$**1**) **Finite second and fourth moment:** The second moment matrix $\mathbf{H}$ and the fourth moment tensor $\mathcal{M}$ $(= \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[\mathbf{a}\otimes\mathbf{a}\otimes\mathbf{a}\otimes\mathbf{a}\right])$ of the input $\mathbf{a}\sim\mathcal{D}$ exist and are finite.

($\mathcal{A}$**2**) **Positive Definiteness**: The second moment matrix $\mathbf{H}$ is strictly positive definite, i.e. $\mathbf{H} \succ 0$.

We assume ($\mathcal{A}$**1**) and ($\mathcal{A}$**2**). ($\mathcal{A}$**2**) implies that $P(\mathbf{x})$ is *strongly convex* and admits a unique minimizer $\mathbf{x}^*$. Denote the noise ϵ in a sample $(\mathbf{a}, b) \sim \mathcal{D}$ as: $\epsilon \stackrel{\text{def}}{=} b - \langle \mathbf{a}, \mathbf{x}^*\rangle$. First order optimality conditions of $\mathbf{x}^*$ imply $\nabla P(\mathbf{x}^*) = \mathbb{E}\left[\epsilon\cdot\mathbf{a}\right] = 0$. Let $\mathbf{\Sigma}$ denote the covariance of gradient at optimum $\mathbf{x}^*$ (or *noise covariance matrix*), $\mathbf{\Sigma} \stackrel{\text{def}}{=} \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[\widehat{\nabla} P(\mathbf{x}^*)\otimes\widehat{\nabla} P(\mathbf{x}^*)\right] = \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[\epsilon^2\cdot\mathbf{a}\otimes\mathbf{a}\right]$.
We define the *noise level* σ^2, *condition number* κ, *statistical condition number* $\widetilde{\kappa}$ below.
Noise level: The *noise level* is defined to be the smallest positive number σ^2 such that $\mathbf{\Sigma} \preceq \sigma^2\mathbf{H}$.The noise level σ^2 quantifies the amount of noise in the stochastic gradient oracle and has been utilized in previous work (e.g., see Bach and Moulines (2011, 2013)) in providing non-asymptotic bounds for the stochastic approximation problem. In the *homoscedastic* (additive noise) case, where ϵ is independent of the input $\mathbf{a}$, this condition is satisfied with equality, i.e. $\mathbf{\Sigma} = \sigma^2\,\mathbf{H}$ with $\sigma^2 = \mathbb{E}\left[\epsilon^2\right]$.
Condition number: Let $\mu \stackrel{\text{def}}{=} \lambda_{\text{min}}(\mathbf{H})$. $\mu > 0$ by ($\mathcal{A}$**2**). Now, let R^2 be the smallest positive number such that $\mathbb{E}\left[\|\mathbf{a}\|^2\,\mathbf{a}\mathbf{a}^\top\right] \preceq R^2\,\mathbf{H}$. The *condition number* κ of the distribution $\mathcal{D}$ (Défossez and Bach, 2015; Jain et al., 2016) is $\kappa \stackrel{\text{def}}{=} R^2/\mu$.
Statistical condition number: The *statistical condition number* $\widetilde{\kappa}$ is defined as the smallest positive number such that $\mathbb{E}\left[\|\mathbf{a}\|_{\mathbf{H}^{-1}}^2\,\mathbf{a}\mathbf{a}^\top\right] \preceq \widetilde{\kappa}\,\mathbf{H}$.

Algorithm 1 (Tail-Averaged) **A**ccelerated **S**tochastic **G**radient **D**escent (ASGD)

Input: n oracle calls to 2, initial iterate $\mathbf{x}_0 = \mathbf{v}_0$, Unaveraged (burn-in) phase t, Step sizes $\alpha, \beta, \gamma, \delta$

Output: $\bar{\mathbf{x}}_{t,n} \leftarrow \frac{1}{n-t}\sum_{j=t+1}^{n} \mathbf{x}_j$

for $j \leftarrow 1$ **to** n **do**

$\quad \mathbf{y}_{j-1} \leftarrow \alpha \mathbf{x}_{j-1} + (1-\alpha)\mathbf{v}_{j-1}$

$\quad \mathbf{x}_j \leftarrow \mathbf{y}_{j-1} - \delta \widehat{\nabla} P(\mathbf{y}_{j-1})$

$\quad \mathbf{z}_{j-1} \leftarrow \beta \mathbf{y}_{j-1} + (1-\beta)\mathbf{v}_{j-1}$

$\quad \mathbf{v}_j \leftarrow \mathbf{z}_{j-1} - \gamma \widehat{\nabla} P(\mathbf{y}_{j-1})$

end

Remarks on $\widetilde{\kappa}$ and κ: Unlike κ, it is straightforward to see that $\widetilde{\kappa}$ is affine invariant (i.e. unchanged with linear transformations over $\mathbf{a}$). Since $\mathbb{E}\left[\|\mathbf{a}\|^2_{\mathbf{H}^{-1}}\, \mathbf{a}\mathbf{a}^\top\right] \preceq \frac{1}{\mu}\mathbb{E}\left[\|\mathbf{a}\|_2^2\, \mathbf{a}\mathbf{a}^\top\right] \preceq \kappa \mathbf{H}$, we note $\widetilde{\kappa} \leq \kappa$. For the discrete case (from Section 1.2), it is straightforward to see that both κ and $\widetilde{\kappa}$ are equal to $1/\min_i p_i$. In contrast, for the Gaussian case (from Section 1.2), $\widetilde{\kappa}$ is $\mathcal{O}(d)$, while κ is $\mathcal{O}(\text{Trace}(\mathbf{H})/\mu)$ which may be arbitrarily large (based on choice of the coordinate system).

$\widetilde{\kappa}$ governs how many samples $\mathbf{a}_i$ require to be drawn from $\mathcal{D}$ so that the empirical covariance is spectrally close to $\mathbf{H}$, i.e. for some constant $c > 1$, $\mathbf{H}/c \preceq \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}_i\mathbf{a}_i^\top \preceq c\mathbf{H}$. In comparison to the matrix Bernstein inequality where stronger (yet related) moment conditions are assumed in order to obtain high probability results, our results hold only in expectation (refer to Hsu et al. (2014) for this definition, wherein $\widetilde{\kappa}$ is referred to as bounded statistical leverage in theorem 1 and remark 1).

2.2. Algorithm and Main Theorem

Algorithm 1 presents the pseudo code of the proposed algorithm. ASGD can be viewed as a variant of Nesterov's accelerated gradient method (Nesterov, 2012), working with a stochastic gradient oracle (equation 2) and with tail-averaging the final $n-t$ iterates. The main result now follows:

Theorem 1 *Suppose* ($\mathcal{A}$**1**) *and* ($\mathcal{A}$**2**) *hold. Set* $\alpha = \frac{3\sqrt{5}\cdot\sqrt{\kappa\widetilde{\kappa}}}{1+3\sqrt{5}\cdot\sqrt{\kappa\widetilde{\kappa}}}, \beta = \frac{1}{9\sqrt{\kappa\widetilde{\kappa}}}, \gamma = \frac{1}{3\sqrt{5}\cdot\mu\sqrt{\kappa\widetilde{\kappa}}}, \delta = \frac{1}{5R^2}$. *After* n *calls to the stochastic first order oracle (equation 2), ASGD outputs* $\bar{\mathbf{x}}_{t,n}$ *satisfying:*

$$\mathbb{E}\left[P(\bar{\mathbf{x}}_{t,n})\right] - P(\mathbf{x}^*) \leq \underbrace{C\cdot\frac{(\kappa\widetilde{\kappa})^{9/4} d\kappa}{(n-t)^2}\cdot\exp\left(\frac{-t}{9\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\big(P(\mathbf{x}_0)-P(\mathbf{x}^*)\big)}_{\text{Leading order bias error}} + \underbrace{5\frac{\sigma^2 d}{n-t}}_{\text{Leading order variance error}} +$$

$$\underbrace{C\cdot(\kappa\widetilde{\kappa})^{5/4} d\kappa\cdot\exp\left(\frac{-n}{9\sqrt{\kappa\widetilde{\kappa}}}\right)\big(P(\mathbf{x}_0)-P(\mathbf{x}^*)\big)}_{\text{Exponentially vanishing lower order bias term}} + \underbrace{C\cdot\frac{\sigma^2 d}{(n-t)^2}\sqrt{\kappa\widetilde{\kappa}}}_{\text{Lower order variance error term}} +$$

$$\underbrace{C\cdot\exp\left(-\frac{n}{9\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\left(\sigma^2 d\cdot(\kappa\widetilde{\kappa})^{7/4} + \frac{\sigma^2 d}{(n-t)^2}\cdot(\kappa\widetilde{\kappa})^{7/2}\widetilde{\kappa}\right) + C\cdot\frac{\sigma^2 d}{n-t}(\kappa\widetilde{\kappa})^{11/4}\exp\left(-\frac{(n-t-1)}{30\sqrt{\kappa\widetilde{\kappa}}}\right)}_{\text{Exponentially vanishing lower order variance error terms}},$$

where C *is a universal constant,* σ^2, κ *and* $\widetilde{\kappa}$ *are the noise level, condition number and statistical condition number respectively.*

The following corollary holds if the iterates are tail-averaged over the last $n/2$ samples and $n > \mathcal{O}(\sqrt{\kappa\widetilde{\kappa}}\log(d\kappa\widetilde{\kappa}))$. The second condition lets us absorb lower order terms into leading order terms.

Corollary 2 *Assume the parameter settings of theorem 1 and let* $t = \lfloor n/2 \rfloor$ *and* $n > C'\sqrt{\kappa\widetilde{\kappa}}\log(d\kappa\widetilde{\kappa})$ *(for an appropriate universal constants* C, C'*). We have that with* n *calls to the stochastic first order oracle, ASGD outputs a vector* $\bar{\mathbf{x}}_{t,n}$ *satisfying:*

$$\mathbb{E}\left[P(\bar{\mathbf{x}}_{t,n})\right] - P(\mathbf{x}^*) \leq C \cdot \exp\left(-\frac{n}{20\sqrt{\kappa\widetilde{\kappa}}}\right) \cdot \big(P(\mathbf{x}_0) - P(\mathbf{x}^*)\big) + 11\frac{\sigma^2 d}{n}.$$

A few remarks about the result of theorem 1 are due: (i) ASGD decays the initial error at a geometric rate of $\mathcal{O}(1/\sqrt{\kappa\widetilde{\kappa}})$ during the unaveraged phase of t iterations, which presents the first improvement over the $\mathcal{O}(1/\kappa)$ rate offered by SGD (Robbins and Monro, 1951)/averaged SGD (Polyak and Juditsky, 1992; Jain et al., 2016) for the least squares stochastic approximation problem, (ii) the second term in the error bound indicates that ASGD obtains (up to constants) the minimax rate once $n > \mathcal{O}(\sqrt{\kappa\widetilde{\kappa}}\log(d\kappa\widetilde{\kappa}))$. Note that this implies that Theorem 1 provides a sharp non-asymptotic analysis (up to $\log$ factors) of the behavior of Algorithm 1.

2.3. Discussion and Open Problems

A challenging problem in this context is in formalizing a finite sample size lower bound in the oracle model considered in this work. Lower bounds in stochastic oracle models have been considered in the literature (see Nemirovsky and Yudin (1983); Raginsky and Rakhlin (2011); Agarwal et al. (2012)), though it is not evident these oracle models and lower bounds are sharp enough to imply statements in our setting (see section 1.4 for a discussion of these oracle models).

Let us now understand theorem 1 in the broader context of stochastic approximation. Under certain regularity conditions, it is known that (Lehmann and Casella, 1998; van der Vaart, 2000) that the rate described in equation 3 for the homoscedastic case holds for a broader set of misspecified models (i.e., heteroscedastic noise case), with an appropriate definition of the noise variance. By defining $\sigma^2_{\text{ERM}} \stackrel{\text{def}}{=} \mathbb{E}\left[\left\|\widehat{\nabla}P(\mathbf{x}^*)\right\|^2_{\mathbf{H}^{-1}}\right]$, the rate of the ERM is guaranteed to approach σ^2_{ERM}/n (Lehmann and Casella, 1998; van der Vaart, 2000) in the limit of large n, i.e.:

$$\lim_{n\to\infty} \frac{\mathbb{E}_{\mathcal{S}_n}[P_n(\widehat{\mathbf{x}}_n^{\text{ERM}})] - P(\mathbf{x}^*)}{\sigma^2_{\text{ERM}}/n} = 1, \tag{7}$$

where $\widehat{\mathbf{x}}_n^{\text{ERM}}$ is the ERM over samples $\mathcal{S}_n = \{\mathbf{a}_i, b_i\}_{i=1}^n$. Averaged SGD (Jain et al., 2016) and streaming SVRG (Frostig et al., 2015b) are known to achieve these rates for the heteroscedastic case. Neglecting constants, Theorem 1 is guaranteed to achieve the rate of the ERM for the *homoscedastic* case (where $\mathbf{\Sigma} = \sigma^2\mathbf{H}$) and is tight when the bound $\mathbf{\Sigma} \preceq \sigma^2\mathbf{H}$ is nearly tight (upto constants). We conjecture ASGD achieves the rate of the ERM in the heteroscedastic case by appealing to a more refined analysis as is the case for averaged SGD (see Jain et al. (2016)). It is also an open question to understand acceleration for smooth stochastic approximation (beyond least squares), in situations where the rate represented by equation 7 holds (Polyak and Juditsky, 1992).

3. Proof Outline

Recall the variables in Algorithm 1. We begin by defining the centered estimate $\boldsymbol{\theta}_j$ as:

$$\boldsymbol{\theta}_j \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{x}_j - \mathbf{x}^* \\ \mathbf{y}_j - \mathbf{x}^* \end{bmatrix} \in \mathbb{R}^{2d}.$$

Recall that the stepsizes in Algorithm 1 are $\alpha = \frac{3\sqrt{5}\cdot\sqrt{\kappa\widetilde{\kappa}}}{1+3\sqrt{5}\cdot\sqrt{\kappa\widetilde{\kappa}}}, \beta = \frac{1}{9\sqrt{\kappa\widetilde{\kappa}}}, \gamma = \frac{1}{3\sqrt{5}\cdot\mu\sqrt{\kappa\widetilde{\kappa}}}, \delta = \frac{1}{5R^2}$. The accelerated SGD updates of Algorithm 1 can be written in terms of $\boldsymbol{\theta}_j$ as:

$$\boldsymbol{\theta}_j = \widehat{\mathbf{A}}_j \boldsymbol{\theta}_{j-1} + \boldsymbol{\zeta}_j, \quad \text{where},$$
$$\widehat{\mathbf{A}}_j \stackrel{\text{def}}{=} \begin{bmatrix} 0 & (\mathbf{I} - \delta \mathbf{a}_j \mathbf{a}_j^\top) \\ -\alpha(1-\beta)\,\mathbf{I} & (1+\alpha(1-\beta))\mathbf{I} - (\alpha\delta + (1-\alpha)\gamma)\mathbf{a}_j\mathbf{a}_j^\top \end{bmatrix}, \boldsymbol{\zeta}_j \stackrel{\text{def}}{=} \begin{bmatrix} \delta \cdot \epsilon_j \mathbf{a}_j \\ (\alpha\delta + (1-\alpha)\gamma) \cdot \epsilon_j \mathbf{a}_j \end{bmatrix},$$

where $\epsilon_j = b_j - \langle \mathbf{a}_j, \mathbf{x}^* \rangle$. The tail-averaged iterate $\bar{\mathbf{x}}_{t,n}$ is associated with $\bar{\boldsymbol{\theta}}_{t,n} \stackrel{\text{def}}{=} \frac{1}{n-t}\sum_{j=t+1}^{n} \boldsymbol{\theta}_j$. Let $\mathbf{A} \stackrel{\text{def}}{=} \mathbb{E}\left[\widehat{\mathbf{A}}_j | \mathcal{F}_{j-1}\right]$, where $\mathcal{F}_{j-1}$ is a filtration generated by $(\mathbf{a}_1, b_1), \cdots, (\mathbf{a}_{j-1}, b_{j-1})$. Let $\mathcal{B}, \mathcal{A}_{\mathcal{L}}, \mathcal{A}_{\mathcal{R}}$ be linear operators acting on a matrix $\mathbf{S} \in \mathbb{R}^{2d\times 2d}$ so that $\mathcal{B}\mathbf{S} \stackrel{\text{def}}{=} \mathbb{E}\left[\widehat{\mathbf{A}}_j \mathbf{S} \widehat{\mathbf{A}}_j^\top | \mathcal{F}_{j-1}\right]$, $\mathcal{A}_{\mathcal{L}}\mathbf{S} \stackrel{\text{def}}{=} \mathbf{A}\mathbf{S}$, $\mathcal{A}_{\mathcal{R}}\mathbf{S} \stackrel{\text{def}}{=} \mathbf{S}\mathbf{A}$. Denote $\widehat{\boldsymbol{\Sigma}} \stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\zeta}_j \boldsymbol{\zeta}_j^\top | \mathcal{F}_{j-1}\right]$ and matrices $\mathbf{G}, \mathbf{Z}, \widetilde{\mathbf{G}}$ as:

$$\mathbf{G} \stackrel{\text{def}}{=} \widetilde{\mathbf{G}}^\top \mathbf{Z} \widetilde{\mathbf{G}}, \text{where}, \widetilde{\mathbf{G}} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I} & 0 \\ \frac{-\alpha}{1-\alpha}\mathbf{I} & \frac{1}{1-\alpha}\mathbf{I} \end{bmatrix}, \quad \mathbf{Z} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mu \mathbf{H}^{-1} \end{bmatrix}.$$

Bias-variance decomposition: The proof of theorem 1 employs the *bias-variance* decomposition, which is well known in the context of stochastic approximation (see Bach and Moulines (2011); Frostig et al. (2015b); Jain et al. (2016)) and is re-derived in the appendix. The bias-variance decomposition allows for the generalization error to be upper-bounded by analyzing two sub-problems: (a) *bias*, analyzing the algorithm's behavior on the *noiseless* problem (i.e. $\boldsymbol{\zeta}_j = 0 \;\forall\; j$ a.s.) while starting at $\boldsymbol{\theta}_0^{\text{bias}} = \boldsymbol{\theta}_0$ and (b) *variance*, analyzing the algorithm's behavior by starting at the solution (i.e. $\boldsymbol{\theta}_0^{\text{variance}} = 0$) and allowing the noise $\boldsymbol{\zeta}_\cdot$ to drive the process. In a similar manner as $\bar{\boldsymbol{\theta}}_{t,n}$, the bias and variance sub-problems are associated with $\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}$ and $\bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}}$, and these are related as:

$$\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n} \otimes \bar{\boldsymbol{\theta}}_{t,n}\right] \preceq 2 \cdot \left(\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}\right] + \mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}}\right] \right). \tag{8}$$

Since we deal with the square loss, the generalization error of the output $\bar{\mathbf{x}}_{t,n}$ of algorithm 1 is:

$$\mathbb{E}\left[P(\bar{\mathbf{x}}_{t,n})\right] - P(\mathbf{x}^*) = \frac{1}{2} \cdot \left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n} \otimes \bar{\boldsymbol{\theta}}_{t,n}\right] \right\rangle, \tag{9}$$

indicating that the generalization error can be bounded by analyzing the bias and variance subproblem. We now present the lemmas that bound the bias error.

Lemma 3 *The covariance* $\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{bias} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{bias}\right]$ *of the bias part of averaged iterate* $\bar{\boldsymbol{\theta}}_{t,n}^{bias}$ *satisfies:*

$$\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{bias} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{bias}\right] = \frac{1}{(n-t)^2}\left(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathcal{A}_{\mathcal{R}}^\top\right)(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\left(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0\right)$$
$$- \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left((\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}(\mathcal{A}_{\mathcal{R}}^\top)^{n+1-j}\right)\mathcal{B}^j(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0).$$

The quantity that needs to be bounded in the term above is $\mathcal{B}^{t+1}\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0$. Lemma 4 presents a result that can be applied recursively to bound $\mathcal{B}^{t+1}\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0$ ($= \mathcal{B}^{t+1}\boldsymbol{\theta}_0^{\text{bias}} \otimes \boldsymbol{\theta}_0^{\text{bias}}$ since $\boldsymbol{\theta}_0^{\text{bias}} = \boldsymbol{\theta}_0$).

Lemma 4 (Bias contraction) *For any two vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, *let* $\boldsymbol{\theta} \stackrel{def}{=} \begin{bmatrix} \mathbf{x} - \mathbf{x}^* \\ \mathbf{y} - \mathbf{x}^* \end{bmatrix} \in \mathbb{R}^{2d}$. *We have:*

$$\left\langle \mathbf{G}, \mathcal{B}\left(\boldsymbol{\theta}\boldsymbol{\theta}^\top\right) \right\rangle \leq \left(1 - \frac{1}{9\sqrt{\kappa\widetilde{\kappa}}}\right)\left\langle \mathbf{G}, \boldsymbol{\theta}\boldsymbol{\theta}^\top \right\rangle$$

Remarks: (i) the matrices $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{G}}^\top$ appearing in $\mathbf{G}$ are due to the fact that we prove contraction using the variables $\mathbf{x}-\mathbf{x}^*$ and $\mathbf{v}-\mathbf{x}^*$ instead of $\mathbf{x}-\mathbf{x}^*$ and $\mathbf{y}-\mathbf{x}^*$, as used in defining $\boldsymbol{\theta}$. (ii) The key novelty in lemma 4 is that while standard analyses of accelerated gradient descent (in the exact first order oracle) use the potential function $\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{H}}^2 + \mu\,\|\mathbf{v}-\mathbf{x}^*\|_2^2$ (e.g. Wilson et al. (2016)), we consider it crucial for employing the potential function $\|\mathbf{x}-\mathbf{x}^*\|_2^2 + \mu\,\|\mathbf{v}-\mathbf{x}^*\|_{\mathbf{H}^{-1}}^2$ (this potential function is captured using the matrix $\mathbf{Z}$) to prove accelerated rates (of $\mathcal{O}\left(1/\sqrt{\kappa\widetilde{\kappa}}\right)$) for bias decay.

We now present the lemmas associated with bounding the variance error:

Lemma 5 *The covariance $\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{variance}\otimes\bar{\boldsymbol{\theta}}_{t,n}^{variance}\right]$ of the variance error $\bar{\boldsymbol{\theta}}_{t,n}^{variance}$ satisfies:*

$$\begin{aligned}\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{variance}\otimes\bar{\boldsymbol{\theta}}_{t,n}^{variance}\right] &= \frac{1}{n-t}\big(\mathcal{I}+(\mathcal{I}-\mathcal{A}_\mathcal{L})^{-1}\mathcal{A}_\mathcal{L}+(\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-1}\mathcal{A}_\mathcal{R}^\top\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}}\\ &-\frac{1}{(n-t)^2}\big((\mathcal{I}-\mathcal{A}_\mathcal{L})^{-2}(\mathcal{A}_\mathcal{L}-\mathcal{A}_\mathcal{L}^{n+1-t})+(\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-2}(\mathcal{A}_\mathcal{R}^\top-(\mathcal{A}_\mathcal{R}^\top)^{n+1-t})\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}}\\ &-\frac{1}{(n-t)^2}\big(\mathcal{I}+(\mathcal{I}-\mathcal{A}_\mathcal{L})^{-1}\mathcal{A}_\mathcal{L}+(\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-1}\mathcal{A}_\mathcal{R}^\top\big)(\mathcal{I}-\mathcal{B})^{-2}(\mathcal{B}^{t+1}-\mathcal{B}^{n+1})\widehat{\boldsymbol{\Sigma}}\\ &+\frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\big((\mathcal{I}-\mathcal{A}_\mathcal{L})^{-1}\mathcal{A}_\mathcal{L}^{n+1-j}+(\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-1}(\mathcal{A}_\mathcal{R}^\top)^{n+1-j}\big)(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^j\widehat{\boldsymbol{\Sigma}}.\end{aligned}$$

The covariance of the stationary distribution of the iterates i.e., $\lim_{j\to\infty}\boldsymbol{\theta}_j^{\text{variance}}$ requires a precise bound to obtain statistically optimal error rates. Lemma 6 presents a bound on this quantity.

Lemma 6 (Stationary covariance) *The covariance of limiting distribution of $\boldsymbol{\theta}^{variance}$ satisfies:*

$$\mathbb{E}\left[\boldsymbol{\theta}_\infty^{variance}\otimes\boldsymbol{\theta}_\infty^{variance}\right] = (\mathbf{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}} \preceq 5\sigma^2\left((2/3)\cdot\big(\frac{1}{\widetilde{\kappa}}\mathbf{H}^{-1}\big)+(5/6)\cdot(\delta\mathbf{I})\right)\otimes\begin{bmatrix}1&0\\0&1\end{bmatrix}.$$

A crucial implication of this lemma is that the limiting final iterate $\boldsymbol{\theta}_\infty^{\text{variance}}$ has an excess risk $\mathcal{O}(\sigma^2)$. This result naturally lends itself to the (tail-)averaged iterate achieving the minimax optimal rate of $\mathcal{O}(d\sigma^2/n)$. Refer to the appendix E and lemma 17 for more details in this regard.

4. Conclusion

This paper introduces an accelerated stochastic gradient method, which presents the first improvement in achieving minimax rates faster than averaged SGD (Robbins and Monro, 1951; Polyak and Juditsky, 1992; Jain et al., 2016)/Streaming SVRG (Frostig et al., 2015b) for the stochastic approximation problem of least squares regression. To obtain this result, the paper presented the need to rethink what acceleration has to offer when working with a stochastic gradient oracle: the statistical condition number (an affine invariant distributional quantity) is shown to characterize the improvements that acceleration offers in the stochastic first order oracle model. In essence, this paper serves to provide the first provable analysis of the claim that fast gradient methods are stable when dealing with statistical errors, in stark contrast to efforts that date to several decades indicating negative results in various statistical or non-statistical settings.

Acknowledgments

Sham Kakade acknowledges funding from Washington Research Foundation Fund for Innovation in Data-Intensive Discovery and the NSF through awards CCF-1703574 and CCF-1740551.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 2012.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *CoRR*, abs/1603.05953, 2016.

Dan Anbar. *On Optimal Estimation Methods Using Stochastic Approximation Procedures*. University of California, 1971. URL http://books.google.com/books?id=MmpHJwAACAAJ.

Francis R. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research (JMLR)*, volume 15, 2014.

Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NIPS 24*, 2011.

Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *NIPS 26*, 2013.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *NIPS 20*, 2007.

Louis Augustin Cauchy. Méthode générale pour la résolution des systémes d'équations simultanees. *C. R. Acad. Sci. Paris*, 1847.

Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

Alexandre Défossez and Francis R. Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *AISTATS*, volume 38, 2015.

Olivier Devolder, François Glineur, and Yurii E. Nesterov. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Papers 2013016*, 2013.

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

Aymeric Dieuleveut and Francis R. Bach. Non-parametric stochastic approximation with large step sizes. *The Annals of Statistics*, 2015.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis R. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *CoRR*, abs/1602.05419, 2016.

Vaclav Fabian. Asymptotically efficient stochastic approximation; the RM case. *Annals of Statistics*, 1(3), 1973.

Roy Frostig, Rong Ge, Sham Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *ICML*, 2015a.

Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *COLT*, 2015b.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 2012.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 2013.

Anne Greenbaum. Behavior of slightly perturbed lanczos and conjugate-gradient recurrences. *Linear Algebra and its Applications*, 1989.

Magnus R. Hestenes and Eduard Stiefel. Methods of conjuate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 1952.

Daniel J. Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Chonghai Hu, James T. Kwok, and Weike Pan. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS 22*, 2009.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. *CoRR*, abs/1610.03774, 2016.

Harold J. Kushner and Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer-Verlag, 1978.

Harold J. Kushner and George Yin. Stochastic approximation and recursive algorithms and applications. *Springer-Verlag*, 2003.

G. Lan. An optimal method for stochastic composite optimization. *Tech. Report, GaTech.*, 2008.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *CoRR*, abs/1507.02000, 2015.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 1998.

Hongzhou Lin, Julien Mairal, and Zaïd Harchaoui. A universal catalyst for first-order optimization. In *NIPS*, 2015.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 2016.

Arkadii S. Nemirovsky and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley, 1983.

Yurii E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269, 1983.

Yurii E. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

Yurii E. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4, 1964.

Boris T. Polyak. *Introduction to Optimization*. Optimization Software, 1987.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, volume 30, 1992.

John G. Proakis. Channel identification for high speed digital communications. *IEEE Transactions on Automatic Control*, 1974.

Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 2011.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, vol. 22, 1951.

Sumit Roy and John J. Shynk. Analysis of the momentum lms algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1990.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. *Tech. Report, ORIE, Cornell University*, 1988.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *ICML*, 2014.

Rajesh Sharma, William A. Sethares, and James A. Bucklew. Analysis of momentum adaptive filtering algorithms. *IEEE Transactions on Signal Processing*, 1998.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Publishers, 2000.

Bernard Widrow and Samuel D. Stearns. *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A lyapunov analysis of momentum methods in optimization. *CoRR*, abs/1611.02635, 2016.

Blake Woodworth and Nathan Srebro. Tight complexity bounds for optimizing composite objectives. *CoRR*, abs/1605.08003, 2016.

Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research (JMLR)*, volume 17, 2016.

Appendix A. Appendix setup

We will first provide a note on the organization of the appendix and follow that up with introducing the notations.

A.1. Organization

- In subsection A.2, we will recall notation from the main paper and introduce some new notation that will be used across the appendix.
- In section B, we will write out expressions that characterize the generalization error of the proposed accelerated SGD method. In order to bound the generalization error, we require developing an understanding of two terms namely the bias error and the variance error.
- In section C, we prove lemmas that will be used in subsequent sections to prove bounds on the bias and variance error.
- In section D, we will bound the bias error of the proposed accelerated stochastic gradient method. In particular, lemma 4 is the key lemma that provides a new potential function with which this paper achieves acceleration. Further, lemma 16 is the lemma that bounds all the terms of the bias error.
- In section E, we will bound the variance error of the proposed accelerated stochastic gradient method. In particular, lemma 6 is the key lemma that considers a stochastic process view of the proposed accelerated stochastic gradient method and provides a sharp bound on the covariance of the stationary distribution of the iterates. Furthermore, lemma 20 bounds all terms of the variance error.
- Section F presents the proof of Theorem 1. In particular, this section aggregates the result of lemma 16 (which bounds all terms of the bias error) and lemma 20 (which bounds all terms of the variance error) to present the guarantees of Algorithm 1.

A.2. Notations

We begin by introducing $\mathcal{M}$, which is the fourth moment tensor of the input $\mathbf{a} \sim \mathcal{D}$, i.e.:

$$\mathcal{M} \stackrel{\text{def}}{=} \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[\mathbf{a}\otimes\mathbf{a}\otimes\mathbf{a}\otimes\mathbf{a}\right]$$

Applying the fourth moment tensor $\mathcal{M}$ to any matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$ produces another matrix in $\mathbb{R}^{d\times d}$ that is expressed as:

$$\mathcal{M}\mathbf{S} \stackrel{\text{def}}{=} \mathbb{E}\left[(\mathbf{a}^\top\mathbf{S}\mathbf{a})\mathbf{a}\mathbf{a}^\top\right].$$

With this definition in place, we recall R^2 as the smallest number, such that $\mathcal{M}$ applied to the identity matrix $\mathbf{I}$ satisfies:

$$\mathcal{M}\mathbf{I} = \mathbb{E}\left[\|\mathbf{a}\|_2^2\,\mathbf{a}\mathbf{a}^\top\right] \preceq R^2\,\mathbf{H}$$

Moreover, we recall that the condition number of the distribution $\kappa = R^2/\mu$, where μ is the smallest eigenvalue of $\mathbf{H}$. Furthermore, the definition of the statistical condition number $\widetilde{\kappa}$ of the distribution follows by applying the fourth moment tensor $\mathcal{M}$ to $\mathbf{H}^{-1}$, i.e.:

$$\mathcal{M}\mathbf{H}^{-1} = \mathbb{E}\left[(\mathbf{a}^\top \mathbf{H}^{-1}\mathbf{a}) \cdot \mathbf{a}\mathbf{a}^\top\right] \preceq \widetilde{\kappa}\,\mathbf{H}$$

We denote by $\mathcal{A}_\mathcal{L}$ and $\mathcal{A}_\mathcal{R}$ the left and right multiplication operator of any matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, i.e. for any matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$, $\mathcal{A}_\mathcal{L}\mathbf{S} = \mathbf{A}\mathbf{S}$ and $\mathcal{A}_\mathcal{R}\mathbf{S} = \mathbf{S}\mathbf{A}$.

Parameter choices: In all of appendix we choose the parameters in Algorithm 1 as

$$\alpha = \frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2\sqrt{2c_1 - c_1^2} + \sqrt{\kappa\widetilde{\kappa}}},\ \ \beta = c_3 \frac{c_2\sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}},\ \ \gamma = c_2\frac{\sqrt{2c_1 - c_1^2}}{\mu\sqrt{\kappa\widetilde{\kappa}}},\ \ \delta = \frac{c_1}{R^2}$$

where c_1 is an arbitrary constant satisfying $0 < c_1 < \frac{1}{2}$. Furthermore, we note that $c_3 = \frac{c_2\sqrt{2c_1 - c_1^2}}{c_1}$, $c_2^2 = \frac{c_4}{2-c_1}$ and $c_4 < 1/6$. Note that we recover Theorem 1 by choosing $c_1 = 1/5, c_2 = \sqrt{5}/9, c_3 = \sqrt{5}/3, c_4 = 1/9$. We denote

$$c \stackrel{\text{def}}{=} \alpha(1-\beta) \text{ and, } q \stackrel{\text{def}}{=} \alpha\delta + (1-\alpha)\gamma.$$

Recall that $\mathbf{x}^*$ denotes unique minimizer of $P(\mathbf{x})$, i.e. $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\left[(b - \langle \mathbf{x}, \mathbf{a}\rangle)^2\right]$. We track $\boldsymbol{\theta}_k = \begin{bmatrix}\mathbf{x}_k - \mathbf{x}^* \\ \mathbf{y}_k - \mathbf{x}^*\end{bmatrix}$. The following equation captures the updates of Algorithm 1:

$$\begin{aligned}\boldsymbol{\theta}_{k+1} &= \begin{bmatrix} 0 & \mathbf{I} - \delta\widehat{\mathbf{H}}_{k+1} \\ -c\cdot\mathbf{I} & (1+c)\cdot\mathbf{I} - q\cdot\widehat{\mathbf{H}}_{k+1}\end{bmatrix}\boldsymbol{\theta}_k + \begin{bmatrix}\delta\cdot\epsilon_{k+1}\mathbf{a}_{k+1} \\ q\cdot\epsilon_{k+1}\mathbf{a}_{k+1}\end{bmatrix} \\ &\stackrel{\text{def}}{=} \widehat{\mathbf{A}}_{k+1}\boldsymbol{\theta}_k + \boldsymbol{\zeta}_{k+1}, \end{aligned} \tag{10}$$

where, $\widehat{\mathbf{H}}_{k+1} \stackrel{\text{def}}{=} \mathbf{a}_{k+1}\mathbf{a}_{k+1}^\top$, $\widehat{\mathbf{A}}_{k+1} \stackrel{\text{def}}{=} \begin{bmatrix} 0 & \mathbf{I} - \delta\widehat{\mathbf{H}}_{k+1} \\ -c\cdot\mathbf{I} & (1+c)\cdot\mathbf{I} - q\cdot\widehat{\mathbf{H}}_{k+1}\end{bmatrix}$ and $\boldsymbol{\zeta}_{k+1} \stackrel{\text{def}}{=} \begin{bmatrix}\delta\cdot\epsilon_{k+1}\mathbf{a}_{k+1} \\ q\cdot\epsilon_{k+1}\mathbf{a}_{k+1}\end{bmatrix}$. Furthermore, we denote by $\boldsymbol{\Phi}_k$ the expected covariance of $\boldsymbol{\theta}_k$, i.e.:

$$\boldsymbol{\Phi}_k \stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\theta}_k \otimes \boldsymbol{\theta}_k\right].$$

Next, let $\mathcal{F}_k$ denote the filtration generated by samples $\{(\mathbf{a}_1, b_1), \cdots, (\mathbf{a}_k, b_k)\}$. Then,

$$\mathbf{A} \stackrel{\text{def}}{=} \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1}|\mathcal{F}_k\right] = \begin{bmatrix} 0 & \mathbf{I} - \delta\mathbf{H} \\ -c\mathbf{I} & (1+c)\mathbf{I} - q\mathbf{H}\end{bmatrix}.$$

By iterated conditioning, we also have

$$\mathbb{E}\left[\boldsymbol{\theta}_{k+1}|\mathcal{F}_k\right] = \mathbf{A}\boldsymbol{\theta}_k. \tag{11}$$

Without loss of generality, we assume that $\mathbf{H}$ is a diagonal matrix. We now note that we can rearrange the coordinates through an eigenvalue decomposition so that $\mathbf{A}$ becomes a block-diagonal matrix with 2×2 blocks. We denote the j^{th} block by $\mathbf{A}_j$:

$$\mathbf{A}_j \stackrel{\text{def}}{=} \begin{bmatrix} 0 & 1 - \delta\lambda_j \\ -c & 1 + c - q\lambda_j\end{bmatrix},$$

where λ_j denotes the j^{th} eigenvalue of $\mathbf{H}$. Next,

$$\mathcal{B} \stackrel{\text{def}}{=} \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1} \otimes \widehat{\mathbf{A}}_{k+1} | \mathcal{F}_k\right], \text{ and}$$
$$\widehat{\mathbf{\Sigma}} \stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\zeta}_{k+1} \otimes \boldsymbol{\zeta}_{k+1} | \mathcal{F}_k\right] = \begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{\Sigma} \preceq \sigma^2 \cdot \begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H}.$$

Finally, we observe the following:

$$\begin{aligned} \mathbb{E}\left[(\mathbf{A} - \widehat{\mathbf{A}}_{k+1}) \otimes (\mathbf{A} - \widehat{\mathbf{A}}_{k+1}) | \mathcal{F}_k\right] &= \mathbf{A} \otimes \mathbf{A} - \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1} \otimes \mathbf{A} | \mathcal{F}_k\right] \\ &\quad - \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1} \otimes \mathbf{A} | \mathcal{F}_k\right] + \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1} \otimes \widehat{\mathbf{A}}_{k+1} | \mathcal{F}_k\right] \\ &= -\mathbf{A} \otimes \mathbf{A} + \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1} \otimes \widehat{\mathbf{A}}_{k+1} | \mathcal{F}_k\right] \\ \implies \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1} \otimes \widehat{\mathbf{A}}_{k+1} | \mathcal{F}_k\right] &= \mathbb{E}\left[(\mathbf{A} - \widehat{\mathbf{A}}_{k+1}) \otimes (\mathbf{A} - \widehat{\mathbf{A}}_{k+1}) | \mathcal{F}_k\right] + \mathbf{A} \otimes \mathbf{A} \end{aligned}$$

We now define:

$$\mathcal{R} \stackrel{\text{def}}{=} \mathbb{E}\left[(\mathbf{A} - \widehat{\mathbf{A}}_{k+1}) \otimes (\mathbf{A} - \widehat{\mathbf{A}}_{k+1}) | \mathcal{F}_k\right], \text{ and}$$
$$\mathcal{D} \stackrel{\text{def}}{=} \mathbf{A} \otimes \mathbf{A}.$$

Thus implying the following relation between the operators $\mathcal{B}, \mathcal{D}$ and $\mathcal{R}$:

$$\mathcal{B} = \mathcal{D} + \mathcal{R}.$$

Appendix B. The Tail-Average Iterate: Covariance and bias-variance decomposition

We begin by considering the first-order Markovian recursion as defined by equation 10:

$$\boldsymbol{\theta}_j = \widehat{\mathbf{A}}_j \boldsymbol{\theta}_{j-1} + \boldsymbol{\zeta}_j.$$

We refer by $\mathbf{\Phi}_j$ the covariance of the j^{th} iterate, i.e.:

$$\mathbf{\Phi}_j \stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right] \tag{12}$$

Consider a decomposition of $\boldsymbol{\theta}_j$ as $\boldsymbol{\theta}_j = \boldsymbol{\theta}_j^{\text{bias}} + \boldsymbol{\theta}_j^{\text{variance}}$, where $\boldsymbol{\theta}_j^{\text{bias}}$ and $\boldsymbol{\theta}_j^{\text{variance}}$ are defined as follows:

$$\boldsymbol{\theta}_j^{\text{bias}} \stackrel{\text{def}}{=} \widehat{\mathbf{A}}_j \boldsymbol{\theta}_{j-1}^{\text{bias}}; \qquad \boldsymbol{\theta}_0^{\text{bias}} \stackrel{\text{def}}{=} \boldsymbol{\theta}_0, \text{ and} \tag{13}$$
$$\boldsymbol{\theta}_j^{\text{variance}} \stackrel{\text{def}}{=} \widehat{\mathbf{A}}_j \boldsymbol{\theta}_{j-1}^{\text{variance}} + \boldsymbol{\zeta}_j; \qquad \boldsymbol{\theta}_0^{\text{variance}} \stackrel{\text{def}}{=} 0. \tag{14}$$

We note that

$$\mathbb{E}\left[\boldsymbol{\theta}_j^{\text{bias}}\right] = \mathbf{A}\mathbb{E}\left[\boldsymbol{\theta}_{j-1}^{\text{bias}}\right], \tag{15}$$
$$\mathbb{E}\left[\boldsymbol{\theta}_j^{\text{variance}}\right] = \mathbf{A}\mathbb{E}\left[\boldsymbol{\theta}_{j-1}^{\text{variance}}\right]. \tag{16}$$

Note equation 16 follows using a conditional expectation argument with the fact that $\mathbb{E}\left[\boldsymbol{\zeta}_k\right] = 0\,\forall\, k$ owing to first order optimality conditions.

Before we prove the decomposition holds using an inductive argument, let us understand what the bias and variance sub-problem intuitively mean.

Note that the *bias* sub-problem (defined by equation 13) refers to running algorithm on the noiseless problem (i.e., where, $\boldsymbol{\zeta}_{\boldsymbol{\cdot}} = 0$ a.s.) by starting it at $\boldsymbol{\theta}_0^{\text{bias}} = \boldsymbol{\theta}_0$. The bias essentially measures the dependence of the generalization error on the excess risk of the initial point $\boldsymbol{\theta}_0$ and bears similarities to convergence rates studied in the context of offline optimization.

The *variance* sub-problem (defined by equation 14) measures the dependence of the generalization error on the noise introduced during the course of optimization, and this is associated with the statistical aspects of the optimization problem. The variance can be understood as starting the algorithm at the solution ($\boldsymbol{\theta}_0^{\text{variance}} = 0$) and running the optimization driven solely by noise. Note that the variance is associated with sharp statistical lower bounds which dictate its rate of decay as a function of the number of oracle calls n.

Now, we will prove that the decomposition $\boldsymbol{\theta}_j = \boldsymbol{\theta}_j^{\text{bias}} + \boldsymbol{\theta}_j^{\text{variance}}$ captures the recursion expressed in equation 10 through induction. For the base case $j = 1$, we see that

$$\begin{aligned}\boldsymbol{\theta}_1 &= \widehat{\mathbf{A}}_1\boldsymbol{\theta}_0 + \boldsymbol{\zeta}_1\\ &= \underbrace{\widehat{\mathbf{A}}_1\boldsymbol{\theta}_0^{\text{bias}}}_{\because\ \boldsymbol{\theta}_0^{\text{bias}}=\boldsymbol{\theta}_0} + \underbrace{\widehat{\mathbf{A}}_1\boldsymbol{\theta}_0^{\text{variance}}}_{=0,\ \because\ \boldsymbol{\theta}_0^{\text{variance}}=0} + \boldsymbol{\zeta}_1\\ &= \boldsymbol{\theta}_1^{\text{bias}} + \boldsymbol{\theta}_1^{\text{variance}}\end{aligned}$$

Now, for the inductive step, let us assume that the decomposition holds in the $j-1^{st}$ iteration, i.e. we assume $\boldsymbol{\theta}_{j-1} = \boldsymbol{\theta}_{j-1}^{\text{bias}} + \boldsymbol{\theta}_{j-1}^{\text{variance}}$. We will then prove that this relation holds in the j^{th} iteration. Towards this, we will write the recursion:

$$\begin{aligned}\boldsymbol{\theta}_j &= \widehat{\mathbf{A}}_j\boldsymbol{\theta}_{j-1} + \boldsymbol{\zeta}_j\\ &= \widehat{\mathbf{A}}_j(\boldsymbol{\theta}_{j-1}^{\text{bias}} + \boldsymbol{\theta}_{j-1}^{\text{variance}}) + \boldsymbol{\zeta}_j \quad \text{(using the inductive hypothesis)}\\ &= \widehat{\mathbf{A}}_j\boldsymbol{\theta}_{j-1}^{\text{bias}} + \widehat{\mathbf{A}}_j\boldsymbol{\theta}_{j-1}^{\text{variance}} + \boldsymbol{\zeta}_j\\ &= \boldsymbol{\theta}_j^{\text{bias}} + \boldsymbol{\theta}_j^{\text{variance}}.\end{aligned}$$

This proves the decomposition holds through a straight forward inductive argument.

In a similar manner as $\boldsymbol{\theta}_j$, the tail-averaged iterate $\bar{\boldsymbol{\theta}}_{t,n} \stackrel{\text{def}}{=} \frac{1}{n-t}\sum_{j=t+1}^{n}\boldsymbol{\theta}_j$ can also be written as $\bar{\boldsymbol{\theta}}_{t,n} = \bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}} + \bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}}$, where $\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}} \stackrel{\text{def}}{=} \frac{1}{n-t}\sum_{j=t+1}^{n}\boldsymbol{\theta}_j^{\text{bias}}$ and $\bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}} \stackrel{\text{def}}{=} \frac{1}{n-t}\sum_{j=t+1}^{n}\boldsymbol{\theta}_j^{\text{variance}}$. Furthermore, the tail-averaged iterate $\bar{\boldsymbol{\theta}}_{t,n}$ and its bias and variance counterparts $\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}, \bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}}$ are associated with their corresponding covariance matrices $\bar{\boldsymbol{\Phi}}_{t,n}, \bar{\boldsymbol{\Phi}}_{t,n}^{\text{bias}}, \bar{\boldsymbol{\Phi}}_{t,n}^{\text{variance}}$ respectively. Note that $\bar{\boldsymbol{\Phi}}_{t,n}$ can be upper bounded using Cauchy-Shwartz inequality as:

$$\begin{aligned}\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}\otimes\bar{\boldsymbol{\theta}}_{t,n}\right] &\preceq 2\cdot\left(\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}\otimes\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}\right] + \mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}}\otimes\bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}}\right]\right)\\ \implies \bar{\boldsymbol{\Phi}}_{t,n} &\preceq 2\cdot(\bar{\boldsymbol{\Phi}}_{t,n}^{\text{bias}} + \bar{\boldsymbol{\Phi}}_{t,n}^{\text{variance}}).\end{aligned} \tag{17}$$

The above inequality is referred to as the *bias-variance* decomposition and is well known from previous work Bach and Moulines (2013); Frostig et al. (2015b); Jain et al. (2016), and we re-derive this decomposition for the sake of completeness. We will now derive an expression for the covariance of the tail-averaged iterate and apply it to obtain the covariance of the bias ($\bar{\mathbf{\Phi}}^{\text{bias}}_{t,n}$) and variance ($\bar{\mathbf{\Phi}}^{\text{variance}}_{t,n}$) error of the tail-averaged iterate.

B.1. The tail-averaged iterate and its covariance

We begin by writing out an expression for the tail-averaged iterate $\bar{\boldsymbol{\theta}}_{t,n}$ as:

$$\bar{\boldsymbol{\theta}}_{t,n} = \frac{1}{n-t}\sum_{j=t+1}^{n} \boldsymbol{\theta}_j$$

To get the excess risk of the tail-averaged iterate $\bar{\boldsymbol{\theta}}_{t,n}$, we track its covariance $\bar{\mathbf{\Phi}}_{t,n}$:

$$\begin{aligned}
\bar{\mathbf{\Phi}}_{t,n} &= \mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n} \otimes \bar{\boldsymbol{\theta}}_{t,n}\right] \\
&= \frac{1}{(n-t)^2}\sum_{j,l=t+1}^{n} \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_l\right] \\
&= \frac{1}{(n-t)^2}\sum_{j}\left(\sum_{l=t+1}^{j-1} \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_l\right] + \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right] + \sum_{l=j+1}^{n} \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_l\right]\right) \\
&= \frac{1}{(n-t)^2}\sum_{j}\left(\sum_{l=t+1}^{j-1} \mathbf{A}^{j-l}\mathbb{E}\left[\boldsymbol{\theta}_l \otimes \boldsymbol{\theta}_l\right] + \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right] + \sum_{l=j+1}^{n} \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right](\mathbf{A}^\top)^{l-j}\right) (\text{ from } (11)) \\
&= \frac{1}{(n-t)^2}\left(\sum_{l=t+1}^{n}\sum_{j=l+1}^{n} \mathbf{A}^{j-l}\mathbb{E}\left[\boldsymbol{\theta}_l \otimes \boldsymbol{\theta}_l\right] + \sum_{j=t+1}^{n}\mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right] + \sum_{j=t+1}^{n}\sum_{l=j+1}^{n} \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right](\mathbf{A}^\top)^{l-j}\right) \\
&= \frac{1}{(n-t)^2}\left(\sum_{j=t+1}^{n}\sum_{l=j+1}^{n} \mathbf{A}^{l-j}\mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right] + \sum_{j=t+1}^{n}\mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right] + \sum_{j=t+1}^{n}\sum_{l=j+1}^{n} \mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right](\mathbf{A}^\top)^{l-j}\right) \\
&= \frac{1}{(n-t)^2}\left(\sum_{j=t+1}^{n}(\mathbf{I}-\mathbf{A})^{-1}(\mathbf{A}-\mathbf{A}^{n+1-j})\mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right] + \sum_{j=t+1}^{n}\mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right]\right. \\
&\qquad\qquad \left. + \sum_{j=t+1}^{n}\mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right](\mathbf{I}-\mathbf{A}^\top)^{-1}(\mathbf{A}^\top-(\mathbf{A}^\top)^{n+1-j})\right) \\
&= \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_\mathcal{L})^{-1}(\mathcal{A}_\mathcal{L}-\mathcal{A}_\mathcal{L}^{n+1-j}) + (\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-1}(\mathcal{A}_\mathcal{R}^\top-(\mathcal{A}_\mathcal{R}^\top)^{n+1-j})\right)\mathbb{E}\left[\boldsymbol{\theta}_j \otimes \boldsymbol{\theta}_j\right] \\
&= \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_\mathcal{L})^{-1}(\mathcal{A}_\mathcal{L}-\mathcal{A}_\mathcal{L}^{n+1-j}) + (\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-1}(\mathcal{A}_\mathcal{R}^\top-(\mathcal{A}_\mathcal{R}^\top)^{n+1-j})\right)\mathbf{\Phi}_j.
\end{aligned} \tag{18}$$

Note that the above recursion can be applied to obtain the covariance of the tail-averaged iterate for the bias ($\bar{\mathbf{\Phi}}^{\text{bias}}_{t,n}$) and variance ($\bar{\mathbf{\Phi}}^{\text{variance}}_{t,n}$) error, since the conditional expectation arguments employed in obtaining equation 18 are satisfied by both the recursion used in tracking the bias error (i.e.

equation 13) and the variance error (i.e. equation 14). This implies that,

$$\bar{\mathbf{\Phi}}_{t,n}^{\text{bias}} \stackrel{\text{def}}{=} \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{A}_{\mathcal{L}} - \mathcal{A}_{\mathcal{L}}^{n+1-j}) + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top} - (\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j})\right)\mathbf{\Phi}_j^{\text{bias}} \tag{19}$$

$$\bar{\mathbf{\Phi}}_{t,n}^{\text{variance}} \stackrel{\text{def}}{=} \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{A}_{\mathcal{L}} - \mathcal{A}_{\mathcal{L}}^{n+1-j}) + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top} - (\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j})\right)\mathbf{\Phi}_j^{\text{variance}} \tag{20}$$

B.2. Covariance of Bias error of the tail-averaged iterate

Proof [Proof of Lemma 3] To obtain the covariance of the bias error of the tail-averaged iterate, we first need to obtain $\mathbf{\Phi}_j^{\text{bias}}$, which we will by unrolling the recursion of equation 13:

$$\begin{aligned}
\boldsymbol{\theta}_k^{\text{bias}} &= \widehat{\mathbf{A}}_k \boldsymbol{\theta}_{k-1}^{\text{bias}} \\
\implies \mathbf{\Phi}_k^{\text{bias}} &= \mathbb{E}\left[\boldsymbol{\theta}_k^{\text{bias}} \otimes \boldsymbol{\theta}_k^{\text{bias}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\theta}_k^{\text{bias}} \otimes \boldsymbol{\theta}_k^{\text{bias}} | \mathcal{F}_{k-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\widehat{\mathbf{A}}_k \boldsymbol{\theta}_{k-1}^{\text{bias}} \otimes \boldsymbol{\theta}_{k-1}^{\text{bias}} \widehat{\mathbf{A}}_k^{\top} | \mathcal{F}_{k-1}\right]\right] \\
&= \mathcal{B}\,\mathbb{E}\left[\boldsymbol{\theta}_{k-1}^{\text{bias}} \otimes \boldsymbol{\theta}_{k-1}^{\text{bias}}\right] = \mathcal{B}\,\mathbf{\Phi}_{k-1}^{\text{bias}} \\
\implies \mathbf{\Phi}_k^{\text{bias}} &= \mathcal{B}^k\,\mathbf{\Phi}_0^{\text{bias}}
\end{aligned} \tag{21}$$

Next, we recount the equation for the covariance of the bias of the tail-averaged iterate from equation 19:

$$\bar{\mathbf{\Phi}}_{t,n}^{\text{bias}} = \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{A}_{\mathcal{L}} - \mathcal{A}_{\mathcal{L}}^{n+1-j}) + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top} - (\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j})\right)\mathbf{\Phi}_j^{\text{bias}}$$

Now, we substitute $\mathbf{\Phi}_j^{\text{bias}}$ from equation 21:

$$\begin{aligned}
\bar{\mathbf{\Phi}}_{t,n}^{\text{bias}} &= \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{A}_{\mathcal{L}} - \mathcal{A}_{\mathcal{L}}^{n+1-j}) + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top} - (\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j})\right)\mathcal{B}^j\mathbf{\Phi}_0 \\
&= \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\right)\mathcal{B}^j\mathbf{\Phi}_0 \\
&\quad - \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\left((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j}\right)\mathcal{B}^j\mathbf{\Phi}_0 \\
&= \underbrace{\frac{1}{(n-t)^2}\left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\right)(\mathcal{I}-\mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\mathbf{\Phi}_0}_{\text{Leading order term}}
\end{aligned}$$

$$-\frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\Big((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j}+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j}\Big)\mathcal{B}^j\mathbf{\Phi}_0. \tag{22}$$

There are two points to note here: (a) The second line consists of terms that constitute the lower-order terms of the bias. We will bound the summation by taking a supremum over j. (b) Note that the burn-in phase consisting of t unaveraged iterations allows for a geometric decay of the bias, followed by the tail-averaged phase that allows for a sublinear rate of bias decay. ■

B.3. Covariance of Variance error of the tail-averaged iterate

Proof [Proof of Lemma 5] Before obtaining the covariance of the tail-averaged iterate, we note that $\mathbb{E}\left[\boldsymbol{\theta}_j^{\text{variance}}\right]=0\ \forall\ j$. This can be easily seen since $\boldsymbol{\theta}_0^{\text{variance}}=0$ and $\mathbb{E}\left[\boldsymbol{\theta}_k^{\text{variance}}\right]=\mathbf{A}\mathbb{E}\left[\boldsymbol{\theta}_{k-1}^{\text{variance}}\right]$ (from equation 16).

Next, in order to obtain the covariance of the variance of the tail-averaged iterate, we first need to obtain $\mathbf{\Phi}_j^{\text{variance}}$, and we will obtain this by unrolling the recursion of equation 14:

$$\begin{aligned}
\boldsymbol{\theta}_k^{\text{variance}} &= \widehat{\mathbf{A}}_k\boldsymbol{\theta}_{k-1}^{\text{variance}}+\boldsymbol{\zeta}_k\\
\implies \mathbf{\Phi}_k^{\text{variance}} &= \mathbb{E}\left[\boldsymbol{\theta}_k^{\text{variance}}\otimes\boldsymbol{\theta}_k^{\text{variance}}\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\theta}_k^{\text{variance}}\otimes\boldsymbol{\theta}_k^{\text{variance}}|\mathcal{F}_{k-1}\right]\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\widehat{\mathbf{A}}_k\boldsymbol{\theta}_{k-1}^{\text{variance}}\otimes\boldsymbol{\theta}_{k-1}^{\text{variance}}\widehat{\mathbf{A}}_k^{\top}+\boldsymbol{\zeta}_k\otimes\boldsymbol{\zeta}_k|\mathcal{F}_{k-1}\right]\right]\\
&= \mathcal{B}\ \mathbb{E}\left[\boldsymbol{\theta}_{k-1}^{\text{variance}}\otimes\boldsymbol{\theta}_{k-1}^{\text{variance}}\right]+\widehat{\mathbf{\Sigma}}=\mathcal{B}\ \mathbf{\Phi}_{k-1}^{\text{variance}}+\widehat{\mathbf{\Sigma}}\\
\implies \mathbf{\Phi}_k^{\text{variance}} &= \sum_{j=0}^{k-1}\mathcal{B}^j\ \widehat{\mathbf{\Sigma}}\\
&= (\mathbf{I}-\mathcal{B})^{-1}(\mathcal{I}-\mathcal{B}^k)\widehat{\mathbf{\Sigma}}
\end{aligned} \tag{23}$$

Note that the cross terms in the outer product computations vanish owing to the fact that $\mathbb{E}\left[\boldsymbol{\theta}_{k-1}^{\text{variance}}\right]=0\ \forall\ k$. We then recount the expression for the covariance of the variance error from equation 20:

$$\bar{\mathbf{\Phi}}_{t,n}^{\text{variance}}=\frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\Big(\mathcal{I}+(\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{A}_{\mathcal{L}}-\mathcal{A}_{\mathcal{L}}^{n+1-j})+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top}-(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j})\Big)\mathbf{\Phi}_j^{\text{variance}}$$

We will substitute the expression for $\mathbf{\Phi}_j^{\text{variance}}$ from equation 23.

$$\bar{\mathbf{\Phi}}_{t,n}^{\text{variance}}=\frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\Big(\mathcal{I}+(\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{A}_{\mathcal{L}}-\mathcal{A}_{\mathcal{L}}^{n+1-j})+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top}-(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j})\Big)(\mathcal{I}-\mathcal{B})^{-1}(\mathcal{I}-\mathcal{B}^j)\widehat{\mathbf{\Sigma}}$$

Evaluating this summation, we have:

$$\begin{aligned}
\bar{\mathbf{\Phi}}_{t,n}^{\text{variance}} &= \underbrace{\frac{1}{n-t}\big(\mathcal{I}+(\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}}_{\text{Leading order term}}\\
&\quad-\frac{1}{(n-t)^2}\big((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-2}(\mathcal{A}_{\mathcal{L}}-\mathcal{A}_{\mathcal{L}}^{n+1-t})+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-2}(\mathcal{A}_{\mathcal{R}}^{\top}-(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-t})\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}
\end{aligned}$$

$$-\frac{1}{(n-t)^2}\big(\mathcal{I}+(\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathcal{A}_{\mathcal{R}}^\top\big)(\mathcal{I}-\mathcal{B})^{-2}(\mathcal{B}^{t+1}-\mathcal{B}^{n+1})\widehat{\mathbf{\Sigma}}$$
$$+\frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\big((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j}+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^\top)^{-1}(\mathcal{A}_{\mathcal{R}}^\top)^{n+1-j}\big)(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^j\widehat{\mathbf{\Sigma}} \tag{24}$$

■

Equations 17, 22, 24 wrap up the proof of lemmas 3, 5.

The parameter error of the (tail-)averaged iterate can be obtained using a trace operator $\langle\bullet,\bullet\rangle$ to the tail-averaged iterate's covariance $\bar{\mathbf{\Phi}}_{t,n}$ with the matrix $\begin{bmatrix}\mathbf{I} & 0\\ 0 & 0\end{bmatrix}$, i.e.

$$\|\bar{\mathbf{x}}_{t,n}-\mathbf{x}^*\|_2^2=\left\langle\begin{bmatrix}\mathbf{I} & 0\\ 0 & 0\end{bmatrix},\bar{\mathbf{\Phi}}_{t,n}\right\rangle$$

In order to obtain the function error, we note the following taylor expansion of the function $P(\bullet)$ around the minimizer $\mathbf{x}^*$:

$$\begin{aligned}P(\mathbf{x})&=P(\mathbf{x}^*)+\frac{1}{2}\|\mathbf{x}-\mathbf{x}^*\|^2_{\nabla^2P(\mathbf{x}^*)}\\&=P(\mathbf{x}^*)+\frac{1}{2}\|\mathbf{x}-\mathbf{x}^*\|^2_{\mathbf{H}}\end{aligned}$$

This implies the excess risk can be obtained as:

$$\begin{aligned}P(\bar{\mathbf{x}}_{t,n})-P(\mathbf{x}^*)&=\frac{1}{2}\cdot\left\langle\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix},\bar{\mathbf{\Phi}}_{t,n}\right\rangle\\&\leq\left\langle\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix},\bar{\mathbf{\Phi}}^{\text{bias}}_{t,n}\right\rangle+\left\langle\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix},\bar{\mathbf{\Phi}}^{\text{variance}}_{t,n}\right\rangle\end{aligned}$$

Appendix C. Useful lemmas

In this section, we will state and prove some useful lemmas that will be helpful in the later sections.

Lemma 7

$$\left(\mathbf{I}-\mathbf{A}^\top\right)^{-1}\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}=\frac{1}{q-c\delta}\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H}) & 0\\ (\mathbf{I}-\delta\mathbf{H}) & 0\end{bmatrix}$$

Proof Since we assumed that $\mathbf{H}$ is a diagonal matrix (with out loss of generality), we note that $\mathbf{A}$ is a block diagonal matrix after a rearrangement of the co-ordinates (via an eigenvalue decomposition).

In particular, by considering the j^{th} block (denoted by $\mathbf{A}_j$ corresponding to the j^{th} eigenvalue λ_j of $\mathbf{H}$), we have:

$$\mathbf{I}-\mathbf{A}_j^\top=\begin{bmatrix}1 & c\\ -(1-\delta\lambda_j) & -(c-q\lambda_j)\end{bmatrix}$$

Implying that the determinant $\left|\mathbf{I}-\mathbf{A}_j^\top\right| = (q-c\delta)\lambda_j$, using which:

$$(\mathbf{I}-\mathbf{A}_j^\top)^{-1} = \frac{1}{(q-c\delta)\lambda_j}\begin{bmatrix}-(c-q\lambda_j) & -c\\ 1-\delta\lambda_j & 1\end{bmatrix} \tag{25}$$

Thus,

$$(\mathbf{I}-\mathbf{A}_j^\top)^{-1}\begin{bmatrix}\lambda_j & 0\\ 0 & 0\end{bmatrix} = \frac{1}{q-c\delta}\begin{bmatrix}-(c-q\lambda_j) & 0\\ (1-\delta\lambda_j) & 0\end{bmatrix}$$

Accumulating the results of each of the blocks and by rearranging the co-ordinates, the result follows. ■

Lemma 8

$$\left(\mathbf{I}-\mathbf{A}^\top\right)^{-1}\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}(\mathbf{I}-\mathbf{A})^{-1} = \frac{1}{(q-c\delta)^2}\left(\otimes_2\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2}\\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix}\right)$$

Proof

In a similar manner as in lemma 7, we decompose the computation into each of the eigen-directions and subsequently re-arrange the results. In particular, we note:

$$(\mathbf{I}-\mathbf{A}_j)^{-1} = \frac{1}{(q-c\delta)\lambda_j}\begin{bmatrix}-(c-q\lambda_j) & (1-\delta\lambda_j)\\ -c & 1\end{bmatrix}$$

Multiplying the above with the result of lemma 7, we have:

$$(\mathbf{I}-\mathbf{A}_j^\top)^{-1}\begin{bmatrix}\lambda_j & 0\\ 0 & 0\end{bmatrix}(\mathbf{I}-\mathbf{A}_j)^{-1} = \frac{1}{(q-c\delta)^2}\left(\otimes_2\begin{bmatrix}-(c-q\lambda_j)\lambda_j^{-1/2}\\ (1-\delta\lambda_j)\lambda_j^{-1/2}\end{bmatrix}\right)$$

From which the statement of the lemma follows through a simple re-arrangement. ■

Lemma 9

$$\left(\mathbf{I}-\mathbf{A}^\top\right)^{-2}\mathbf{A}^\top\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix} = \frac{1}{(q-c\delta)^2}\begin{bmatrix}\mathbf{H}^{-1}(-c(1-c)\mathbf{I}-cq\mathbf{H})(\mathbf{I}-\delta\mathbf{H}) & 0\\ \mathbf{H}^{-1}((1-c)\mathbf{I}-c\delta\mathbf{H})(\mathbf{I}-\delta\mathbf{H}) & 0\end{bmatrix}$$

Proof In a similar argument as in previous two lemmas, we analyze the expression in each eigendirection of $\mathbf{H}$ through a rearrangement of the co-ordinates. Utilizing the expression of $\mathbf{I}-\mathbf{A}_j^\top$ from equation 25, we get:

$$(\mathbf{I}-\mathbf{A}_j^\top)^{-1}\mathbf{A}_j^\top\begin{bmatrix}\lambda_j & 0\\ 0 & 0\end{bmatrix} = \frac{1}{(q-c\delta)}\begin{bmatrix}-c(1-\delta\lambda_j) & 0\\ (1-\delta\lambda_j) & 0\end{bmatrix} \tag{26}$$

thus implying:

$$(\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top\begin{bmatrix}\lambda_j & 0\\ 0 & 0\end{bmatrix} = \frac{(1-\delta\lambda_j)}{(q-c\delta)^2\lambda_j}\begin{bmatrix}-c(1-c)-cq\lambda_j & 0\\ (1-c)-c\delta\lambda_j & 0\end{bmatrix}$$

Rearranging the co-ordinates, the statement of the lemma follows. ■

Lemma 10 *The matrix* $\mathbf{A}$ *satisfies the following properties:*

1. *Eigenvalues* q *of* $\mathbf{A}$ *satisfy* $|q| \leq \sqrt{\alpha}$*, and*
2. $\left\|\mathbf{A}^k\right\|_2 \leq 3\sqrt{2} \cdot k \cdot \alpha^{\frac{k-1}{2}} \;\forall\; k \geq 1.$

Proof Since the matrix is block-diagonal with 2×2 blocks, after a rearranging the coordinates, we will restrict ourselves to bounding the eigenvalues and eigenvectors of each of these 2×2 blocks. Combining the results for different blocks then proves the lemma. Recall that $\mathbf{A}_j = \begin{bmatrix} 0 & 1-\delta\lambda_j \\ -c & 1+c-q\lambda_j \end{bmatrix}$.

Part I: Let us first prove the statement about the eigenvalues of $\mathbf{A}$. There are two scenarios here:

1. *Complex eigenvalues*: In this case, both eigenvalues of $\mathbf{A}_j$ have the same magnitude which is given by $\sqrt{\det(\mathbf{A}_j)} = \sqrt{c(1-\delta\lambda_j)} \leq \sqrt{c} \leq \sqrt{\alpha}$.
2. *Real eigenvalues*: Let q_1 and q_2 be the two real eigenvalues of $\mathbf{A}_j$. We know that $q_1 + q_2 = \mathrm{Tr}\left(\mathbf{A}_j\right) = 1 + c - q\lambda_j > 0$ and $q_1 \cdot q_2 = \det(\mathbf{A}_j) > 0$. This means that $q_1 > 0$ and $q_2 > 0$. Now, consider the matrix $\mathbf{G}_j \stackrel{\text{def}}{=} (1-\beta)\mathbf{I} - \mathbf{A}_j = \begin{bmatrix} (1-\beta) & -1+\delta\lambda_j \\ c & -1+(1-\beta)(1-\alpha)+q\lambda_j \end{bmatrix}$. We see that $((1-\beta)-q_1)((1-\beta)-q_2) = \det(\mathbf{G}_j) = (1-\beta)(1-\alpha)\left((1-\beta)-1\right) + (1-\beta)\left(q-\alpha\delta\right)\lambda_j = (1-\beta)\left(1-\alpha\right)\left(\gamma\lambda_j - \beta\right) \geq 0$. This means that there are two possibilities: either $q_1, q_2 \geq (1-\beta)$ or $q_1, q_2 \leq (1-\beta)$. If the second condition is true, then we are done. If not, if $q_1, q_2 \geq (1-\beta)$, then $\max_i q_i = \frac{\det(\mathbf{A}_j)}{\min_i q_i} \leq \frac{c(1-\delta\lambda_j)}{(1-\beta)} \leq \alpha(1-\delta\lambda_j)$. Since $\sqrt{\alpha} \geq \alpha \geq 1-\beta$, this proves the first part of the lemma.

Part II: Let $\mathbf{A}_j = \mathbf{V}\mathbf{Q}\mathbf{V}^\top$ be the Schur decomposition of $\mathbf{A}_j$ where $\mathbf{Q} = \begin{bmatrix} q_1 & q \\ 0 & q_2 \end{bmatrix}$ is an upper triangular matrix with eigenvalues q_1 and q_2 of $\mathbf{A}_j$ on the diagonal and $\mathbf{V}$ is a unitary matrix i.e., $\mathbf{V}\mathbf{V}^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$. We first observe that $|q| \leq \left\|\mathbf{Q}\right\|_2 \stackrel{(\zeta_1)}{=} \left\|\mathbf{A}_j\right\|_2 \leq \left\|\mathbf{A}_j\right\|_F \leq \sqrt{6}$, where (ζ_1) follows from the fact that $\mathbf{V}$ is a unitary matrix. $\mathbf{V}$ being unitary also implies that $\mathbf{A}_j^k = \mathbf{V}\mathbf{Q}^k\mathbf{V}^\top$. On the other hand, a simple proof via induction tells us that

$$\mathbf{Q}^k = \begin{bmatrix} q_1^k & q\left(\sum_{\ell=1}^{k-1} q_1^\ell q_2^{k-\ell}\right) \\ 0 & q_2^k \end{bmatrix}.$$

So, we have $\left\|\mathbf{A}_j^k\right\|_2 = \left\|\mathbf{Q}^k\right\|_2 \leq \left\|\mathbf{Q}^k\right\|_F \leq \sqrt{3}k\left|q\right|\max\left(\left|q_1\right|^{k-1}, \left|q_2\right|^{k-1}\right) \leq 3\sqrt{2} \cdot k \cdot \alpha^{\frac{k-1}{2}}$, where we used $|q| \leq \sqrt{6}$ and $\max\left(\left|q_1\right|, \left|q_2\right|\right) \leq \sqrt{\alpha}$. ∎

Finally, we state and prove the following lemma which is a relation between left and right multiplication operators.

Lemma 11 *Let* $\mathbf{A}$ *be any matrix with* $\mathcal{A}_{\mathcal{L}} = \mathbf{A} \otimes \mathbf{I}$ *and* $\mathcal{A}_{\mathcal{R}} = \mathbf{I} \otimes \mathbf{A}$ *representing its left and right multiplication operators. Then, the following expression holds:*

$$\left(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathcal{A}_{\mathcal{R}}^\top\right)(\mathcal{I} - \mathcal{A}_{\mathcal{L}}\mathcal{A}_{\mathcal{R}}^\top)^{-1} = (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}$$

Proof Let us assume that $\mathbf{A}$ can be written in terms of its eigen decomposition as $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}$. Then the first claim is that $\mathcal{I}, \mathcal{A}_{\mathcal{L}}, \mathcal{A}_{\mathcal{R}}$ are diagonalized by the same basis consisting of the eigenvectors of $\mathbf{A}$, i.e. in particular, the matrix of eigenvectors of $\mathcal{I}, \mathcal{A}_{\mathcal{L}}, \mathcal{A}_{\mathcal{R}}$ can be written as $\mathbf{V} \otimes \mathbf{V}$. In particular, this implies, $\forall\, i, j \in \{1, 2, ..., d\} \times \{1, 2, ..., d\}$, we have, applying $\mathbf{v}_i \otimes \mathbf{v}_j$ to the LHS, we have:

$$\begin{aligned}
&\left(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathcal{A}_{\mathcal{R}}^\top\right)(\mathcal{I} - \mathcal{A}_{\mathcal{L}}\mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathbf{v}_i \otimes \mathbf{v}_j \\
&= (1 - \lambda_i\lambda_j)^{-1}\left(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathcal{A}_{\mathcal{R}}^\top\right)\mathbf{v}_i \otimes \mathbf{v}_j \\
&= (1 + \lambda_i(1 - \lambda_i)^{-1} + \lambda_j(1 - \lambda_j)^{-1}) \cdot (1 - \lambda_i\lambda_j)^{-1}\mathbf{v}_i \otimes \mathbf{v}_j
\end{aligned}$$

Applying $\mathbf{v}_i \otimes \mathbf{v}_j$ to the RHS, we have:

$$\begin{aligned}
&(\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathbf{v}_i \otimes \mathbf{v}_j \\
&= (1 - \lambda_i)^{-1}(1 - \lambda_j)^{-1}\mathbf{v}_i \otimes \mathbf{v}_j
\end{aligned}$$

The next claim is that for any scalars (real/complex) $x, y \neq 1$, the following statement holds implying the statement of the lemma:

$$(1 + (1 - x)^{-1}x + (1 - y)^{-1}y) \cdot (1 - xy)^{-1} = (1 - x)^{-1}(1 - y)^{-1}$$

■

Lemma 12 *Recall the matrix* $\mathbf{G}$ *defined as* $\mathbf{G} \stackrel{def}{=} \begin{bmatrix} \mathbf{I} & \frac{-\alpha}{1-\alpha}\mathbf{I} \\ 0 & \frac{1}{1-\alpha}\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mu\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ \frac{-\alpha}{1-\alpha}\mathbf{I} & \frac{1}{1-\alpha}\mathbf{I} \end{bmatrix}$. *The condition number of* $\mathbf{G}$, $\kappa(\mathbf{G})$ *satisfies* $\kappa(\mathbf{G}) \leq \frac{4\kappa}{\sqrt{1-\alpha^2}}$.

Proof Since the above matrix is block-diagonal after a rearrangement of coordinates, it suffices to compute the smallest and largest singular values of each block. Let λ_i be the i^{th} eigenvalue of $\mathbf{H}$. Let $\mathbf{C} \stackrel{\text{def}}{=} \begin{bmatrix} 1 & 0 \\ \frac{-\alpha}{1-\alpha} & \frac{1}{1-\alpha} \end{bmatrix}$ and consider the matrix $\mathbf{G}_i \stackrel{\text{def}}{=} \mathbf{C}\begin{bmatrix} 1 & 0 \\ 0 & \frac{\mu}{\lambda_i} \end{bmatrix}\mathbf{C}^\top$. The largest eigenvalue of $\mathbf{G}_i$ is at most $\sigma_{\text{max}}(\mathbf{C})^2$, while the smallest eigenvalue, $\sigma_{\text{min}}(\mathbf{G}_i)$ is at least $\frac{\mu}{\lambda_i} \cdot \sigma_{\text{min}}(\mathbf{C})^2$. We obtain the following bounds on $\sigma_{\text{min}}(\mathbf{C})$ and $\sigma_{\text{max}}(\mathbf{C})$.

$$\begin{aligned}
\sigma_{\text{max}}(\mathbf{C}) &\leq \|\mathbf{C}\|_F \leq \frac{2}{\sqrt{1-\alpha^2}} \quad (\because\ \alpha \leq 1) \\
\sigma_{\text{min}}(\mathbf{C}) &\geq \frac{\sqrt{\det(\mathbf{C}\mathbf{C}^\top)}}{\|\mathbf{C}\|_F} \geq \frac{1}{2}, \\
&\qquad \left(\because \det\left(\mathbf{C}\mathbf{C}^\top\right) = \sigma_{\text{max}}(\mathbf{C})^2\,\sigma_{\text{min}}(\mathbf{C})^2\right)
\end{aligned}$$

where we used the computation that $\det\left(\mathbf{C}\mathbf{C}^\top\right) = \frac{1}{1-\alpha}$. This means that $\sigma_{\text{min}}(\mathbf{G}_i) \geq \frac{\mu}{2\lambda_i}$ and $\sigma_{\text{max}}(\mathbf{G}_i) \leq \frac{2}{\sqrt{1-\alpha^2}}$. Combining all the blocks, we see that the condition number of $\mathbf{G}$ is at most $\frac{4\kappa}{\sqrt{1-\alpha^2}}$, proving the lemma. ■

Appendix D. Lemmas and proofs for bias contraction

Proof [Proof of Lemma 4] Let $\mathbf{v} \stackrel{\text{def}}{=} \frac{1}{1-\alpha}(\mathbf{y}-\alpha\mathbf{x})$ and consider the following update rules corresponding to the noiseless versions of the updates in Algorithm 1:

$$\begin{aligned}\mathbf{x}^+ &= \mathbf{y} - \delta\widehat{\mathbf{H}}(\mathbf{y}-\mathbf{x}^*)\\ \mathbf{z} &= \beta\mathbf{y} + (1-\beta)\mathbf{v}\\ \mathbf{v}^+ &= \mathbf{z} - \gamma\widehat{\mathbf{H}}(\mathbf{y}-\mathbf{x}^*)\\ \mathbf{y}^+ &= \alpha\mathbf{x}^+ + (1-\alpha)\mathbf{v}^+,\end{aligned}$$

where $\widehat{\mathbf{H}} \stackrel{\text{def}}{=} \mathbf{a}\mathbf{a}^\top$ where $\mathbf{a}$ is sampled from the marginal on $(\mathbf{a}, b) \sim \mathcal{D}$. We first note that

$$\begin{aligned}\mathbb{E}\left[\otimes_2 \begin{bmatrix}\mathbf{x}^+ - \mathbf{x}^*\\ \mathbf{y}^+ - \mathbf{x}^*\end{bmatrix}\right] &= \mathbb{E}\left[\widehat{\mathbf{A}}\left(\otimes_2 \begin{bmatrix}\mathbf{x} - \mathbf{x}^*\\ \mathbf{y} - \mathbf{x}^*\end{bmatrix}\right)\widehat{\mathbf{A}}^\top\right]\\ &= \mathcal{B}\left(\otimes_2 \begin{bmatrix}\mathbf{x} - \mathbf{x}^*\\ \mathbf{y} - \mathbf{x}^*\end{bmatrix}\right)\end{aligned}$$

Letting $\widetilde{\mathbf{G}} \stackrel{\text{def}}{=} \begin{bmatrix}\mathbf{I} & 0\\ \frac{-\alpha}{1-\alpha}\mathbf{I} & \frac{1}{1-\alpha}\mathbf{I}\end{bmatrix}$, we can verify that $\begin{bmatrix}\mathbf{x}-\mathbf{x}^*\\ \mathbf{v}-\mathbf{x}^*\end{bmatrix} = \widetilde{\mathbf{G}}\begin{bmatrix}\mathbf{x}-\mathbf{x}^*\\ \mathbf{y}-\mathbf{x}^*\end{bmatrix}$, similarly $\begin{bmatrix}\mathbf{x}^+-\mathbf{x}^*\\ \mathbf{v}^+-\mathbf{x}^*\end{bmatrix} = \widetilde{\mathbf{G}}\begin{bmatrix}\mathbf{x}^+-\mathbf{x}^*\\ \mathbf{y}^+-\mathbf{x}^*\end{bmatrix}$. Recall that $\mathbf{G} \stackrel{\text{def}}{=} \widetilde{\mathbf{G}}^\top \begin{bmatrix}\mathbf{I} & 0\\ 0 & \mu\mathbf{H}^{-1}\end{bmatrix}\widetilde{\mathbf{G}}$. With this notation in place, we prove the statement below, and substitute the values of c_1, c_2, c_3 to obtain the statement of the lemma:

$$\left\langle \begin{bmatrix}\mathbf{I} & 0\\ 0 & \mu\cdot\mathbf{H}^{-1}\end{bmatrix}, \otimes_2\left(\begin{bmatrix}\mathbf{x}^+-\mathbf{x}^*\\ \mathbf{v}^+-\mathbf{x}^*\end{bmatrix}\right)\right\rangle \le \left(1 - c_3\frac{c_2\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\left\langle \begin{bmatrix}\mathbf{I} & 0\\ 0 & \mu\cdot\mathbf{H}^{-1}\end{bmatrix}, \otimes_2\left(\begin{bmatrix}\mathbf{x}-\mathbf{x}^*\\ \mathbf{v}-\mathbf{x}^*\end{bmatrix}\right)\right\rangle \tag{27}$$

To establish this result, let us define two quantities: $e \stackrel{\text{def}}{=} \|\mathbf{x}-\mathbf{x}^*\|_2^2$, $f \stackrel{\text{def}}{=} \|\mathbf{v}-\mathbf{x}^*\|_{\mathbf{H}^{-1}}^2$ and similarly, $e^+ \stackrel{\text{def}}{=} \|\mathbf{x}^+-\mathbf{x}^*\|_2^2$ and $f^+ \stackrel{\text{def}}{=} \|\mathbf{v}^+-\mathbf{x}^*\|_{\mathbf{H}^{-1}}^2$. The potential function we consider is $e + \mu\cdot f$. Recall that the parameters are chosen as:

$$\alpha = \frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2\sqrt{2c_1-c_1^2}+\sqrt{\kappa\widetilde{\kappa}}},\quad \beta = c_3\frac{c_2\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}},\quad \gamma = c_2\frac{\sqrt{2c_1-c_1^2}}{\mu\sqrt{\kappa\widetilde{\kappa}}},\quad \delta = \frac{c_1}{R^2}$$

with $c_1 < 1/2$, $c_3 = \frac{c_2\sqrt{2c_1-c_1^2}}{c_1}$, $c_2^2 = \frac{c_4}{2-c_1}$. Consider e^+ and employ the simple gradient descent bound:

$$\begin{aligned}e^+ = \mathbb{E}\left[\|\mathbf{x}^+-\mathbf{x}^*\|_2^2\right] &= \mathbb{E}\left[\left\|\mathbf{y}-\delta\cdot\widehat{\mathbf{H}}(\mathbf{y}-\mathbf{x}^*)-\mathbf{x}^*\right\|_2^2\right]\\ &= \mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_2^2\right] - 2\delta\cdot\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_{\mathbf{H}}^2\right] + \delta^2\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_{\mathcal{M}\mathbf{I}}^2\right]\\ &\le \mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_2^2\right] - 2\delta\cdot\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_{\mathbf{H}}^2\right] + R^2\delta^2\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_{\mathbf{H}}^2\right]\\ &= \mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_2^2\right] - \frac{2c_1-c_1^2}{R^2}\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_{\mathbf{H}}^2\right]\end{aligned} \tag{28}$$

Next, consider f^+:

$$f^+ = \mathbb{E}\left[\left\|\mathbf{v}^+ - \mathbf{x}^*\right\|^2_{\mathbf{H}^{-1}}\right] = \mathbb{E}\left[\left\|\mathbf{z} - \gamma\widehat{\mathbf{H}}(\mathbf{y}-\mathbf{x}^*) - \mathbf{x}^*\right\|^2_{\mathbf{H}^{-1}}\right]$$
$$= \mathbb{E}\left[\|\mathbf{z}-\mathbf{x}^*\|^2_{\mathbf{H}^{-1}}\right] + \gamma^2\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_{\mathcal{M}\mathbf{H}^{-1}}\right] - 2\gamma\mathbb{E}\left[\langle \mathbf{z}-\mathbf{x}^*, \mathbf{y}-\mathbf{x}^*\rangle\right]$$
$$\leq \mathbb{E}\left[\|\mathbf{z}-\mathbf{x}^*\|^2_{\mathbf{H}^{-1}}\right] + \gamma^2\widetilde{\kappa}\cdot\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_{\mathbf{H}}\right] - 2\gamma\cdot\mathbb{E}\left[\langle \mathbf{z}-\mathbf{x}^*, \mathbf{y}-\mathbf{x}^*\rangle\right] \tag{29}$$

Where, we use the fact that $\mathcal{M}\mathbf{H}^{-1} \preceq \widetilde{\kappa}\mathbf{H}$, where $\widetilde{\kappa}$ is the *statistical* condition number. Consider $\mathbb{E}\left[\|\mathbf{z}-\mathbf{x}^*\|^2_{\mathbf{H}^{-1}}\right]$ and use convexity of the weighted $2-$norm to get:

$$\mathbb{E}\left[\|\mathbf{z}-\mathbf{x}^*\|^2_{\mathbf{H}^{-1}}\right] \leq \beta\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_{\mathbf{H}^{-1}}\right] + (1-\beta)\mathbb{E}\left[\|\mathbf{v}-\mathbf{x}^*\|^2_{\mathbf{H}^{-1}}\right]$$
$$\leq \frac{\beta}{\mu}\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] + (1-\beta)\cdot f \tag{30}$$

Next, consider $\mathbb{E}\left[\langle \mathbf{z}-\mathbf{x}^*, \mathbf{y}-\mathbf{x}^*\rangle\right]$, and first write $\mathbf{z}$ in terms of $\mathbf{x}$ and $\mathbf{y}$. This can be seen as two steps:

- $\mathbf{v} = \frac{1}{1-\alpha}\cdot\mathbf{y} - \frac{\alpha}{1-\alpha}\cdot\mathbf{x}$
- $\mathbf{z} = \beta\mathbf{y} + (1-\beta)\mathbf{v} = \mathbf{y} + (1-\beta)(\mathbf{v}-\mathbf{y})$. Then substituting $\mathbf{v}$ in terms of $\mathbf{x}$ and $\mathbf{y}$ as in the equation above, we get: $\mathbf{z} = \mathbf{y} + \left(\frac{\alpha\cdot(1-\beta)}{1-\alpha}\right)(\mathbf{y}-\mathbf{x})$

Then, $\mathbb{E}\left[\langle \mathbf{z}-\mathbf{x}^*, \mathbf{y}-\mathbf{x}^*\rangle\right]$ can be written as:

$$\mathbb{E}\left[\langle \mathbf{z}-\mathbf{x}^*, \mathbf{y}-\mathbf{x}^*\rangle\right] = \mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] + \left(\frac{\alpha(1-\beta)}{1-\alpha}\right)\mathbb{E}\left[\langle \mathbf{y}-\mathbf{x}, \mathbf{y}-\mathbf{x}^*\rangle\right] \tag{31}$$

Then, we note:

$$\mathbb{E}\left[\langle \mathbf{y}-\mathbf{x}, \mathbf{y}-\mathbf{x}^*\rangle\right] = \mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] - \mathbb{E}\left[\langle \mathbf{x}-\mathbf{x}^*, \mathbf{y}-\mathbf{x}^*\rangle\right]$$
$$\geq \mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] - \frac{1}{2}\cdot\left(\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] + \mathbb{E}\left[\|\mathbf{x}-\mathbf{x}^*\|^2_2\right]\right)$$
$$= \frac{1}{2}\cdot\left(\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] - \mathbb{E}\left[\|\mathbf{x}-\mathbf{x}^*\|^2_2\right]\right)$$

Re-substituting in equation 31:

$$\mathbb{E}\left[\langle \mathbf{z}-\mathbf{x}^*, \mathbf{y}-\mathbf{x}^*\rangle\right] \geq \left(1 + \frac{1}{2}\cdot\frac{\alpha(1-\beta)}{1-\alpha}\right)\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] - \frac{1}{2}\cdot\frac{\alpha(1-\beta)}{1-\alpha}\mathbb{E}\left[\|\mathbf{x}-\mathbf{x}^*\|^2_2\right]$$
$$= \left(1 + \frac{1}{2}\cdot\frac{\alpha(1-\beta)}{1-\alpha}\right)\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] - \frac{1}{2}\cdot\frac{\alpha(1-\beta)}{1-\alpha}\cdot e \tag{32}$$

Substituting equations 30, 32 into equation 29, we get:

$$\mu\cdot f^+ \leq \left(\beta - 2\gamma\mu - \frac{\gamma\mu\alpha(1-\beta)}{1-\alpha}\right)\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|^2_2\right] + \mu(1-\beta)\cdot f$$

$$+\frac{\gamma\mu\alpha(1-\beta)}{1-\alpha}\cdot e+\mu\gamma^2\widetilde{\kappa}\cdot\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_{\mathbf{H}}^2\right]$$

Rewriting the guarantee on e^+ as in equation 28:

$$e^+\leq\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_2^2\right]-\frac{2c_1-c_1^2}{R^2}\cdot\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_{\mathbf{H}}^2\right]$$

By considering $e^++\mu\cdot f^+$, we see the following:

- The coefficient of $\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_{\mathbf{H}}^2\right]\leq 0$ by setting $\gamma=c_2\frac{\sqrt{2c_1-c_1^2}}{\mu\sqrt{\kappa\widetilde{\kappa}}}$, where, $0<c_2\leq 1$, $\kappa=\frac{R^2}{\mu}$.
- Set $\frac{\gamma\mu\alpha}{1-\alpha}=1$ implying $\alpha=\frac{1}{1+\gamma\mu}=\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2\sqrt{2c_1-c_1^2}+\sqrt{\kappa\widetilde{\kappa}}}$

With these in place, we have the final result:

$$e^++\mu\cdot f^+\leq(2\beta-2\gamma\mu)\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}^*\|_2^2\right]+(1-\beta)\cdot(e+\mu\cdot f)$$

In particular, setting $\beta=c_3\gamma\mu=c_3\frac{c_2\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}$, we have a per-step contraction of $1-\beta$ which is precisely $1-c_3\frac{c_2\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}$, from which the claimed result naturally follows by substituting the values of c_1,c_2,c_3. ■

Lemma 13 *For any psd matrix $\mathbf{Q}\succeq 0$, we have:*

$$\left\|\mathcal{B}^k\mathbf{Q}\right\|_2\leq\frac{4\kappa}{\sqrt{1-\alpha^2}}\left(1-\left(\frac{c_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\right)^k\|\mathbf{Q}\|_2.$$

Proof From Lemma 4, we conclude that $\left\langle\mathbf{G},\mathcal{B}^k\mathbf{Q}\right\rangle\leq\left(1-\left(\frac{c_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\right)^k\langle\mathbf{G},\mathbf{Q}\rangle$. This implies that $\left\|\mathcal{B}^k\mathbf{Q}\right\|_2\leq\left(1-\left(\frac{c_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\right)^k\|\mathbf{Q}\|_2\,\kappa(\mathbf{G})$. Plugging the bound on $\kappa(\mathbf{G})$ from Lemma 12 proves the lemma. ■

Lemma 14 *We have:*

$$\begin{aligned}&(\mathbf{I}-\mathcal{D})\,(\mathbf{I}-\mathcal{B})^{-1}\mathcal{B}^{t+1}\left(\mathbf{I}-\mathcal{B}^{n-t}\right)\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\\&\preceq\frac{4\kappa}{\sqrt{1-\alpha^2}}\exp\left(-tc_2c_3\sqrt{2c_1-c_1^2}/\sqrt{\kappa\widetilde{\kappa}}\right)\|\boldsymbol{\theta}_0\|^2\left(\mathbf{I}+\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2c_3\sqrt{2c_1-c_1^2}}(R^2/\sigma^2)\widehat{\boldsymbol{\Sigma}}\right).\end{aligned}$$

Proof The proof follows from Lemma 4. Since $\mathcal{B}=\mathcal{D}+\mathcal{R}$, we have $(\mathcal{I}-\mathcal{D})\,(\mathcal{I}-\mathcal{B})^{-1}=\mathcal{I}+\mathcal{R}(\mathcal{I}-\mathcal{B})^{-1}$. Since $\mathcal{R},\mathcal{B}$ and $(\mathcal{I}-\mathcal{B})^{-1}$ are all PSD operators, we have

$$(\mathcal{I}-\mathcal{D})\,(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{t+1}\left(\mathcal{I}-\mathcal{B}^{n-t}\right)\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$$

$$= \left(\mathcal{I} + \mathcal{R}(\mathcal{I}-\mathcal{B})^{-1}\right)\mathcal{B}^{t+1}\left(\mathcal{I}-\mathcal{B}^{n-t}\right)\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$$
$$\preceq \underbrace{\mathcal{B}^{t+1}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top}_{\mathbf{S}_1 \stackrel{\text{def}}{=}} + \underbrace{\mathcal{R}(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{t+1}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top}_{\mathbf{S}_2 \stackrel{\text{def}}{=}}.$$

Applying Lemma 13 with $\mathbf{Q} = \boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$ tells us that $\mathbf{S}_1 \preceq \frac{4\kappa}{\sqrt{1-\alpha^2}} \exp\left(-tc_2c_3\sqrt{2c_1-c_1^2}/\sqrt{\kappa\widetilde{\kappa}}\right)\|\boldsymbol{\theta}_0\|_2^2\,\mathbf{I}$. For $\mathbf{S}_2$, we have

$$\left\langle \mathbf{G}, (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{t+1}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top \right\rangle = \left\langle \mathbf{G}, \sum_{j=t+1}^{\infty} \mathcal{B}^j\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top \right\rangle$$
$$\leq \sum_{j=t+1}^{\infty}\left(1-\left(\frac{c_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\right)^j \left\langle \mathbf{G}, \boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\right\rangle$$
$$\leq \frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2c_3\sqrt{2c_1-c_1^2}} \exp\left(-tc_2c_3\sqrt{2c_1-c_1^2}/\sqrt{4\kappa\widetilde{\kappa}}\right)\left\langle \mathbf{G}, \boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\right\rangle.$$

This implies

$$(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{t+1}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top \preceq \kappa(\mathbf{G})(\sqrt{\kappa\widetilde{\kappa}}/(c_2c_3\sqrt{2c_1-c_1^2}))\exp\left(-tc_2c_3\sqrt{2c_1-c_1^2}/\sqrt{4\kappa\widetilde{\kappa}}\right)\|\boldsymbol{\theta}_0\|^2\,\mathbf{I},$$

which tells us that

$$\mathbf{S}_2 \preceq \kappa(\mathbf{G})(\sqrt{\kappa\widetilde{\kappa}}/(c_2c_3\sqrt{2c_1-c_1^2}))\exp\left(-tc_2c_3\sqrt{2c_1-c_1^2}/\sqrt{4\kappa\widetilde{\kappa}}\right)\|\boldsymbol{\theta}_0\|^2\,(R^2/\sigma^2)\widehat{\boldsymbol{\Sigma}}$$

Combining the bounds on $\mathbf{S}_1$ and $\mathbf{S}_2$, we obtain

$$(\mathcal{I}-\mathcal{D})\,(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{t+1}\left(\mathcal{I}-\mathcal{B}^{n-t}\right)\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$$
$$\preceq \kappa(\mathbf{G})\exp\left(-tc_2c_3\sqrt{2c_1-c_1^2}/\sqrt{4\kappa\widetilde{\kappa}}\right)\|\boldsymbol{\theta}_0\|^2\left(\mathbf{I}+\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2c_3\sqrt{2c_1-c_1^2}}(R^2/\sigma^2)\widehat{\boldsymbol{\Sigma}}\right).$$

Plugging the bound for $\kappa(\mathbf{G})$ from Lemma 12 finishes the proof. ■

Corollary 15 *For any psd matrix* $\mathbf{Q} \succeq 0$*, we have:*

$$\left\|\mathbf{A}^{n+1-j}\mathcal{B}^j\mathbf{Q}\right\| \leq \frac{12\sqrt{2}(n+1-j)\kappa}{\sqrt{1-\alpha^2}}\alpha^{\frac{n-j}{2}}\left(1-\frac{c_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)^j\|\mathbf{Q}\|_2$$
$$\leq \frac{12\sqrt{2}(n+1-j)\kappa}{\sqrt{1-\alpha^2}}\alpha^{\frac{n-j}{2}}\exp\left(\frac{-jc_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\|\mathbf{Q}\|_2.$$

Proof This corollary follows directly from Lemmas 10 and 13 and using the fact that $1-x \leq e^{-x}$ ■

The following lemma bounds the total error of $\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}$.

Lemma 16

$$\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{bias} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{bias}\right] \right\rangle \leq C \cdot \frac{(\kappa\widetilde{\kappa})^{9/4} d\kappa}{(n-t)^2} \cdot \exp\left(-(t+1)\frac{c_2 c_3 \sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}} \right) \cdot \big(P(\mathbf{x}_0) - P(\mathbf{x}^*)\big)$$
$$+ C \cdot (\kappa\widetilde{\kappa})^{5/4} d\kappa \cdot \exp\left(\frac{-nc_2 c_3 \sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}} \right) \cdot \big(P(\mathbf{x}_0) - P(\mathbf{x}^*)\big)$$

Where, C is a universal constant.

Proof Lemma 3 tells us that

$$\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}\right] = \frac{1}{(n-t)^2}\Big(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\Big)(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\,(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0)$$
$$- \frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\Big((\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j}\Big)\mathcal{B}^j\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0. \tag{33}$$

We now use lemmas in this section to bound inner product of the two terms in the above expression with $\begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}$, i.e. we seek to bound,

$$\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}\right] \right\rangle$$
$$= \left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \frac{1}{(n-t)^2}\Big(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\Big)(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\,(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0) \right\rangle$$
$$+ \left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, -\frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\Big((\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j}\Big)\mathcal{B}^j\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0 \right\rangle \tag{34}$$

For the first term of equation 34, we have

$$\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \Big(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\Big)(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\,(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0) \right\rangle$$
$$= \Big\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \Big(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\Big)\Big(\mathcal{I} - \mathcal{A}_{\mathcal{L}}\mathcal{A}_{\mathcal{R}}^{\top}\Big)^{-1}\Big(\mathcal{I} - \mathcal{A}_{\mathcal{L}}\mathcal{A}_{\mathcal{R}}^{\top}\Big)$$
$$(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\,(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0)\Big\rangle$$
$$= \left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}(\mathcal{I} - \mathcal{A}_{\mathcal{R}}^{\top})^{-1}\Big(\mathcal{I} - \mathcal{A}_{\mathcal{L}}\mathcal{A}_{\mathcal{R}}^{\top}\Big)(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\,(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0) \right\rangle$$
(using Lemma 11)
$$= \left\langle (\mathbf{I} - \mathbf{A}^{\top})^{-1}\begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}(\mathbf{I} - \mathbf{A})^{-1}, (\mathcal{I} - \mathcal{D})\,(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\,(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0) \right\rangle$$
$$\leq \frac{1}{(q - c\delta)^2}\frac{4\kappa}{\sqrt{1-\alpha^2}}\exp\left(-(t+1)c_2 c_3\sqrt{2c_1 - c_1^2}/\sqrt{\kappa\widetilde{\kappa}}\right)\|\boldsymbol{\theta}_0\|^2$$

$$\left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I} - q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I} - \delta\mathbf{H})\mathbf{H}^{-1/2} \end{bmatrix} \right), \mathbf{I} + 2\sqrt{\kappa\widetilde{\kappa}}(R^2/\sigma^2)\widehat{\mathbf{\Sigma}} \right\rangle.$$

The two terms above can be bounded as

$$\left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I} - q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I} - \delta\mathbf{H})\mathbf{H}^{-1/2} \end{bmatrix} \right), \mathbf{I} \right\rangle \leq 7 \cdot \mathrm{Tr}\left(\mathbf{H}^{-1}\right) \leq \frac{7d}{\mu} \text{ and,}$$
$$2\sqrt{\kappa\widetilde{\kappa}}(R^2/\sigma^2) \left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I} - q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I} - \delta\mathbf{H})\mathbf{H}^{-1/2} \end{bmatrix} \right), \widehat{\mathbf{\Sigma}} \right\rangle = 2\sqrt{\kappa\widetilde{\kappa}}R^2(q - c\delta)^2 d.$$

Combining the above and noting the fact that $2\sqrt{\kappa\widetilde{\kappa}}R^2(q-c\delta)^2 d < \frac{7d}{\mu}$, we have

$$\begin{aligned} &\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \left(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathcal{A}_{\mathcal{R}}^\top\right)(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\left(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0\right) \right\rangle \\ &\quad \leq \frac{56\kappa d}{\sqrt{1-\alpha^2}} \cdot \frac{\|\boldsymbol{\theta}_0\|^2}{\mu\left(q - c\delta\right)^2} \cdot \exp\left(-(t+1)c_2 c_3\sqrt{2c_1 - c_1^2}/\sqrt{\kappa\widetilde{\kappa}}\right). \end{aligned} \tag{35}$$

We now note the following facts:

$$\frac{1}{1-\alpha} = \frac{c_2\sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}} + c_2\sqrt{2c_1 - c_1^2}} \leq \frac{2}{\sqrt{c_1 c_4}} \cdot \sqrt{\kappa\widetilde{\kappa}}$$
$$\frac{1}{q - c\delta} \leq \frac{1}{\gamma(1-\alpha)} \leq \frac{\mu}{(1-\alpha)^2} \leq \frac{4\widetilde{\kappa}}{c_4\delta}$$

This implies, equation 35 can be bounded as:

$$\begin{aligned} &\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \left(\mathcal{I} + (\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathcal{A}_{\mathcal{R}}^\top\right)(\mathcal{I} - \mathcal{B})^{-1}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\left(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0\right) \right\rangle \\ &\quad \leq \frac{1792}{(c_1c_4)^{5/4}} \cdot \frac{(\kappa\widetilde{\kappa})^{9/4} d}{\delta c_4} \cdot \exp\left(- (t+1)\frac{c_2c_3\sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right) \|\boldsymbol{\theta}_0\|^2 \\ &\quad \leq \frac{1792}{(c_1c_4)^{5/4}} \cdot \frac{(\kappa\widetilde{\kappa})^{9/4} d\kappa}{c_1 c_4} \cdot \exp\left(- (t+1)\frac{c_2c_3\sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right) \mu\|\boldsymbol{\theta}_0\|^2 \\ &\quad \leq \frac{3584}{(c_1c_4)^{5/4}} \cdot \frac{(\kappa\widetilde{\kappa})^{9/4} d\kappa}{c_1 c_4} \cdot \exp\left(- (t+1)\frac{c_2c_3\sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right) \cdot \left(P(\mathbf{x}_0) - P(\mathbf{x}^*)\right) \\ &\quad \leq C \cdot (\kappa\widetilde{\kappa})^{9/4} d\kappa \cdot \exp\left(- (t+1)\frac{c_2c_3\sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right) \cdot \left(P(\mathbf{x}_0) - P(\mathbf{x}^*)\right). \end{aligned} \tag{36}$$

Where, C is a universal constant.

Consider now a term in the summation in the second term of (34).

$$\begin{aligned} &\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \left((\mathcal{I} - \mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j} + (\mathcal{I} - \mathcal{A}_{\mathcal{R}}^\top)^{-1}(\mathcal{A}_{\mathcal{R}}^\top)^{n+1-j}\right)\mathcal{B}^j\left(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0\right) \right\rangle \\ &= \left\langle (\mathbf{I} - \mathbf{A}^\top)^{-1} \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{A}^{n+1-j}\mathcal{B}^j\left(\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0\right) \right\rangle \end{aligned}$$

$$
\begin{aligned}
&+\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}(\mathbf{I}-\mathbf{A})^{-1}, \Big(\mathcal{B}^j\left(\boldsymbol{\theta}_0\otimes\boldsymbol{\theta}_0\right)\Big)(\mathbf{A}^\top)^{n+1-j}\right\rangle\\
&\leq 4d\left\|(\mathbf{I}-\mathbf{A}^\top)^{-1}\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}\right\|\left\|\mathbf{A}^{n+1-j}\mathcal{B}^j\left(\boldsymbol{\theta}_0\otimes\boldsymbol{\theta}_0\right)\right\|\\
&\leq \frac{4d}{q-c\delta}\left\|\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H}) & 0\\ (\mathbf{I}-\delta\mathbf{H}) & 0\end{bmatrix}\right\|\cdot\frac{12\sqrt{2}(n+1-j)\kappa}{\sqrt{1-\alpha^2}}\alpha^{\frac{n-j}{2}}\exp\left(\frac{-jc_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\|\boldsymbol{\theta}_0\|^2\\
&\qquad\qquad\qquad\qquad\text{(Lemma 8 and Corollary 15)}\\
&\leq \frac{672(n-t)d\kappa}{(q-c\delta)\sqrt{1-\alpha^2}}\cdot\exp\left(\frac{-nc_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\|\boldsymbol{\theta}_0\|^2\\
&\leq \frac{5376}{(c_1c_4)^{1/4}}\frac{(\kappa\widetilde{\kappa})^{5/4}d}{\delta c_4}(n-t)\exp\left(\frac{-nc_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\|\boldsymbol{\theta}_0\|^2\\
&\leq \frac{5376}{(c_1c_4)^{1/4}}\frac{(\kappa\widetilde{\kappa})^{5/4}d\kappa}{c_1c_4}(n-t)\exp\left(\frac{-nc_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\mu\|\boldsymbol{\theta}_0\|^2\\
&\leq \frac{10752}{(c_1c_4)^{1/4}}\frac{(\kappa\widetilde{\kappa})^{5/4}d\kappa}{c_1c_4}(n-t)\exp\left(\frac{-nc_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\big(P(\mathbf{x}_0)-P(\mathbf{x}^*)\big)\\
&\leq C\cdot(\kappa\widetilde{\kappa})^{5/4}d\kappa\cdot(n-t)\exp\left(\frac{-nc_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\big(P(\mathbf{x}_0)-P(\mathbf{x}^*)\big). \qquad (37)
\end{aligned}
$$

Where, C is a universal constant. Plugging (36) and (37) into (34), we obtain

$$
\begin{aligned}
&\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}\otimes\bar{\boldsymbol{\theta}}_{t,n}^{\text{bias}}\right]\right\rangle\\
&\leq C\cdot\frac{(\kappa\widetilde{\kappa})^{9/4}d\kappa}{(n-t)^2}\cdot\exp\left(-(t+1)\frac{c_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\big(P(\mathbf{x}_0)-P(\mathbf{x}^*)\big)\\
&\qquad\qquad + C\cdot(\kappa\widetilde{\kappa})^{5/4}d\kappa\cdot\exp\left(\frac{-nc_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\big(P(\mathbf{x}_0)-P(\mathbf{x}^*)\big)
\end{aligned}
$$

This proves the lemma. ∎

Appendix E. Lemmas and proofs for Bounding variance error

Before we prove lemma 6, we recall old notation and introduce new notations that will be employed in these proofs.

E.1. Notations

We begin with by recalling that we track $\boldsymbol{\theta}_k = \begin{bmatrix}\mathbf{x}_k-\mathbf{x}^*\\ \mathbf{y}_k-\mathbf{x}^*\end{bmatrix}$. Given $\boldsymbol{\theta}_k$, we recall the recursion governing the evolution of $\boldsymbol{\theta}_k$:

$$
\boldsymbol{\theta}_{k+1} = \begin{bmatrix}0 & \mathbf{I}-\delta\widehat{\mathbf{H}}_{k+1}\\ -c\cdot\mathbf{I} & (1+c)\mathbf{I}-q\cdot\widehat{\mathbf{H}}_{k+1}\end{bmatrix}\boldsymbol{\theta}_k + \begin{bmatrix}\delta\cdot\epsilon_{k+1}\mathbf{a}_{k+1}\\ q\cdot\epsilon_{k+1}\mathbf{a}_{k+1}\end{bmatrix}
$$

$$= \widehat{\mathbf{A}}_{k+1}\boldsymbol{\theta}_k + \boldsymbol{\zeta}_{k+1} \tag{38}$$

where, recall, $c = \alpha(1-\beta),\ q = \alpha\delta + (1-\alpha)\gamma$, and $\widehat{\mathbf{H}}_{k+1} = \mathbf{a}_{k+1}\mathbf{a}_{k+1}^\top$. Furthermore, we recall the following definitions, which will be heavily used in the following proofs:

$$\mathbf{A} = \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1}|\mathcal{F}_k\right]$$
$$\mathcal{B} = \mathbb{E}\left[\widehat{\mathbf{A}}_{k+1} \otimes \widehat{\mathbf{A}}_{k+1}|\mathcal{F}_k\right]$$
$$\widehat{\mathbf{\Sigma}} = \mathbb{E}\left[\boldsymbol{\zeta}_{k+1} \otimes \boldsymbol{\zeta}_{k+1}|\mathcal{F}_k\right] = \begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix} \otimes \mathbf{\Sigma} \preceq \sigma^2 \cdot \begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix} \otimes \mathbf{H}$$

We recall:

$$\mathcal{R} = \mathbb{E}\left[(\mathbf{A} - \widehat{\mathbf{A}}_{k+1}) \otimes (\mathbf{A} - \widehat{\mathbf{A}}_{k+1})|\mathcal{F}_k\right]$$
$$\mathcal{D} = \mathbf{A} \otimes \mathbf{A}$$

And the operators $\mathcal{B}, \mathcal{D}, \mathcal{R}$ being related by:

$$\mathcal{B} = \mathcal{D} + \mathcal{R}$$

Furthermore, in order to compute the steady state distribution with the fourth moment quantities in the mix, we need to rely on the following re-parameterization of the update matrix $\widehat{\mathbf{A}}$:

$$\begin{aligned}\widehat{\mathbf{A}} &= \begin{bmatrix}0 & \mathbf{I} - \delta\widehat{\mathbf{H}}\\ -c\cdot\mathbf{I} & (1+c)\cdot\mathbf{I} - q\cdot\widehat{\mathbf{H}}\end{bmatrix}\\ &= \begin{bmatrix}0 & \mathbf{I}\\ -c\cdot\mathbf{I} & (1+c)\cdot\mathbf{I}\end{bmatrix} + \begin{bmatrix}0 & -\delta\cdot\widehat{\mathbf{H}}\\ 0 & -q\cdot\widehat{\mathbf{H}}\end{bmatrix}\\ &\stackrel{\text{def}}{=} \mathbf{V}_1 + \hat{\mathbf{V}}_2\end{aligned}$$

This implies in particular:

$$\begin{aligned}\widehat{\mathbf{A}} \otimes \widehat{\mathbf{A}} &= (\mathbf{V}_1 + \hat{\mathbf{V}}_2) \otimes (\mathbf{V}_1 + \hat{\mathbf{V}}_2)\\ &= \mathbf{V}_1 \otimes \mathbf{V}_1 + \mathbf{V}_1 \otimes \hat{\mathbf{V}}_2 + \hat{\mathbf{V}}_2 \otimes \mathbf{V}_1 + \hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\end{aligned}$$

Note in particular, the fourth moment part resides in the operator $\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2$. Terms such as $\mathbf{V}_1 \otimes \mathbf{V}_1$ are deterministic, or terms such as $\mathbf{V}_1 \otimes \hat{\mathbf{V}}_2$ or $\hat{\mathbf{V}}_2 \otimes \mathbf{V}_1$ contain second moment quantities. Furthermore, note that the operator $\mathcal{B} = \mathbb{E}\left[\widehat{\mathbf{A}} \otimes \widehat{\mathbf{A}}\right]$ where the expectation is taken with respect to a single random draw from the distribution $\mathcal{D}$.

Considering the expectation of $\widehat{\mathbf{A}} \otimes \widehat{\mathbf{A}}$ with respect to a single draw from the distribution $\mathcal{D}$, we have:

$$\begin{aligned}\mathcal{B} = \mathbb{E}\left[\widehat{\mathbf{A}} \otimes \widehat{\mathbf{A}}\right] &= \mathbf{V}_1 \otimes \mathbf{V}_1 + \mathbb{E}\left[\mathbf{V}_1 \otimes \hat{\mathbf{V}}_2\right] + \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \mathbf{V}_1\right] + \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right]\\ &= \mathbf{V}_1 \otimes \mathbf{V}_1 + \mathbf{V}_1 \otimes \mathbf{V}_2 + \mathbf{V}_2 \otimes \mathbf{V}_1 + \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right],\end{aligned}$$

where $\mathbf{V}_2 \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{\mathbf{V}}_2\right] = \begin{bmatrix}0 & -\delta\cdot\mathbf{H}\\ 0 & -q\cdot\mathbf{H}\end{bmatrix}$.

Finally, we let nr and dr to denote the numerator and denominator respectively.

E.2. An exact expression for the stationary distribution

Note that a key term appearing in the expression for covariance of the variance equation (24) is $(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}$. This is in fact nothing but the covariance of the error when we run accelerated SGD forever starting at $\mathbf{x}^*$ (i.e., at steady state). This can be seen by considering the base variance recursion using equation (38):

$$
\begin{aligned}
&\boldsymbol{\theta}_k = \widehat{\mathbf{A}}_k\boldsymbol{\theta}_{k-1} + \boldsymbol{\zeta}_k\\
\implies\ &\mathbf{\Phi}_k \stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\theta}_k\otimes\boldsymbol{\theta}_k\right]\\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\left(\widehat{\mathbf{A}}_k\boldsymbol{\theta}_{k-1}\otimes\boldsymbol{\theta}_{k-1}\widehat{\mathbf{A}}_k^\top + \boldsymbol{\zeta}_k\otimes\boldsymbol{\zeta}_k\right)|\mathcal{F}_{k-1}\right]\right]\\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\left(\widehat{\mathbf{A}}_k\boldsymbol{\theta}_{k-1}\otimes\boldsymbol{\theta}_{k-1}\widehat{\mathbf{A}}_k^\top\right)|\mathcal{F}_{k-1}\right]\right] + \widehat{\mathbf{\Sigma}}\\
&\quad = \mathcal{B}\cdot\mathbb{E}\left[\boldsymbol{\theta}_{k-1}\otimes\boldsymbol{\theta}_{k-1}\right] + \widehat{\mathbf{\Sigma}}\\
&\quad = \mathcal{B}\cdot\mathbf{\Phi}_{k-1} + \widehat{\mathbf{\Sigma}}
\end{aligned}
$$

This recursion on the covariance operator $\mathbf{\Phi}_k$ can be unrolled until the start i.e. $k=0$ to yield:

$$
\begin{aligned}
\mathbf{\Phi}_k &= \mathcal{B}^k\mathbf{\Phi}_0 + \sum_{l=0}^{k-1}\mathcal{B}^l\cdot\widehat{\mathbf{\Sigma}}\\
&= (\mathcal{I}-\mathcal{B})^{-1}(\mathcal{I}-\mathcal{B}^k)\widehat{\mathbf{\Sigma}} \qquad (\because\ \mathbf{\Phi}_0 = 0)\\
\implies \mathbf{\Phi}_\infty &= \lim_{k\to\infty}\mathbf{\Phi}_k = (\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}} \qquad (39)
\end{aligned}
$$

E.3. Computing the steady state distribution

We now proceed to compute the stationary distribution. Recall that

$$
\begin{aligned}
&\mathcal{B} = \mathbf{V}_1\otimes\mathbf{V}_1 + \mathbf{V}_1\otimes\mathbf{V}_2 + \mathbf{V}_2\otimes\mathbf{V}_1 + \mathbb{E}\left[\hat{\mathbf{V}}_2\otimes\hat{\mathbf{V}}_2\right]\\
\implies\ &\mathcal{I}-\mathcal{B} = \left(\mathcal{I} - \mathbf{V}_1\otimes\mathbf{V}_1 - \mathbf{V}_1\otimes\mathbf{V}_2 - \mathbf{V}_2\otimes\mathbf{V}_1\right) - \mathbb{E}\left[\hat{\mathbf{V}}_2\otimes\hat{\mathbf{V}}_2\right]
\end{aligned}
$$

Where the expectation is over a single sample drawn from the distribution $\mathcal{D}$. This implies in particular,

$$
\begin{aligned}
(\mathcal{I}-\mathcal{B})^{-1} &= \left(\left(\mathcal{I} - \mathbf{V}_1\otimes\mathbf{V}_1 - \mathbf{V}_1\otimes\mathbf{V}_2 - \mathbf{V}_2\otimes\mathbf{V}_1\right) - \mathbb{E}\left[\hat{\mathbf{V}}_2\otimes\hat{\mathbf{V}}_2\right]\right)^{-1}\\
&= \sum_{k=0}^{\infty}\left(\left(\mathcal{I} - \mathbf{V}_1\otimes\mathbf{V}_1 - \mathbf{V}_1\otimes\mathbf{V}_2 - \mathbf{V}_2\otimes\mathbf{V}_1\right)^{-1}\mathbb{E}\left[\hat{\mathbf{V}}_2\otimes\hat{\mathbf{V}}_2\right]\right)^k\\
&\qquad\cdot\left(\mathcal{I} - \mathbf{V}_1\otimes\mathbf{V}_1 - \mathbf{V}_1\otimes\mathbf{V}_2 - \mathbf{V}_2\otimes\mathbf{V}_1\right)^{-1} \qquad (40)
\end{aligned}
$$

Since $\widehat{\mathbf{\Sigma}} \preceq \sigma^2\cdot\begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix}\otimes\mathbf{H}$, and $(\mathcal{I}-\mathcal{B})^{-1}$ is a PSD operator, the steady state distribution $\mathbf{\Phi}_\infty$ is bounded by:

$$
\mathbf{\Phi}_\infty = (\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}} \preceq \sigma^2(\mathcal{I}-\mathcal{B})^{-1}\left(\begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix}\otimes\mathbf{H}\right)
$$

$$= \sigma^2 \sum_{k=0}^{\infty} \left(\left(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\right)^{-1} \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right] \right)^k \cdot \left(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\right)^{-1} \left(\begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H} \right). \tag{41}$$

Note that the Taylor expansion above is guaranteed to be correct if the right hand side is finite. We will understand bounds on the steady state distribution by splitting the analysis into the following parts:

- Obtain $\mathbf{U} \stackrel{\text{def}}{=} \left(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\right)^{-1} \left(\begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H} \right)$ (in section E.3.1).
- Obtain bounds on $\mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right] \mathbf{U}$ (in section E.3.2)
- Combine the above to obtain bounds on $\mathbf{\Phi}_\infty$ (lemma 6).

Before deriving these bounds, we will present some reasoning behind the validity of the upper bounds that we derive on the stationary distribution $\mathbf{\Phi}_\infty$:

$$\begin{aligned}
\mathbf{\Phi}_\infty &= (\mathcal{I} - \mathcal{B})^{-1} \widehat{\mathbf{\Sigma}} \\
&\preceq \sigma^2 \sum_{k=0}^{\infty} \left(\left(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\right)^{-1} \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right] \right)^k \mathbf{U} \quad (***) \\
&= \sigma^2 \mathbf{U} + \sigma^2 \sum_{k=1}^{\infty} \left(\left(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\right)^{-1} \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right] \right)^k \mathbf{U} \\
&= \sigma^2 \mathbf{U} + \sigma^2 \sum_{k=0}^{\infty} \left(\left(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\right)^{-1} \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right] \right)^k \\
&\qquad \cdot \left(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\right)^{-1} \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right] \mathbf{U} \\
&= \sigma^2 \mathbf{U} + \sigma^2 (\mathcal{I} - \mathcal{B})^{-1} \cdot \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right] \mathbf{U} \qquad \text{(using equation 40)},
\end{aligned} \tag{42}$$

with $(***)$ following through using equation 41 and through the definition of $\mathbf{U}$. Now, with this in place, we clearly see that since $(\mathcal{I} - \mathcal{B})^{-1}$ and $\mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right]$ are PSD operators, we can upper bound right hand side to create valid PSD upper bounds on $\mathbf{\Phi}_\infty$. In particular, in section E.3.1, we derive with equality what $\mathbf{U}$ is, and follow that up with computation of an upper bound on $\mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right] \mathbf{U}$ in section E.3.2. Combining this will enable us to present a valid PSD upper bound on $\mathbf{\Phi}_\infty$ owing to equation 42.

E.3.1. UNDERSTANDING THE SECOND MOMENT EFFECTS

This part of the proof deals with deriving the solution to:

$$\mathbf{U} = \left(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\right)^{-1} \left(\begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H} \right)$$

This is equivalent to solving the (linear) equation:

$$\begin{aligned}
&\big(\mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1\big) \cdot \mathbf{U} = \left(\begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H} \right) \\
&\implies \mathbf{U} - \mathbf{V}_1 \mathbf{U} \mathbf{V}_1^\top - \mathbf{V}_1 \mathbf{U} \mathbf{V}_2^\top - \mathbf{V}_2 \mathbf{U} \mathbf{V}_1^\top = \left(\begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H} \right) \qquad (43)
\end{aligned}$$

Note that all the known matrices above i.e., $\mathbf{V}_1, \mathbf{V}_2$ and $\mathbf{H}$ are all diagonalizable with respect to $\mathbf{H}$, and thus, the solution of this system can be computed in each of the eigenspaces $(\lambda_j, \mathbf{u}_j)$ of $\mathbf{H}$. This implies, in reality, we deal with matrices $\mathbf{U}^{(j)}$, one corresponding to each eigenspace. However, for this section, we will neglect the superscript on $\mathbf{U}$, since it is clear from context for the purpose of this section.

$$\begin{aligned}
\mathbf{V}_1 \mathbf{U} \mathbf{V}_1^\top &= \begin{bmatrix} 0 & 1 \\ -c & 1+c \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} 0 & -c \\ 1 & 1+c \end{bmatrix} \\
&= \begin{bmatrix} u_{22} & -cu_{12} + (1+c)u_{22} \\ -cu_{12} + (1+c)u_{22} & c^2 u_{11} - 2c(1+c)u_{12} + (1+c)^2 u_{22} \end{bmatrix}
\end{aligned}$$

Next,

$$\begin{aligned}
\mathbf{V}_1 \mathbf{U} \mathbf{V}_2^\top &= \begin{bmatrix} 0 & 1 \\ -c & 1+c \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ -\delta & -q \end{bmatrix} \lambda_j \\
&= \begin{bmatrix} u_{12} & u_{22} \\ -cu_{11} + (1+c)u_{12} & -cu_{12} + (1+c)u_{22} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ -\delta & -q \end{bmatrix} \lambda_j \\
&= \begin{bmatrix} -\delta u_{22} & -q u_{22} \\ -\delta(-cu_{12} + (1+c)u_{22}) & -q(-cu_{12} + (1+c)u_{22}) \end{bmatrix} \lambda_j
\end{aligned}$$

It follows that:

$$\begin{aligned}
\mathbf{V}_2 \mathbf{U} \mathbf{V}_1^\top &= (\mathbf{V}_1 \mathbf{U} \mathbf{V}_2^\top)^\top \\
&= \begin{bmatrix} -\delta u_{22} & -\delta(-cu_{12} + (1+c)u_{22}) \\ -q u_{22} & -q(-cu_{12} + (1+c)u_{22}) \end{bmatrix} \lambda_j
\end{aligned}$$

Given all these computations, comparing the $(1,1)$ term on both sides of equation 43, we get:

$$\begin{aligned}
u_{11} - u_{22} + 2\delta\lambda_j u_{22} &= \delta^2 \lambda_j \\
u_{11} &= u_{22}(1 - 2\delta\lambda_j) + \delta^2 \lambda_j \qquad (44)
\end{aligned}$$

Next, comparing $(1,2)$ term on both sides of equation 43, we get:

$$\begin{aligned}
&u_{12} - (-cu_{12} + (1+c)u_{22}) + q\lambda_j u_{22} + \delta\lambda_j(-cu_{12} + (1+c)u_{22}) = \delta\, q\lambda_j \\
&u_{12} - (1 - \delta\lambda_j)(-cu_{12} + (1+c)u_{22}) + q\lambda_j u_{22} = \delta\, q\lambda_j \\
&(1 + c(1 - \delta\lambda_j)) \cdot u_{12} + (q\lambda_j - (1+c)(1 - \delta\lambda_j)) \cdot u_{22} = \delta\, q\lambda_j \qquad (45)
\end{aligned}$$

Finally, comparing the $(2,2)$ term on both sides of equation 43, we get:

$$u_{22} - (c^2 u_{11} - 2c(1+c)u_{12} + (1+c)^2 u_{22}) + 2q\lambda_j(-cu_{12} + (1+c)u_{22}) = q^2 \lambda_j$$

$$\implies -c^2 u_{11} + (2c(1+c) - 2cq\lambda_j)u_{12} + (1-(1+c)^2 + 2(1+c)q\lambda_j)u_{22} = q^2\lambda_j \quad \text{(from equation 44)}$$
$$\implies -c^2(u_{22}(1-2\delta\lambda_j) + \delta^2\lambda_j) + (2c(1+c) - 2cq\lambda_j)u_{12} + (1-(1+c)^2 + 2(1+c)q\lambda_j)u_{22} = q^2\lambda_j$$
$$\implies (2c(1+c) - 2cq\lambda_j)u_{12} + (1-(1+c)^2 - c^2(1-2\delta\lambda_j) + 2(1+c)q\lambda_j)u_{22} = (q^2 + c^2\delta^2)\lambda_j$$
$$\implies 2c((1+c) - q\lambda_j)u_{12} + 2((1+c)(q\lambda_j - c) + \delta\lambda_j c^2)u_{22} = (q^2 + c^2\delta^2)\lambda_j \tag{46}$$

Now, we note that equations 45, 46 are linear systems in two variables u_{12} and u_{22}. Denoting the system in the following manner,

$$a_{11}u_{12} + a_{12}u_{22} = b_1$$
$$a_{21}u_{12} + a_{22}u_{22} = b_2$$

For analyzing the variance error, we require u_{22}, u_{12}:

$$u_{22} = \frac{b_1 a_{21} - b_2 a_{11}}{a_{12}a_{21} - a_{11}a_{22}}, \; u_{12} = \frac{b_1 a_{22} - b_2 a_{12}}{a_{11}a_{22} - a_{12}a_{21}}$$

Substituting the values from equations 45 and 46, we get:

$$u_{22} = \frac{2cq\delta\Big(1+c-q\lambda_j\Big) - (q^2+c^2\delta^2)\Big(1+c(1-\delta\lambda_j)\Big)}{2c\Big(\big(1+c-q\lambda_j\big)\cdot\big(\lambda_j q - (1+c)(1-\delta\lambda_j)\big)\Big) - 2\cdot\Big(\big(1+c-c\delta\lambda_j\big)\cdot\big((1+c)(q\lambda_j - c) + \delta\lambda_j c^2\big)\Big)}\cdot\lambda_j \tag{47}$$

$$u_{12} = \frac{2q\delta\Big((1+c)(q\lambda_j - c) + \delta\lambda_j c^2\Big) - (q^2+c^2\delta^2)\Big(\lambda_j q - (1+c)(1-\delta\lambda_j)\Big)}{2\Big(\big(1+c-c\delta\lambda_j\big)\cdot\big((1+c)(q\lambda_j - c) + \delta\lambda_j c^2\big)\Big) - 2c\Big(\big(1+c-q\lambda_j\big)\cdot\big(\lambda_j q - (1+c)(1-\delta\lambda_j)\big)\Big)}\cdot\lambda_j \tag{48}$$

Denominator of u_{22}: Let us consider the denominator of u_{22} (from equation 47) to write it in a concise manner.

$$\text{dr}(u_{22}) = 2\Big(\ \big(1+c-q\lambda_j\big)\cdot k_1 \ - \ \big(1+c-c\delta\lambda_j\big)\cdot k_2\Big)$$

with

$$\begin{aligned}
k_1 &= c\cdot\big(\lambda_j q - (1+c)(1-\delta\lambda_j)\big) \\
&= \big(c\lambda_j q - (c+c^2)(1-\delta\lambda_j)\big) \\
&= \big(cq\lambda_j - c - c^2 + c\delta\lambda_j + c^2\delta\lambda_j\big) \\
k_2 &= \big((1+c)(q\lambda_j - c) + \delta\lambda_j c^2\big) \\
&= \big(q\lambda_j - c + cq\lambda_j - c^2 + \delta\lambda_j c^2\big)
\end{aligned}$$

Plugging in expressions for $q = \alpha\delta + (1-\alpha)\gamma$ and $c = \alpha(1-\beta)$, in dr(u_{22}) we get:

$$\text{dr}(u_{22}) = 2\cdot\Big(\ \big(1+c-\alpha\delta\lambda_j\big)(k_1 - k_2) - \lambda_j\cdot\big((1-\alpha)\gamma k_1 + \alpha\beta\delta k_2\big)\Big) \tag{49}$$

Next, considering $k_1 - k_2$, we have:

$$\begin{aligned} k_1 - k_2 &= c\lambda_j q - c - c^2 + c\delta\lambda_j + c^2\delta\lambda_j - q\lambda_j + c - cq\lambda_j + c^2 - c^2\delta\lambda_j \\ &= (c\delta - q)\lambda_j \\ &= -(\alpha\beta\delta + \gamma(1-\alpha))\lambda_j \end{aligned} \tag{50}$$

Next, considering $\gamma(1-\alpha)k_1 + \alpha\beta\delta\, k_2$, we have:

$$\begin{aligned} &\gamma(1-\alpha)k_1 + \alpha\beta\delta\, k_2 \\ &= \gamma(1-\alpha)(c\lambda_j q - c - c^2 + c^2\delta\lambda_j + c\delta\lambda_j) \\ &+ \alpha\beta\delta(c\lambda_j q - c - c^2 + c^2\delta\lambda_j + q\lambda_j) \\ &= (\alpha\beta\delta + (1-\alpha)\gamma)(c\lambda_j q - c - c^2 + c^2\delta\lambda_j) + \lambda_j\delta(c\gamma(1-\alpha) + \alpha\beta q) \end{aligned}$$

Consider $c\gamma(1-\alpha) + \alpha\beta q$:

$$\begin{aligned} c\gamma(1-\alpha) + \alpha\beta q &= \alpha(1-\beta)\gamma(1-\alpha) + \alpha\beta(\alpha\delta + (1-\alpha)\gamma) \\ &= \alpha(1-\beta)\gamma(1-\alpha) + \alpha\beta\gamma(1-\alpha) + \alpha^2\beta\delta \\ &= \alpha\gamma(1-\alpha) + \alpha^2\beta\delta \\ &= \alpha(\alpha\beta\delta + (1-\alpha)\gamma) \end{aligned}$$

Re-substituting this in the expression for $\gamma(1-\alpha)k_1 + \alpha\beta\delta k_2$, we have:

$$\begin{aligned} \gamma(1-\alpha)k_1 + \alpha\beta\delta\, k_2 &= (\alpha\beta\delta + (1-\alpha)\gamma)(c\lambda_j q - c - c^2 + c^2\delta\lambda_j) + \lambda_j\delta(c\gamma(1-\alpha) + \alpha\beta q) \\ &= (\alpha\beta\delta + (1-\alpha)\gamma)(c\lambda_j q - c - c^2 + c^2\delta\lambda_j) + \alpha\lambda_j\delta(\alpha\beta\delta + (1-\alpha)\gamma) \\ &= (\alpha\beta\delta + (1-\alpha)\gamma)(c\lambda_j q - c - c^2 + c^2\delta\lambda_j + \alpha\lambda_j\delta) \end{aligned} \tag{51}$$

Substituting equations 50, 51 into equation 49, we have:

$$\begin{aligned} \mathrm{dr}(u_{22}) &= -2\lambda_j(\alpha\beta\delta + \gamma(1-\alpha)) \cdot (1 + c - \alpha\delta\lambda_j + c\lambda_j q - c - c^2 + c^2\delta\lambda_j + \alpha\delta\lambda_j) \\ &= -2\lambda_j(\alpha\beta\delta + \gamma(1-\alpha)) \cdot (1 - c^2 + c\lambda_j(q + c\delta)) \end{aligned} \tag{52}$$

We note that the denominator of u_{12} (in equation 48) is just the negative of the denominator of u_{22} as represented in equation 52.

Numerator of u_{22}: We begin by writing out the numerator of u_{22} (from equation 47):

$$\begin{aligned} \mathrm{nr}(u_{22}) &= \lambda_j \cdot \Big(2cq\delta\big(1 + c - q\lambda_j\big) - (q^2 + c^2\delta^2)\big(1 + c(1 - \delta\lambda_j)\big)\Big) \\ &= \lambda_j \cdot \Big(2cq\delta\big(1 + c - \alpha\delta\lambda_j - \gamma(1-\alpha)\lambda_j\big) - (q^2 + c^2\delta^2)\big(1 + c - \alpha\delta\lambda_j + \alpha\beta\delta\lambda_j\big)\Big) \\ &= \lambda_j \cdot \Big(-(1 + c - \alpha\delta\lambda_j)(q - c\delta)^2 - \lambda_j \cdot \big(2cq\delta\gamma(1-\alpha) + (q^2 + (c\delta)^2)\alpha\beta\delta\big)\Big) \end{aligned} \tag{53}$$

We now consider $2cq\delta\gamma(1-\alpha) + (q^2 + (c\delta)^2)\alpha\beta\delta$:

$$2cq\delta\gamma(1-\alpha) + (q^2 + (c\delta)^2)\alpha\beta\delta$$

$$
\begin{aligned}
&= 2cq\delta \cdot (\gamma(1-\alpha) + \alpha\beta\delta) + (q^2 + (c\delta)^2 - 2cq\delta)\alpha\beta\delta \\
&= 2cq\delta(q - c\delta) + (q - c\delta)^2\alpha\beta\delta \qquad (54)
\end{aligned}
$$

Substituting equation 54 into equation 53 and grouping common terms, we obtain:

$$
\begin{aligned}
\text{nr}(u_{22}) &= \lambda_j \cdot \Big(-(1 + c - \alpha\delta\lambda_j)(q - c\delta)^2 - \lambda_j \cdot \big(2cq\delta(q - c\delta) + (q - c\delta)^2\alpha\beta\delta\big)\Big) \\
&= \lambda_j \cdot \Big(-(1 + c - c\delta\lambda_j)(q - c\delta)^2 - \lambda_j \cdot \big(2cq\delta(q - c\delta)\big)\Big) \\
&= -\lambda_j \cdot \Big((1 + c - c\delta\lambda_j)(q - c\delta)^2 + 2cq\delta\lambda_j(q - c\delta)\Big) \qquad (55)
\end{aligned}
$$

With this, we can write out the exact expression for u_{22}:

$$
u_{22} = \frac{\big(1 + c - c\delta\lambda_j\big)(q - c\delta) + 2cq\delta\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \qquad (56)
$$

<u>**Numerator of u_{12}**</u>: We begin by rewriting the numerator of u_{12} (from equation 48):

$$
\text{nr}(u_{12}) = \lambda_j \cdot \Big(2q\delta\big((1+c)(q\lambda_j - c) + \delta\lambda_j c^2\big) - (q^2 + c^2\delta^2)\big(\lambda_j q - (1+c)(1 - \delta\lambda_j)\big)\Big) \qquad (57)
$$

We split the simplification into two parts: one depending on $(1+c)$ and the other part representing terms that don't contain $(1+c)$. In particular, we consider the terms that do not carry a coefficient of $(1+c)$:

$$
\begin{aligned}
&2q\delta^2\lambda_j c^2 - (q^2 + c^2\delta^2) \cdot (q\lambda_j) \\
&= q\lambda_j \cdot (2\delta^2 c^2 - q^2 - \delta^2 c^2) \\
&= -q\lambda_j \cdot (q^2 - (c\delta)^2) \qquad (58)
\end{aligned}
$$

Next, we consider the other term containing the $(1+c)$ part:

$$
\begin{aligned}
&(1+c) \cdot \Big(2q\delta \cdot (q\lambda_j - c) \; + \; (q^2 + (c\delta)^2) \cdot (1 - \delta\lambda_j)\Big) \\
&= (1+c) \cdot \Big(2q^2\delta\lambda_j - 2q\delta c + q^2 + (c\delta)^2 - q^2\delta\lambda_j - c^2\delta^3\lambda_j \Big) \\
&= (1+c) \cdot \Big((q - c\delta)^2 + \delta\lambda_j\,(q^2 - (c\delta)^2) \Big) \qquad (59)
\end{aligned}
$$

Substituting equations 58, 59 into equation 57, we get:

$$
\begin{aligned}
\text{nr}(u_{12}) &= \lambda_j \cdot \big((1+c)\delta\lambda_j(q^2 - (c\delta)^2) + (1+c)(q - c\delta)^2 - q\lambda_j(q^2 - (c\delta)^2)\big) \\
&= \lambda_j \cdot \big((1+c)(q - c\delta)^2 + \lambda_j\big((1+c)\delta - q\big) \cdot (q^2 - (c\delta)^2)\big) \\
&= \lambda_j \cdot \big((1+c)(q - c\delta)^2 + \lambda_j\big(\delta - (q - c\delta)\big) \cdot (q^2 - (c\delta)^2)\big) \\
&= \lambda_j \cdot \big((1+c)(q - c\delta)^2 + \delta\lambda_j \cdot (q^2 - (c\delta)^2) - \lambda_j(q + c\delta)(q - c\delta)^2\big)
\end{aligned}
$$

$$= \lambda_j \cdot \big((1 + c - \lambda_j \cdot (q + c\delta)) \cdot (q - c\delta)^2 + \delta\lambda_j \cdot (q^2 - (c\delta)^2)\big) \tag{60}$$

With which, we can now write out the expression for u_{12}:

$$u_{12} = \frac{\big(1 + c - \lambda_j(q + c\delta)\big)(q - c\delta) + \delta\lambda_j(q + c\delta)}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \tag{61}$$

Obtaining u_{11}: We revisit equation 44 and substitute u_{22} from equation 56:

$$\begin{aligned} u_{11} &= u_{22}(1 - 2\delta\lambda_j) + \delta^2\lambda_j \\ &= \frac{\big(1 + c - c\delta\lambda_j\big)(q - c\delta) + 2cq\delta\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \cdot (1 - 2\delta\lambda_j) + \delta^2\lambda_j \end{aligned}$$

From which, we consider the numerator of u_{11} and begin simplifying it:

$$\begin{aligned} \mathrm{nr}(u_{11}) &= (1 + c - c\delta\lambda_j)(q - c\delta)(1 - 2\delta\lambda_j) + 2cq\delta\lambda_j(1 - 2\delta\lambda_j) + 2\delta^2\lambda_j(1 - c^2 + c\lambda_j(q + c\delta)) \\ &= (1 + c - c\delta\lambda_j)(q - c\delta)(1 - 2\delta\lambda_j) + 2\delta^2\lambda_j + 2c\delta\lambda_j(q - c\delta)(1 - \delta\lambda_j) \\ &= (1 + c + c\delta\lambda_j)(q - c\delta)(1 - \delta\lambda_j) + 2\delta^2\lambda_j - \delta\lambda_j(1 + c - c\delta\lambda_j)(q - c\delta) \\ &= (1 + c + c\delta\lambda_j)(q - c\delta) - 2\delta\lambda_j(q - c\delta)(1 + c) + 2\delta^2\lambda_j \\ &= (1 + c - c\delta\lambda_j)(q - c\delta) - 2\delta\lambda_j(q - c\delta) + 2\delta^2\lambda_j \end{aligned} \tag{62}$$

This implies,

$$u_{11} = \frac{(1 + c - c\delta\lambda_j)(q - c\delta) - 2\delta\lambda_j(q - c\delta) + 2\delta^2\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \tag{63}$$

Obtaining a bound on $\mathbf{U}_{22}$

For obtaining a PSD upper bound on $\mathbf{U}_{22}$, we will write out a sharp bound of u_{22} in each eigen space:

$$\begin{aligned} u_{22} &= \frac{\big(1 + c - c\lambda_j\delta\big)(q - c\delta) + 2cq\delta\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \\ &= \frac{\big(1 - c^2 + c\lambda_j(q + c\delta) + q\lambda_j + (1 + c)(c - \lambda_j(q + c\delta))\big)(q - c\delta) + 2cq\delta\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \\ &= \frac{q - c\delta}{2} + \frac{q\lambda_j(q - c\delta)}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} + \frac{(1 + c)(c - \lambda_j(q + c\delta))(q - c\delta) + 2cq\delta\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \\ &\leq \frac{q - c\delta}{2} + \frac{q\lambda_j(q - c\delta)}{2 \cdot (c\lambda_j \cdot (q + c\delta))} + \frac{(1 + c)(c - \lambda_j(q + c\delta))(q - c\delta) + 2cq\delta\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \\ &\leq \frac{q - c\delta}{2} \cdot \frac{1 + c}{c} + \frac{(1 + c)(c - \lambda_j(q + c\delta))(q - c\delta) + 2cq\delta\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \end{aligned}$$

Let us consider bounding the numerator of the 2nd term:

$$\begin{aligned} &(1 + c)(c - \lambda_j(q + c\delta))(q - c\delta) + 2cq\delta\lambda_j \\ &= c(1 + c)(q - c\delta) - (1 + c)\lambda_j(q + c\delta)(q - c\delta) + 2cq\delta\lambda_j \end{aligned}$$

$$
\begin{aligned}
&= c(1+c)(q-c\delta) - (1+c)\lambda_j(q-c\delta)^2 - 2c\delta\lambda_j(1+c)(q-c\delta) + 2cq\delta\lambda_j \\
&= c(1+c)(q-c\delta) - (1+c)\lambda_j(q-c\delta)^2 - 2c\delta\lambda_j(1+c)(q-c\delta) + 2c(q-c\delta)\delta\lambda_j + 2c^2\delta^2\lambda_j \\
&= c(1+c)(q-c\delta) + 2c^2\delta^2\lambda_j - (1+c)\lambda_j(q-c\delta)^2 - 2c^2\delta\lambda_j(q-c\delta) \\
&\leq c(1+c)(q-c\delta) + 2c^2\delta^2\lambda_j
\end{aligned}
$$

Implying,

$$
\begin{aligned}
u_{22} &\leq \frac{q-c\delta}{2}\cdot\frac{1+c}{c} + \frac{c(1+c)(q-c\delta)+2c^2\delta^2\lambda_j}{2\cdot(1-c^2+c\lambda_j\cdot(q+c\delta))} \\
&\leq \frac{q-c\delta}{2}\cdot\frac{1+c}{c} + \frac{c(1+c)(q-c\delta)}{2\cdot(1-c^2+c\lambda_j\cdot(q+c\delta))} + \frac{c^2\delta^2\lambda_j}{(1-c^2+c\lambda_j\cdot(q+c\delta))}
\end{aligned}
$$

We will first upper bound the third term:

$$
\begin{aligned}
\frac{c^2\delta^2\lambda_j}{(1-c^2+c\lambda_j\cdot(q+c\delta))} &\leq \frac{c\delta^2}{(q+c\delta)} \\
&= \frac{c\delta^2}{(q-c\delta+2c\delta)} \\
&\leq \frac{c\delta^2}{2c\delta} = \frac{\delta}{2}
\end{aligned}
$$

This implies,

$$
\begin{aligned}
u_{22} &\leq \frac{q-c\delta}{2}\cdot\frac{1+c}{c} + \frac{\delta}{2} + \frac{c(1+c)(q-c\delta)}{2\cdot(1-c^2+c\lambda_j\cdot(q+c\delta))} \\
&= \frac{q-c\delta}{2}\cdot\frac{1+c}{c} + \frac{\delta}{2} + \frac{c^2(q-c\delta)}{1-c^2+c\lambda_j\cdot(q+c\delta)} + \frac{c(1-c)(q-c\delta)}{2\cdot(1-c^2+c\lambda_j\cdot(q+c\delta))} \\
&\leq \frac{q-c\delta}{2}\cdot\frac{1+c}{c} + \frac{\delta}{2} + \frac{c^2(q-c\delta)}{1-c^2+c\lambda_j\cdot(q+c\delta)} + \frac{c(1-c)(q-c\delta)}{2\cdot(1-c^2)} \\
&= \frac{q-c\delta}{2}\cdot\frac{1+c}{c} + \frac{\delta}{2} + \frac{c^2(q-c\delta)}{1-c^2+c\lambda_j\cdot(q+c\delta)} + \frac{c(q-c\delta)}{2\cdot(1+c)} \\
&= \frac{q-c\delta}{2}\cdot\left(\frac{1+c}{c}+\frac{c}{1+c}\right) + \frac{\delta}{2} + \frac{c^2(q-c\delta)}{1-c^2+c\lambda_j\cdot(q+c\delta)} \\
&\leq \frac{q-c\delta}{2}\cdot\frac{3}{c} + \frac{\delta}{2} + \frac{c^2(q-c\delta)}{1-c^2+c\lambda_j\cdot(q+c\delta)} \\
&\leq \frac{q-c\delta}{2}\cdot\frac{3}{c} + \frac{\delta}{2} + \frac{c(q-c\delta)}{\lambda_j\cdot(q+c\delta)} \\
&= \frac{q-c\delta}{2}\cdot\frac{3}{c} + \frac{\delta}{2} + \frac{c(q-c\delta)}{\lambda_j\cdot(q-c\delta+2c\delta)} \\
&\leq \frac{q-c\delta}{2}\cdot\frac{3}{c} + \frac{\delta}{2} + \frac{q-c\delta}{2\lambda_j\delta} \\
&\leq \frac{4}{c}\cdot\frac{q-c\delta}{2\delta\lambda_j} + \frac{\delta}{2}
\end{aligned}
$$

Let us consider bounding $\frac{q-c\delta}{2\delta\lambda_j}$:

$$\frac{q-c\delta}{2\delta\lambda_j} = \frac{\alpha\beta\delta + \gamma(1-\alpha)}{2\delta\lambda_j}$$

Substituting the values for $\alpha, \beta, \gamma, \delta$ applying $\frac{1}{1+\gamma\mu} \le 1$, $c_3 = \frac{c_2\sqrt{2c_1 - c_1^2}}{c_1}$ and, $c_2^2 = \frac{c_4}{2-c_1}$ with $0 < c_4 < 1/6$ we get:

$$\begin{aligned}
\frac{q-c\delta}{2\delta\lambda_j} &\le \left(\frac{c_3c_2\sqrt{2c_1 - c_1^2}}{2}\sqrt{\frac{\widetilde{\kappa}}{\kappa}} + \frac{c_2^2(2c_1 - c_1^2)}{2c_1}\right) \cdot \frac{1}{\lambda_j\widetilde{\kappa}} \\
&\le \left(\frac{c_3c_2\sqrt{2c_1 - c_1^2}}{2} + \frac{c_2^2(2c_1 - c_1^2)}{2c_1}\right) \cdot \frac{1}{\lambda_j\widetilde{\kappa}} \\
&= c_2^2(2-c_1) \cdot \frac{1}{\widetilde{\kappa}\lambda_j} = c_4 \cdot \frac{1}{\lambda_j\widetilde{\kappa}}
\end{aligned}$$

Which implies the bound on u_{22}:

$$u_{22} \le \frac{4}{c} \cdot \frac{c_4}{\lambda_j\widetilde{\kappa}} + \frac{\delta}{2}$$

Now, consider the following bound on $1/c$:

$$\begin{aligned}
\frac{1}{c} &= \frac{1}{\alpha(1-\beta)} \\
&= 1 + \frac{(1+c_3)c_2\sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}} - c_2c_3\sqrt{2c_1 - c_1^2}} \\
&\le 1 + \frac{(1+c_3)c_2\sqrt{2c_1 - c_1^2}}{1 - c_2c_3\sqrt{2c_1 - c_1^2}} \\
&= 1 + \frac{\sqrt{c_1c_4} + c_4}{1 - c_4} \\
&= \frac{1 + \sqrt{c_1c_4}}{1 - c_4}
\end{aligned} \tag{64}$$

Substituting values of c_1, c_4 we have: $1/c \le 1.5$. This implies the following bound on u_{22}:

$$u_{22} \le 6 \cdot \frac{c_4}{\lambda_j\widetilde{\kappa}} + \frac{\delta}{2} \tag{65}$$

Alternatively, this implies that $\mathbf{U}_{22}$ can be upper bounded in a psd sense as:

$$\mathbf{U}_{22} \preceq \frac{6c_4}{\widetilde{\kappa}} \cdot \mathbf{H}^{-1} + \frac{\delta}{2} \cdot \mathbf{I}$$

E.3.2. UNDERSTANDING FOURTH MOMENT EFFECTS

We wish to obtain a bound on:

$$\mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right]\mathbf{U} = \mathbb{E}\left[\hat{\mathbf{V}}_2\mathbf{U}\hat{\mathbf{V}}_2^\top\right]$$

$$= \begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathcal{M}\mathbf{U}_{22}$$

We need to understand $\mathcal{M}\mathbf{U}_{22}$.

$$\begin{aligned}\mathcal{M}\mathbf{U}_{22} &\preceq \frac{6c_4}{\widetilde{\kappa}} \cdot \mathcal{M}\mathbf{H}^{-1} + \frac{\delta}{2} \cdot \mathcal{M}\mathbf{I} \\ &\preceq (6c_4 + \frac{\delta R^2}{2}) \cdot \mathbf{H} \\ &= s \cdot \mathbf{H} \end{aligned} \tag{66}$$

where, $s \stackrel{\text{def}}{=} (6c_4 + \frac{\delta R^2}{2}) = 23/30 \leq \frac{4}{5}$. This implies (along with the fact that for any PSD matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$, if $\mathbf{A} \preceq \mathbf{B}$, then, $\mathbf{A} \otimes \mathbf{C} \preceq \mathbf{B} \otimes \mathbf{C}$)),

$$\mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right]\mathbf{U} \preceq s \cdot \begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H} \preceq \frac{4}{5} \cdot \begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H}. \tag{67}$$

This will lead us to obtaining a PSD upper bound on $\mathbf{\Phi}_\infty$, i.e., the proof of lemma 6

Proof [Proof of lemma 6] We begin by recounting the expression for the steady state covariance operator $\mathbf{\Phi}_\infty$ and applying results derived from previous subsections:

$$\begin{aligned} \mathbf{\Phi}_\infty &= (\mathcal{I} - \mathcal{B})^{-1}\widehat{\mathbf{\Sigma}} \\ &\preceq \sigma^2\mathbf{U} + \sigma^2(\mathcal{I} - \mathcal{B})^{-1} \cdot \mathbb{E}\left[\hat{\mathbf{V}}_2 \otimes \hat{\mathbf{V}}_2\right]\mathbf{U} \quad \text{(from equation 42)} \\ &\preceq \sigma^2\mathbf{U} + \frac{4}{5}\sigma^2(\mathcal{I} - \mathcal{B})^{-1}\left(\begin{bmatrix} \delta^2 & \delta \cdot q \\ \delta \cdot q & q^2 \end{bmatrix} \otimes \mathbf{H}\right) \quad \text{(from equation 67)} \\ &= \sigma^2\mathbf{U} + \frac{4}{5}(\mathcal{I} - \mathcal{B})^{-1}\widehat{\mathbf{\Sigma}} \\ &= \sigma^2\mathbf{U} + \frac{4}{5} \cdot \mathbf{\Phi}_\infty \\ \implies \mathbf{\Phi}_\infty &\preceq 5\sigma^2\mathbf{U}. \end{aligned} \tag{68}$$

Now, given the upper bound provided by equation 68, we can now obtain a (mildly) looser upper PSD bound on $\mathbf{U}$ that is more interpretable, and this is by providing an upper bound on $\mathbf{U}_{11}$ and $\mathbf{U}_{22}$ by considering their magnitude along each eigen direction of $\mathbf{H}$. In particular, let us consider the max of u_{11} and u_{22} along the j^{th} eigen direction (as implied by equations 63, 56):

$$\begin{aligned} \max(u_{11}, u_{22}) &= \frac{(1 + c - c\delta\lambda_j)(q - c\delta) + 2\delta^2\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \\ &= \frac{(1 + c - c\delta\lambda_j)(q - c\delta) + 2\delta^2\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \\ &= \frac{(1 + c - c\delta\lambda_j)(q - c\delta) + 2cq\lambda_j - 2cq\lambda_j + 2\delta^2\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \\ &= u_{22} + \frac{-2cq\lambda_j + 2\delta^2\lambda_j}{2 \cdot (1 - c^2 + c\lambda_j \cdot (q + c\delta))} \\ &\leq \frac{6c_4}{\widetilde{\kappa}\lambda_j} + \frac{\delta}{2} + \frac{\delta^2\lambda_j - cq\lambda_j}{(1 - c^2 + c\lambda_j \cdot (q + c\delta))} \quad \text{(using equation 65)} \end{aligned}$$

This implies, we can now consider upper bounding the term in the equation above and this will yield us the result:

$$\begin{aligned}
\frac{\delta^2\lambda_j - cq\lambda_j}{(1-c^2+c\lambda_j\cdot(q+c\delta))} &\leq \frac{\delta^2\lambda_j - cq\lambda_j}{c\lambda_j\cdot(q+c\delta)} \\
&\leq \frac{\delta^2\lambda_j - cq\lambda_j}{2c^2\delta\lambda_j} \\
&= \frac{\delta^2\lambda_j - c(\alpha\delta+\gamma(1-\alpha))\lambda_j}{2c^2\delta\lambda_j} \\
&\leq \frac{\delta^2\lambda_j - c\alpha\delta\lambda_j}{2c^2\delta\lambda_j} = \frac{1-c\alpha}{c^2}\cdot\frac{\delta}{2} \\
&= \big(\frac{1-c}{c^2}+\frac{1-\alpha}{c}\big)\cdot\frac{\delta}{2} \\
&= \big(\frac{(1+c_3)(1-\alpha)}{c^2}+\frac{1-\alpha}{c}\big)\cdot\frac{\delta}{2} \\
&= \frac{1-\alpha}{c}\big(\frac{(1+c_3)}{c}+1\big)\cdot\frac{\delta}{2} \\
&\leq 3\frac{1-\alpha}{c}\cdot\frac{1}{c}\cdot\frac{\delta}{2} \\
&\leq 3\frac{1-\alpha}{c}\cdot\frac{1+\sqrt{c_1c_4}}{1-c_4}\cdot\frac{\delta}{2} \\
&= 3\cdot\frac{c_1c_3}{\sqrt{\kappa\widetilde{\kappa}}-c_1c_3^2}\cdot\frac{1+\sqrt{c_1c_4}}{1-c_4}\cdot\frac{\delta}{2} \\
&\leq 3\cdot\frac{c_1c_3}{1-c_1c_3^2}\cdot\frac{1+\sqrt{c_1c_4}}{1-c_4}\cdot\frac{\delta}{2} \\
&\leq (2/3)\frac{\delta}{2}
\end{aligned}$$

Plugging this into the bound for $\max u_{11}, u_{22}$, we get:

$$\max(u_{11},u_{22}) \leq \frac{6c_4}{\widetilde{\kappa}\lambda_j} + (5/3)\frac{\delta}{2} = (2/3)\frac{1}{\widetilde{\kappa}\lambda_j} + (5/3)\frac{\delta}{2}$$

This implies the bound written out in the lemma, that is,

$$\mathbf{U} \preceq \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \otimes \left(\frac{2}{3}\big(\frac{1}{\widetilde{\kappa}}\mathbf{H}^{-1}\big) + \frac{5}{6}\cdot(\delta\,\mathbf{I})\right)$$

■

Lemma 17

$$\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\right)\cdot\mathbb{E}\left[\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right]\right\rangle \leq$$
$$\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\right)\cdot\mathbb{E}\left[\boldsymbol{\theta}_\infty\otimes\boldsymbol{\theta}_\infty\right]\right\rangle \leq 5\sigma^2 d.$$

Where, d is the dimension of the problem.

Before proving Lemma 17, we note that the sequence of expected covariances of the centered parameters $\mathbb{E}\left[\boldsymbol{\theta}_l \otimes \boldsymbol{\theta}_l\right]$ when initialized at the zero covariance (as in the case of variance analysis) only grows (in a psd sense) as a function of time and settles at the steady state covariance.

Lemma 18 *Let $\boldsymbol{\theta}_0 = 0$. Then, by running the stochastic process defined using the recursion as in equation 38, the covariance of the resulting process is monotonically increasing until reaching the stationary covariance $\mathbb{E}\left[\boldsymbol{\theta}_\infty \otimes \boldsymbol{\theta}_\infty\right]$.*

Proof As long as the process does not diverge (as defined by spectral norm bounds of the expected update $\mathcal{B} = \mathbb{E}\left[\hat{\mathbf{A}} \otimes \hat{\mathbf{A}}\right]$ being less than 1), the first-order Markovian process converges geometrically to its unique stationary distribution $\boldsymbol{\theta}_\infty \otimes \boldsymbol{\theta}_\infty$. In particular,

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\theta}_l \otimes \boldsymbol{\theta}_l\right] &= \mathcal{B}\mathbb{E}\left[\boldsymbol{\theta}_{l-1} \otimes \boldsymbol{\theta}_{l-1}\right] + \widehat{\boldsymbol{\Sigma}} \\ &= (\sum_{k=0}^{l-1} \mathcal{B}^k)\widehat{\boldsymbol{\Sigma}}\end{aligned}$$

Thus implying the fact that

$$\mathbb{E}\left[\boldsymbol{\theta}_l \otimes \boldsymbol{\theta}_l\right] = \mathbb{E}\left[\boldsymbol{\theta}_{l-1} \otimes \boldsymbol{\theta}_{l-1}\right] + \mathcal{B}^{l-1}\widehat{\boldsymbol{\Sigma}}$$

Owing to the PSD'ness of the operators in the equation above, the lemma concludes with the claim that $\mathbb{E}\left[\boldsymbol{\theta}_l \otimes \boldsymbol{\theta}_l\right] \succeq \mathbb{E}\left[\boldsymbol{\theta}_{l-1} \otimes \boldsymbol{\theta}_{l-1}\right]$ ■

Given these lemmas, we are now in a position to prove lemma 17.
Proof [Proof of Lemma 17]

$$\begin{aligned}
&\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_\mathcal{L})^{-1}\mathcal{A}_\mathcal{L} + (\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-1}\mathcal{A}_\mathcal{R}^\top\right)\cdot\mathbb{E}\left[\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right]\right\rangle\\
&= \left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \left(\mathcal{I} + (\mathcal{I}-\mathcal{A}_\mathcal{L})^{-1}\mathcal{A}_\mathcal{L} + (\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-1}\mathcal{A}_\mathcal{R}^\top\right)(\mathcal{I}-\mathcal{A}_\mathcal{L}\mathcal{A}_\mathcal{R}^\top)^{-1}(\mathcal{I}-\mathcal{A}_\mathcal{L}\mathcal{A}_\mathcal{R}^\top)\cdot\mathbb{E}\left[\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right]\right\rangle\\
&= \left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \left((\mathcal{I}-\mathcal{A}_\mathcal{L})^{-1}(\mathcal{I}-\mathcal{A}_\mathcal{R}^\top)^{-1}\right)(\mathcal{I}-\mathcal{A}_\mathcal{L}\mathcal{A}_\mathcal{R}^\top)\cdot\mathbb{E}\left[\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right]\right\rangle \quad \text{(using Lemma 11)}\\
&= \left\langle \left((\mathcal{I}-\mathcal{A}_\mathcal{L}^\top)^{-1}(\mathcal{I}-\mathcal{A}_\mathcal{R})^{-1}\right)\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, (\mathcal{I}-\mathcal{A}_\mathcal{L}\mathcal{A}_\mathcal{R}^\top)\cdot\mathbb{E}\left[\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right]\right\rangle\\
&= \left\langle (\mathbf{I}-\mathbf{A}^\top)^{-1}\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}(\mathbf{I}-\mathbf{A})^{-1}, (\mathcal{I}-\mathcal{A}_\mathcal{L}\mathcal{A}_\mathcal{R}^\top)\cdot\mathbb{E}\left[\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right]\right\rangle\\
&= \left\langle (\mathbf{I}-\mathbf{A}^\top)^{-1}\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}(\mathbf{I}-\mathbf{A})^{-1}, (\mathcal{I}-\mathcal{D})\cdot\mathbb{E}\left[\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right]\right\rangle\\
&= \frac{1}{(q-c\delta)^2}\left\langle \left(\otimes_2\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2}\\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix}\right), (\mathcal{I}-\mathcal{D})\cdot\mathbb{E}\left[\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right]\right\rangle \quad \text{(using lemma 8)}\\
&= \frac{1}{(q-c\delta)^2}\left\langle \left(\otimes_2\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2}\\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix}\right), (\mathcal{I}-\mathcal{D})(\mathcal{I}-\mathcal{B})^{-1}(\mathcal{I}-\mathcal{B}^l)\widehat{\boldsymbol{\Sigma}}\right\rangle\\
&= \frac{1}{(q-c\delta)^2}\left\langle \left(\otimes_2\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2}\\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix}\right), (\mathcal{I}-\mathcal{B}+\mathcal{R})(\mathcal{I}-\mathcal{B})^{-1}(\mathcal{I}-\mathcal{B}^l)\widehat{\boldsymbol{\Sigma}}\right\rangle
\end{aligned}$$

$$= \frac{1}{(q-c\delta)^2}\left\langle\left(\otimes_2\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2}\\(\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix}\right), \widehat{\mathbf{\Sigma}} - \mathcal{B}^l\widehat{\mathbf{\Sigma}} + \mathcal{R}(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}} - \mathcal{R}(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^l\widehat{\mathbf{\Sigma}}\right\rangle$$

$$\leq \frac{1}{(q-c\delta)^2}\left\langle\left(\otimes_2\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2}\\(\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix}\right), \widehat{\mathbf{\Sigma}} + \sigma^2\mathcal{R}\cdot(5\mathbf{U})\right\rangle \tag{69}$$

So, we need to understand $\mathcal{R}\mathbf{U}$:

$$\begin{aligned}
\mathcal{R}\mathbf{U} &= \mathbb{E}\left(\begin{bmatrix}0 & \delta\cdot(\mathbf{H}-\mathbf{a}\mathbf{a}^\top)\\0 & q\cdot(\mathbf{H}-\mathbf{a}\mathbf{a}^\top)\end{bmatrix}\mathbf{U}\begin{bmatrix}0 & 0\\ \delta\cdot(\mathbf{H}-\mathbf{a}\mathbf{a}^\top) & q\cdot(\mathbf{H}-\mathbf{a}\mathbf{a}^\top)\end{bmatrix}\right)\\
&= \begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix}\otimes\mathbb{E}\left[(\mathbf{H}-\mathbf{a}\mathbf{a}^\top)\mathbf{U}_{22}(\mathbf{H}-\mathbf{a}\mathbf{a}^\top)\right]\\
&= \begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix}\otimes\left(\mathcal{M}-\mathcal{H}_{\mathcal{L}}\mathcal{H}_{\mathcal{R}}\right)\mathbf{U}_{22}\\
&\preceq \begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix}\otimes\mathcal{M}\mathbf{U}_{22}\\
&\preceq \frac{4}{5}\cdot\begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix}\otimes\mathbf{H} \quad \text{(from equation 66)}.
\end{aligned}$$

Then,

$$\begin{aligned}
&\left\langle\begin{bmatrix}\mathbf{H} & 0\\0 & 0\end{bmatrix}, \left(\mathcal{I}+(\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^\top)^{-1}\mathcal{A}_{\mathcal{R}}^\top\right)\cdot\boldsymbol{\theta}_l\otimes\boldsymbol{\theta}_l\right\rangle\\
&\leq \frac{1}{(q-c\delta)^2}\left\langle\left(\otimes_2\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2}\\(\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix}\right), \widehat{\mathbf{\Sigma}}+\sigma^2\mathcal{R}\cdot(5\mathbf{U})\right\rangle \qquad \text{(from equation 69)}\\
&\leq \frac{5\sigma^2}{(q-c\delta)^2}\cdot\left\langle\left(\otimes_2\begin{bmatrix}-(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2}\\(\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix}\right), \begin{bmatrix}\delta^2 & \delta\cdot q\\ \delta\cdot q & q^2\end{bmatrix}\otimes\mathbf{H}\right\rangle\\
&= \frac{5}{(q-c\delta)^2}\cdot d\,\sigma^2\cdot(q-c\delta)^2\\
&= 5\sigma^2 d.
\end{aligned}\tag{70}$$

■

Lemma 19

$$\left|\left\langle\begin{bmatrix}\mathbf{H} & 0\\0 & 0\end{bmatrix}, \left((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-2}\mathcal{A}_{\mathcal{L}}+(\mathcal{I}-\mathcal{A}_{\mathcal{R}}^\top)^{-2}\mathcal{A}_{\mathcal{R}}^\top\right)\mathbf{\Phi}_\infty\right\rangle\right| \leq C\cdot\sigma^2 d\sqrt{\kappa\widetilde{\kappa}}$$

Where, C is a universal constant.

Proof We begin by noting the following while considering the left side of the above expression:

$$\begin{aligned}
&\left\langle\begin{bmatrix}\mathbf{H} & 0\\0 & 0\end{bmatrix}, \left((\mathcal{I}-\mathcal{A}_{\mathcal{R}}^\top)^{-2}\mathcal{A}_{\mathcal{R}}^\top+(\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-2}\mathcal{A}_{\mathcal{L}}\right)\mathbf{\Phi}_\infty\right\rangle\\
&= \left\langle\begin{bmatrix}\mathbf{H} & 0\\0 & 0\end{bmatrix}\mathbf{A}(\mathbf{I}-\mathbf{A})^{-2}+(\mathbf{I}-\mathbf{A}^\top)^{-2}\mathbf{A}^\top\begin{bmatrix}\mathbf{H} & 0\\0 & 0\end{bmatrix}, \mathbf{\Phi}_\infty\right\rangle
\end{aligned}$$

The inner product above is a sum of two terms, so let us consider the first of the terms:

$$\begin{aligned}
&\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix} \mathbf{A}(\mathbf{I}-\mathbf{A})^{-2}, \mathbf{\Phi}_\infty \right\rangle \\
&= \mathrm{Tr}\bigg((\mathbf{I}-\mathbf{A}^\top)^{-2}\mathbf{A}^\top \begin{bmatrix}\mathbf{H}^{1/2}\\ 0\end{bmatrix} \begin{bmatrix}\mathbf{H}^{1/2} & 0\end{bmatrix} \mathbf{\Phi}_\infty \bigg) \\
&= \mathrm{Tr}\bigg(\bigg(\begin{bmatrix}\mathbf{H}^{1/2}\\ 0\end{bmatrix}^\top \mathbf{\Phi}_\infty (\mathbf{I}-\mathbf{A}^\top)^{-2}\mathbf{A}^\top \begin{bmatrix}\mathbf{H}^{1/2}\\ 0\end{bmatrix} \bigg)\bigg) \\
&= \sum_{j=1}^d \mathrm{Tr}\bigg(\bigg(\begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix}^\top (\mathbf{\Phi}_\infty)_j (\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top \begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix} \bigg)\bigg) \\
&= \sum_{j=1}^d \mathrm{Tr}\bigg(\bigg(\begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix}^\top (\mathbf{\Phi}_\infty^{1/2})_j \bigg) \cdot \bigg((\mathbf{\Phi}_\infty^{1/2})_j^\top (\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top \begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix} \bigg)\bigg),
\end{aligned}$$

where $(\mathbf{\Phi}_\infty)_j$ is the 2×2 block of $\mathbf{\Phi}_\infty$ corresponding to the j^{th} eigensubspace of $\mathbf{H}$, $(\mathbf{\Phi}_\infty^{1/2})_j$ denotes the $2\times 2d$ submatrix (i.e., 2 rows) of $\mathbf{\Phi}_\infty^{1/2}$ corresponding to the j^{th} eigensubspace and $\mathbf{A}_j$ denotes the j^{th} diagonal block of $\mathbf{A}$. Note that $(\mathbf{\Phi}_\infty^{1/2})_j(\mathbf{\Phi}_\infty^{1/2})_j^\top = (\mathbf{\Phi}_\infty)_j$. It is very easy to observe that the second term in the dot product can be written in a similar manner, i.e.:

$$\begin{aligned}
&\left\langle (\mathbf{I}-\mathbf{A}^\top)^{-2}\mathbf{A}^\top \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \mathbf{\Phi}_\infty \right\rangle \\
&= \sum_{j=1}^d \mathrm{Tr}\bigg(\bigg((\mathbf{\Phi}_\infty^{1/2})_j^\top \begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix} \bigg) \cdot \bigg(\begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix}^\top \mathbf{A}_j(\mathbf{I}-\mathbf{A}_j)^{-2}(\mathbf{\Phi}_\infty^{1/2})_j \bigg)\bigg)
\end{aligned}$$

So, essentially, the expression in the left side of the lemma can be upper bounded by using cauchy-shwartz inequality:

$$\begin{aligned}
&\mathrm{Tr}\bigg(\bigg(\begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix}^\top (\mathbf{\Phi}_\infty^{1/2})_j \bigg) \cdot \bigg((\mathbf{\Phi}_\infty^{1/2})_j^\top (\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top \begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix} \bigg)\bigg) \\
&+ \mathrm{Tr}\bigg(\bigg((\mathbf{\Phi}_\infty^{1/2})_j^\top \begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix} \bigg) \cdot \bigg(\begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix}^\top \mathbf{A}_j(\mathbf{I}-\mathbf{A}_j)^{-2}(\mathbf{\Phi}_\infty^{1/2})_j \bigg)\bigg) \\
&\qquad \le 2 \left\| \begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix} \right\|_{(\mathbf{\Phi}_\infty)_j} \cdot \left\| (\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top \begin{bmatrix}\lambda_j^{1/2}\\ 0\end{bmatrix} \right\|_{(\mathbf{\Phi}_\infty)_j} \qquad (71)
\end{aligned}$$

The advantage with the above expression is that we can now begin to employ psd upper bounds on the covariance of the steady state distribution $\mathbf{\Phi}_\infty$ and provide upper bounds on the expression on the right hand side. In particular, we employ the following bound provided by the taylor expansion that gives us an upper bound on $\mathbf{\Phi}_\infty$:

$$\mathbf{\Phi}_\infty \stackrel{\text{def}}{=} \begin{bmatrix}\hat{\mathbf{U}}_{11} & \hat{\mathbf{U}}_{12}\\ \hat{\mathbf{U}}_{12}^\top & \hat{\mathbf{U}}_{22}\end{bmatrix} \preceq 5\sigma^2\mathbf{U} = 5\sigma^2 \begin{bmatrix}\mathbf{U}_{11} & \mathbf{U}_{12}\\ \mathbf{U}_{12}^\top & \mathbf{U}_{22}\end{bmatrix} \quad \text{(using equation 68)}$$

This implies in particular that $(\mathbf{\Phi}_\infty)_j \preceq 5\sigma^2\mathbf{U}_j$ for every $j \in [d]$ and hence, for any vector $\|\mathbf{a}\|_{(\mathbf{\Phi}_\infty)_j} \leq \sqrt{5\sigma^2}\,\|\mathbf{a}\|_{\mathbf{U}_j}$. The important property of the matrix $\mathbf{U}$ that serves as a PSD upper bound is that it is diagonalizable using the basis of $\mathbf{H}$, thus allowing us to bound the computations in each of the eigen directions of $\mathbf{H}$.

$$\begin{aligned}
&\left\|(\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top\begin{bmatrix}\lambda_j^{1/2}\\0\end{bmatrix}\right\|_{(\mathbf{\Phi}_\infty)_j}\\
=&\sqrt{\begin{bmatrix}\lambda_j^{1/2} & 0\end{bmatrix}\mathbf{A}_j(\mathbf{I}-\mathbf{A}_j)^{-2}(\mathbf{\Phi}_\infty)_j(\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top\begin{bmatrix}\lambda_j^{1/2}\\0\end{bmatrix}}\\
\leq&\sqrt{5\sigma^2\begin{bmatrix}\lambda_j^{1/2} & 0\end{bmatrix}\mathbf{A}_j(\mathbf{I}-\mathbf{A}_j)^{-2}\mathbf{U}_j(\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top\begin{bmatrix}\lambda_j^{1/2}\\0\end{bmatrix}}\\
=&\sqrt{5\sigma^2}\left\|(\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top\begin{bmatrix}\lambda_j^{1/2}\\0\end{bmatrix}\right\|_{\mathbf{U}_j}
\end{aligned}\tag{72}$$

So, let us consider $\begin{bmatrix}\lambda_j^{1/2} & 0\end{bmatrix}\mathbf{A}_j(\mathbf{I}-\mathbf{A}_j)^{-2}$ and write out the following series of equations:

$$\begin{aligned}
\begin{bmatrix}\lambda_j^{1/2} & 0\end{bmatrix}\mathbf{A}_j &= \begin{bmatrix}0 & \sqrt{\lambda_j}(1-\delta\lambda_j)\end{bmatrix}\\
\mathbf{I}-\mathbf{A}_j &= \begin{bmatrix}1 & -(1-\delta\lambda_j)\\ c & -(c-q\lambda_j)\end{bmatrix}\\
\det(\mathbf{I}-\mathbf{A}_j) &= (q-c\delta)\lambda_j\\
(\mathbf{I}-\mathbf{A}_j)^{-1} &= \frac{1}{(q-c\delta)\lambda_j}\begin{bmatrix}-(c-q\lambda_j) & 1-\delta\lambda_j\\ -c & 1\end{bmatrix}\\
\implies \begin{bmatrix}\lambda_j^{1/2} & 0\end{bmatrix}\mathbf{A}_j(\mathbf{I}-\mathbf{A}_j)^{-1} &= \frac{\sqrt{\lambda_j}(1-\delta\lambda_j)}{(q-c\delta)\lambda_j}\begin{bmatrix}-c & 1\end{bmatrix}\\
\implies \begin{bmatrix}\lambda_j^{1/2} & 0\end{bmatrix}\mathbf{A}_j(\mathbf{I}-\mathbf{A}_j)^{-2} &= \frac{\sqrt{\lambda_j}(1-\delta\lambda_j)}{((q-c\delta)\lambda_j)^2}\begin{bmatrix}-c(1-c+q\lambda_j) & 1-c+c\delta\lambda_j\end{bmatrix}\\
&= \frac{\sqrt{\lambda_j}(1-\delta\lambda_j)}{((q-c\delta)\lambda_j)^2}\cdot\Big((1-c+c\delta\lambda_j)\begin{bmatrix}-c & 1\end{bmatrix} - c\lambda_j(q-c\delta)\begin{bmatrix}1 & 0\end{bmatrix}\Big)
\end{aligned}$$

This implies,

$$\left\|(\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top\begin{bmatrix}\lambda_j^{1/2}\\0\end{bmatrix}\right\|_{\mathbf{U}_j} \leq \frac{\sqrt{\lambda_j}(1-\delta\lambda_j)}{((q-c\delta)\lambda_j)^2}\cdot(1-c+c\delta\lambda_j)\left\|\begin{bmatrix}-c\\1\end{bmatrix}\right\|_{\mathbf{U}_j} + \frac{c\sqrt{\lambda_j}(1-\delta\lambda_j)}{((q-c\delta)\lambda_j)}\left\|\begin{bmatrix}1\\0\end{bmatrix}\right\|_{\mathbf{U}_j}\tag{73}$$

Next, let us consider $\left\|\begin{bmatrix}-c\\1\end{bmatrix}\right\|_{\mathbf{U}_j}^2$:

$$\left\|\begin{bmatrix}-c\\1\end{bmatrix}\right\|_{\mathbf{U}_j}^2 = c^2u_{11}+u_{22}-2c\cdot u_{12}$$

Note that u_{11}, u_{12}, u_{22} share the same denominator, so let us evaluate the numerator $\text{nr}(c^2 u_{11} - 2cu_{12} + u_{22})$. For this, we have, from equations 63, 61, 56 respectively: Furthermore,

$$\begin{aligned}
\text{nr}(u_{11}) &= (1 + c - c\delta\lambda_j)(q - c\delta) - 2\delta\lambda_j(q - c\delta) + 2\delta^2\lambda_j \\
\text{nr}(u_{12}) &= (1 + c - \lambda_j(q + c\delta))(q - c\delta) + \delta\lambda_j(q + c\delta) \\
\text{nr}(u_{22}) &= (1 + c - c\delta\lambda_j)(q - c\delta) + 2cq\delta\lambda_j
\end{aligned}$$

Combining these, we have:

$$\begin{aligned}
&c^2\text{nr}(u_{11}) - 2c \cdot \text{nr}(u_{12}) + \text{nr}(u_{22}) \\
=&\big((1 + c - c\delta\lambda_j)(1 - c)^2 + 2cq\lambda_j\big)(q - c\delta) - 2c^2\delta\lambda_j(q - c\delta) \\
=&\big((1 + c - c\delta\lambda_j)(1 - c)^2(q - c\delta)\big) + 2c\lambda_j(q - c\delta)^2
\end{aligned}$$

Implying,

$$\left\| \begin{bmatrix} -c \\ 1 \end{bmatrix} \right\|_{\mathbf{U}_j}^2 = \frac{(1 + c - c\delta\lambda_j)(1 - c)^2(q - c\delta) + 2c\lambda_j(q - c\delta)^2}{1 - c^2 + c\lambda_j(q + c\delta)}$$

In a very similar manner,

$$\begin{aligned}
\left\| \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j}^2 &= u_{11} \\
&= \frac{(1 + c - c\delta\lambda_j)(q - c\delta) - 2\delta\lambda_j(q - c\delta) + 2\delta^2\lambda_j}{1 - c^2 + c\lambda_j(q + c\delta)}
\end{aligned}$$

This implies, plugging into equation 73

$$\begin{aligned}
&\left\| (\mathbf{I} - \mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j} \\
&\leq \frac{\sqrt{\lambda_j}(1 - \delta\lambda_j)}{((q - c\delta)\lambda_j)^2} \cdot (1 - c + c\delta\lambda_j)\sqrt{\frac{(1 + c - c\delta\lambda_j)(1 - c)^2(q - c\delta) + 2c\lambda_j(q - c\delta)^2}{1 - c^2 + c\lambda_j(q + c\delta)}} \\
&+ \frac{c\sqrt{\lambda_j}(1 - \delta\lambda_j)}{((q - c\delta)\lambda_j)}\sqrt{\frac{(1 + c - c\delta\lambda_j)(q - c\delta) - 2\delta\lambda_j(q - c\delta) + 2\delta^2\lambda_j}{1 - c^2 + c\lambda_j(q + c\delta)}} \qquad (74)
\end{aligned}$$

Finally, we need,

$$\left\| \begin{bmatrix} \mathbf{H}^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{\Phi}_\infty} \leq \sqrt{5\sigma^2} \left\| \begin{bmatrix} \mathbf{H}^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}}$$

Again, this can be analyzed in each of the eigen directions $(\lambda_j, \mathbf{u}_j)$ of $\mathbf{H}$ to yield:

$$\left\| \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j} = \sqrt{\lambda_j u_{11}}$$

$$= \sqrt{\lambda_j \cdot \frac{(1+c-c\delta\lambda_j)(q-c\delta) - 2\delta\lambda_j(q-c\delta) + 2\delta^2\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}} \quad (75)$$

Now, we require to bound the product of equation 74 and 75:

$$\left\| (\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j} \cdot \left\| \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j} = T_1 + T_2 \quad (76)$$

Where,

$$T_1 = \frac{\lambda_j(1-\delta\lambda_j)}{((q-c\delta)\lambda_j)^2} \cdot (1-c+c\delta\lambda_j)\left(\sqrt{\frac{(1+c-c\delta\lambda_j)(1-c)^2(q-c\delta)+2c\lambda_j(q-c\delta)^2}{1-c^2+c\lambda_j(q+c\delta)}}\right) \cdot \left(\sqrt{\frac{(1+c-c\delta\lambda_j)(q-c\delta)-2\delta\lambda_j(q-c\delta)+2\delta^2\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}}\right)$$

And,

$$T_2 = \frac{c(1-\delta\lambda_j)}{q-c\delta} \cdot \left(\frac{(1+c-c\delta\lambda_j)(q-c\delta)-2\delta\lambda_j(q-c\delta)+2\delta^2\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}\right)$$

We begin by considering T_1:

$$\begin{aligned}
T_1 &= \frac{\lambda_j(1-\delta\lambda_j)}{((q-c\delta)\lambda_j)^2} \cdot (1-c+c\delta\lambda_j)\left(\sqrt{\frac{(1+c-c\delta\lambda_j)(1-c)^2(q-c\delta)+2c\lambda_j(q-c\delta)^2}{1-c^2+c\lambda_j(q+c\delta)}}\right) \\
&\cdot \left(\sqrt{\frac{(1+c-c\delta\lambda_j)(q-c\delta)-2\delta\lambda_j(q-c\delta)+2\delta^2\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}}\right) \\
&= \left(\frac{\lambda_j(1-\delta\lambda_j)}{((q-c\delta)\lambda_j)^2}\right) \cdot \left(\frac{1-c+c\delta\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}\right) \cdot \\
&\left(\sqrt{(1+c-c\delta\lambda_j)(q-c\delta)-2\delta\lambda_j(q-c\delta)+2\delta^2\lambda_j} \cdot \sqrt{(1+c-c\delta\lambda_j)(1-c)^2(q-c\delta)+2c\lambda_j(q-c\delta)^2}\right) \\
&\le \left(\frac{\lambda_j}{((q-c\delta)\lambda_j)^2}\right) \cdot \left(\frac{1-c+c\delta\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}\right) \cdot \\
&\left(\sqrt{(1+c-c\delta\lambda_j)(q-c\delta)+2\delta^2\lambda_j} \cdot \sqrt{(1+c-c\delta\lambda_j)(1-c)^2(q-c\delta)+2c\lambda_j(q-c\delta)^2}\right)
\end{aligned} \quad (77)$$

We will consider the four terms within the square root and bound them separately:

$$\begin{aligned}
T_1^{11} &= \frac{(1+c-c\delta\lambda_j)(1-c)}{(q-c\delta)\lambda_j} \\
&\le \frac{2(1-c)}{\lambda_j \cdot (q-c\delta)} \le \frac{2(1+c_3)}{\lambda_j\gamma} \\
&\le \frac{2(1+c_3)}{c_2\sqrt{2c_1-c_1^2}}\sqrt{\kappa\widetilde{\kappa}}
\end{aligned}$$

Next,

$$\begin{aligned}
T_1^{21} &= \frac{\sqrt{2\delta^2\lambda_j}\sqrt{(1+c-c\delta\lambda_j)(1-c)^2(q-c\delta)}}{(q-c\delta)^2\lambda_j} \\
&\leq \frac{2\delta(1-c)}{\sqrt{(q-c\delta)^3\lambda_j}} = \frac{2\delta}{\sqrt{(q-c\delta)\lambda_j}}\frac{1-c}{q-c\delta} \\
&= \frac{2(1+c_3)\delta}{\gamma}\cdot\frac{1}{\sqrt{(q-c\delta)\lambda_j}} \\
&\leq \frac{2(1+c_3)\delta}{\gamma}\cdot\frac{1}{\sqrt{\gamma(1-\alpha)\mu}} \\
&\leq \frac{2\sqrt{2}(1+c_3)}{c_2^2(2-c_1)}\cdot\widetilde{\kappa}
\end{aligned}$$

Next,

$$\begin{aligned}
T_1^{12} &= \frac{\sqrt{(1+c-c\delta\lambda_j)(q-c\delta)^3\cdot 2c\lambda_j}}{(q-c\delta)^2\lambda_j} \\
&\leq \frac{2\sqrt{2}}{c_2\sqrt{2c_1-c_1^2}}\cdot\sqrt{\kappa\widetilde{\kappa}}
\end{aligned}$$

Finally,

$$\begin{aligned}
T_1^{22} &= \frac{\sqrt{2\delta^2\lambda_j\cdot 2\,c\lambda_j(q-c\delta)^2}}{(q-c\delta)^2\lambda_j} \\
&\leq \frac{2\delta}{q-c\delta} \leq \frac{4}{c_2^2(2-c_1)}\cdot\widetilde{\kappa}
\end{aligned}$$

Implying,

$$\begin{aligned}
T_1 &\leq \left(\frac{1-c+c\delta\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}\right)\cdot(T_1^{11}+T_1^{12}+T_1^{21}+T_1^{22}) \\
&\leq \left(\frac{1-c+c\delta\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}\right)\cdot 2\cdot(1+\sqrt{2}+c_3)\left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2\sqrt{2c_1-c_1^2}}+\sqrt{2}\frac{\widetilde{\kappa}}{c_2^2(2-c_1)}\right) \\
&\leq \left(\frac{1}{1+c}+\frac{1}{2c}\right)\cdot 2\cdot(1+\sqrt{2}+c_3)\left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2\sqrt{2c_1-c_1^2}}+\sqrt{2}\frac{\widetilde{\kappa}}{c_2^2(2-c_1)}\right) \\
&= \left(\frac{1}{1+c}+\frac{1}{2c}\right)\cdot 2\cdot(1+\sqrt{2}+c_3)\left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{\sqrt{c_1c_4}}+\frac{\sqrt{2}\widetilde{\kappa}}{c_4}\right) \\
&\leq \frac{3}{c}\cdot(1+\sqrt{2}+c_3)\left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{\sqrt{c_1c_4}}+\frac{\sqrt{2}\widetilde{\kappa}}{c_4}\right)
\end{aligned}$$

Recall the bound on 1/c from equation 64:

$$\frac{1}{c} \leq \frac{1+\sqrt{c_1c_4}}{1-c_4}$$

Implying,

$$\begin{aligned}
T_1 &\leq \frac{3}{c} \cdot (1+\sqrt{2}+c_3)\left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{\sqrt{c_1 c_4}} + \frac{\sqrt{2}\widetilde{\kappa}}{c_4}\right) \\
&\leq \frac{3}{c} \cdot (1+\sqrt{2}+c_3)\left(\frac{1}{\sqrt{c_1 c_4}} + \frac{\sqrt{2}}{c_4}\right)\sqrt{\kappa\widetilde{\kappa}} \\
&\leq 3(1+\sqrt{2}+c_3)\left(\frac{1}{\sqrt{c_1 c_4}} + \frac{\sqrt{2}}{c_4}\right) \cdot \frac{1+\sqrt{c_1 c_4}}{1-c_4}\sqrt{\kappa\widetilde{\kappa}} \\
&\leq 3(1+\sqrt{2}+\sqrt{(c_4/c_1)})\left(\frac{1}{\sqrt{c_1 c_4}} + \frac{\sqrt{2}}{c_4}\right) \cdot \frac{1+\sqrt{c_1 c_4}}{1-c_4}\sqrt{\kappa\widetilde{\kappa}} \qquad (78)
\end{aligned}$$

Next, we consider T_2:

$$\begin{aligned}
T_2 &= \frac{c(1-\delta\lambda_j)}{q-c\delta} \cdot \left(\frac{(1+c-c\delta\lambda_j)(q-c\delta) - 2\delta\lambda_j(q-c\delta) + 2\delta^2\lambda_j}{1-c^2+c\lambda_j(q+c\delta)}\right) \\
&\leq \left(\frac{(1+c-c\delta\lambda_j)(q-c\delta) - 2\delta\lambda_j(q-c\delta) + 2\delta^2\lambda_j}{(q-c\delta)\cdot(1-c^2+c\lambda_j(q+c\delta))}\right) \\
&\leq \left(\frac{(1+c-c\delta\lambda_j)(q-c\delta) + 2\delta^2\lambda_j}{(q-c\delta)\cdot(1-c^2+c\lambda_j(q+c\delta))}\right)
\end{aligned}$$

We split T_2 into two parts:

$$\begin{aligned}
T_2^1 &= \frac{(1+c-c\delta\lambda_j)}{(1-c^2+c\lambda_j(q+c\delta))} \\
&\leq \frac{1}{1-c} = \frac{1}{1-\alpha+\alpha\beta} \\
&= \frac{1}{(1+c_3)(1-\alpha)} \\
&\leq \frac{2\sqrt{\kappa\widetilde{\kappa}}}{(1+c_3)c_2\sqrt{2c_1-c_1^2}} \\
&\leq \frac{2\sqrt{\kappa\widetilde{\kappa}}}{(1+\sqrt{c_4/c_1})\sqrt{c_1 c_4}} \\
&= \frac{2\sqrt{\kappa\widetilde{\kappa}}}{\sqrt{c_1 c_4}+c_4}
\end{aligned}$$

Then,

$$\begin{aligned}
T_2^2 &= \frac{2\delta^2\lambda_j}{(q-c\delta)(1-c^2+c\lambda_j(q+c\delta))} \\
&\leq \frac{\delta^2\lambda_j}{\gamma(1-\alpha)c^2\lambda_j\delta} = \frac{\delta}{c^2\gamma(1-\alpha)} \\
&= \frac{2\widetilde{\kappa}}{c_4} \cdot \frac{1}{c^2}
\end{aligned}$$

Implying,

$$
\begin{aligned}
T_2 &\leq 2 \cdot \left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_4 + \sqrt{c_1 c_4}} + \frac{\widetilde{\kappa}}{c^2 c_4} \right) \\
&\leq 2 \cdot \left(\frac{1}{\sqrt{c_1 c_4} + c_4} + (\frac{1 + \sqrt{c_1 c_4}}{1 - c_4})^2 \cdot \frac{1}{c_4} \right) \sqrt{\kappa\widetilde{\kappa}} \\
&\leq \frac{2}{c_4} \cdot \left(1 + (\frac{1 + \sqrt{c_1 c_4}}{1 - c_4})^2 \right) \sqrt{\kappa\widetilde{\kappa}} \qquad (79)
\end{aligned}
$$

We add T_1 and T_2 and revisit equation 76:

$$
\begin{aligned}
&\left\| (\mathbf{I} - \mathbf{A}_j^\top)^{-2} \mathbf{A}_j^\top \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j} \cdot \left\| \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j} \\
&= T_1 + T_2 \\
&\leq \left(\frac{2}{c_4} \cdot \left(1 + (\frac{1 + \sqrt{c_1 c_4}}{1 - c_4})^2 \right) + 3 \cdot \frac{1 + \sqrt{c_1 c_4}}{1 - c_4} \cdot \frac{1 + \sqrt{2} + \sqrt{c_4/c_1}}{c_4} \cdot (\sqrt{2} + \sqrt{c_4/c_1}) \right) \sqrt{\kappa\widetilde{\kappa}} \qquad (80)
\end{aligned}
$$

Then, we revisit equation 71:

$$
\begin{aligned}
&\left(\begin{bmatrix} \mathbf{H}^{1/2} \\ 0 \end{bmatrix}^\top \mathbf{\Phi}_\infty^{1/2} \right) \cdot \left(\mathbf{\Phi}_\infty^{1/2} (\mathbf{I} - \mathbf{A}^\top)^{-2} \mathbf{A}^\top \begin{bmatrix} \mathbf{H}^{1/2} \\ 0 \end{bmatrix} \right) + \left(\mathbf{\Phi}_\infty^{1/2} \begin{bmatrix} \mathbf{H}^{1/2} \\ 0 \end{bmatrix} \right) \cdot \left(\begin{bmatrix} \mathbf{H}^{1/2} \\ 0 \end{bmatrix}^\top \mathbf{A} (\mathbf{I} - \mathbf{A})^{-2} \mathbf{\Phi}_\infty^{1/2} \right) \\
&\leq 2 \sum_{j=1}^d \left\| \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{(\mathbf{\Phi}_\infty)_j} \cdot \left\| (\mathbf{I} - \mathbf{A}_j^\top)^{-2} \mathbf{A}_j^\top \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{(\mathbf{\Phi}_\infty)_j} \\
&\leq 10\sigma^2 \sum_{j=1}^d \left\| \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j} \cdot \left\| (\mathbf{I} - \mathbf{A}_j^\top)^{-2} \mathbf{A}_j^\top \begin{bmatrix} \lambda_j^{1/2} \\ 0 \end{bmatrix} \right\|_{\mathbf{U}_j} \qquad \text{(using equation 68)} \\
&\leq 10\sigma^2 \cdot d \cdot \left(\frac{2}{c_4} \cdot \left(1 + (\frac{1 + \sqrt{c_1 c_4}}{1 - c_4})^2 \right) + 3 \cdot \frac{1 + \sqrt{c_1 c_4}}{1 - c_4} \cdot \frac{1 + \sqrt{2} + \sqrt{c_4/c_1}}{c_4} \cdot (\sqrt{2} + \sqrt{c_4/c_1}) \right) \sqrt{\kappa\widetilde{\kappa}} \\
&\leq C\sigma^2 d \sqrt{\kappa\widetilde{\kappa}} \qquad (81)
\end{aligned}
$$

Where the equation in the penultimate line is obtained by summing over all eigen directions the bound implied by equation 80, and C is a universal constant. ■

Lemma 20

$$
\begin{aligned}
&\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\textit{variance}} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{\textit{variance}} \right] \right\rangle \leq 5 \frac{\sigma^2 d}{n - t} + C \cdot \frac{\sigma^2 d}{(n-t)^2} \cdot \sqrt{\kappa\widetilde{\kappa}} \\
&+ C \cdot \frac{\sigma^2 d}{n - t} (\kappa\widetilde{\kappa})^{11/4} \exp\left(- \frac{(n - t - 1) c_2 \sqrt{2c_1 - c_1^2}}{4\sqrt{\kappa\widetilde{\kappa}}} \right) \\
&+ C \cdot \frac{\sigma^2 d}{(n-t)^2} \cdot \exp\left(-(n+1) \frac{c_1 c_3^2}{\sqrt{\kappa\widetilde{\kappa}}} \right) \cdot (\kappa\widetilde{\kappa})^{7/2} \widetilde{\kappa} + C \cdot \sigma^2 d \cdot (\kappa\widetilde{\kappa})^{7/4} \exp\left(-(n+1) \cdot \frac{c_2 c_3 \sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}} \right)
\end{aligned}
$$

where, C is a universal constant.

Proof We begin by recounting the expression for the covariance of the variance error of the tail-averaged iterate $\bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}}$ from equation 24:

$$\mathbb{E}\left[\bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}} \otimes \bar{\boldsymbol{\theta}}_{t,n}^{\text{variance}}\right] = \underbrace{\frac{1}{n-t}\big(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}}}_{\mathcal{E}_1 \stackrel{\text{def}}{=}}$$
$$\underbrace{-\frac{1}{(n-t)^2}\big((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-2}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-2}\mathcal{A}_{\mathcal{R}}^{\top}\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}}}_{\mathcal{E}_2 \stackrel{\text{def}}{=}}$$
$$\underbrace{+\frac{1}{(n-t)^2}\big((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-2}\mathcal{A}_{\mathcal{L}}^{n+1-t} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-2}(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-t}\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}}}_{\mathcal{E}_3 \stackrel{\text{def}}{=}}$$
$$\underbrace{-\frac{1}{(n-t)^2}\big(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\big)(\mathcal{I}-\mathcal{B})^{-2}(\mathcal{B}^{t+1}-\mathcal{B}^{n+1})\widehat{\boldsymbol{\Sigma}}}_{\mathcal{E}_4 \stackrel{\text{def}}{=}}$$
$$\underbrace{+\frac{1}{(n-t)^2}\sum_{j=t+1}^{n}\big((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-j}\big)(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{j}\widehat{\boldsymbol{\Sigma}}}_{\mathcal{E}_5 \stackrel{\text{def}}{=}}$$

The goal is to bound $\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathcal{E}_i \right\rangle$, for $i = 1, .., 5$.

For the case of $\mathcal{E}_1$, combining the fact that $\mathbb{E}\left[\boldsymbol{\theta}_\infty \otimes \boldsymbol{\theta}_\infty\right] = (\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}}$ and lemma 17, we get:

$$\begin{aligned}
\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathcal{E}_1 \right\rangle &= \frac{1}{n-t}\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \big(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}} \right\rangle \\
&= \frac{1}{n-t}\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \big(\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top}\big)\mathbb{E}\left[\boldsymbol{\theta}_\infty \otimes \boldsymbol{\theta}_\infty\right] \right\rangle \\
&\leq 5\frac{\sigma^2 d}{n-t} \qquad (82)
\end{aligned}$$

For the case of $\mathcal{E}_2$, we employ the result from lemma 19, and this gives us:

$$\left|\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathcal{E}_2 \right\rangle\right| \leq \frac{C \cdot \sigma^2 d\sqrt{\kappa\widetilde{\kappa}}}{(n-t)^2} \qquad (83)$$

For $i = 3$, we have:

$$\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathcal{E}_3 \right\rangle = \frac{1}{(n-t)^2}\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \big((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-2}\mathcal{A}_{\mathcal{L}}^{n+1-t} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-2}(\mathcal{A}_{\mathcal{R}}^{\top})^{n+1-t}\big)(\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}} \right\rangle$$
$$= \frac{1}{(n-t)^2}\left(\left\langle (\mathbf{I}-\mathbf{A}^{\top})^{-2}\mathbf{A}^{\top}\begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{A}^{n-t}(\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}} \right\rangle + \left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}\mathbf{A}(\mathbf{I}-\mathbf{A})^{-2}, (\mathcal{I}-\mathcal{B})^{-1}\widehat{\boldsymbol{\Sigma}}\,(\mathbf{A}^{\top})^{n-t} \right\rangle \right)$$

$$= \frac{4d}{(n-t)^2} \cdot \|(\mathbf{I}-\mathbf{A}^\top)^{-2}\mathbf{A}^\top \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}\| \cdot \|\mathbf{A}^{n-t}(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}\| \tag{84}$$

We will consider bounding $\|\mathbf{A}^{n-t}(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}\|$:

$$\begin{aligned}
\|\mathbf{A}^{n-t}(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}\| &\leq \sum_{i=0}^{\infty}\|\mathbf{A}^{n-t}\mathcal{B}^i\widehat{\mathbf{\Sigma}}\| \\
&\leq \frac{12\sqrt{2}}{\sqrt{1-\alpha^2}}\kappa(n-t)\alpha^{(n-t-1)/2}\left(\sum_i \big(1-\frac{c_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}\big)^i\right)\|\widehat{\mathbf{\Sigma}}\| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(using corollary 15)} \\
&= \frac{12\sqrt{2}}{\sqrt{1-\alpha^2}}\kappa(n-t)\alpha^{(n-t-1)/2}\cdot\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2c_3\sqrt{2c_1-c_1^2}}\cdot\|\widehat{\mathbf{\Sigma}}\| \\
&= \frac{12\sqrt{2}\sigma^2}{\sqrt{1-\alpha^2}}\kappa(n-t)\alpha^{(n-t-1)/2}\cdot\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2c_3\sqrt{2c_1-c_1^2}}\cdot(q+c\delta)^2\|\mathbf{H}\| \\
&\leq \frac{108\sqrt{2}\sigma^2}{\sqrt{1-\alpha^2}}\kappa(n-t)\alpha^{(n-t-1)/2}\cdot\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2c_3\sqrt{2c_1-c_1^2}}\cdot\delta^2\|\mathbf{H}\| \qquad (85)
\end{aligned}$$

We also upper bound α as:

$$\begin{aligned}
\alpha &= 1-\frac{c_2\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}+c_2\sqrt{2c_1-c_1^2}} \\
&\leq 1-\frac{c_2\sqrt{2c_1-c_1^2}}{2\sqrt{\kappa\widetilde{\kappa}}} \\
&= e^{-\frac{c_2\sqrt{2c_1-c_1^2}}{2\sqrt{\kappa\widetilde{\kappa}}}} \qquad (86)
\end{aligned}$$

Furthermore, for $\|(\mathbf{I}-\mathbf{A}^\top)^{-2}\mathbf{A}^\top\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}\|$, we consider a bound in each eigendirection j and accumulate the results subsequently:

$$\begin{aligned}
&\|(\mathbf{I}-\mathbf{A}_j^\top)^{-2}\mathbf{A}_j^\top\begin{bmatrix}\lambda_j & 0\\ 0 & 0\end{bmatrix}\| \\
&\leq \frac{1}{(q-c\delta)^2}\cdot\frac{1-\delta\lambda_j}{\lambda_j}\cdot\sqrt{(1+c^2)(1-c)^2+c^2\lambda_j^2(q^2+\delta^2)} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(using lemma 9)} \\
&\leq \frac{\sqrt{7}}{(q-c\delta)^2}\cdot\frac{1}{\lambda_j} \\
&\leq \frac{\sqrt{7}}{(\gamma(1-\alpha))^2}\cdot\frac{1}{\lambda_j} \\
&\leq \frac{48(\kappa\widetilde{\kappa})^2}{(c_1c_4)^2}\frac{\mu^2}{\lambda_j} = \frac{48\widetilde{\kappa}^2}{(\delta c_4)^2}\frac{1}{\lambda_j}
\end{aligned}$$

$$\implies \|(\mathbf{I}-\mathbf{A}^\top)^{-2}\mathbf{A}^\top\begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}\| \leq \frac{48\widetilde{\kappa}^2}{(\delta c_4)^2}\cdot\frac{1}{\mu}$$

Plugging this into equation 84, we obtain:

$$\begin{aligned}\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \mathcal{E}_3 \right\rangle &\leq 41472 \frac{\sigma^2 d}{n-t} (\kappa\widetilde{\kappa})^{11/4} \alpha^{(n-t-1)/2} \frac{1}{c_3 c_4^2 (c_1 c_3)^{3/2}} \\ &\leq C \frac{\sigma^2 d}{n-t} (\kappa\widetilde{\kappa})^{11/4} \alpha^{(n-t-1)/2} \\ &\leq C \frac{\sigma^2 d}{n-t} (\kappa\widetilde{\kappa})^{11/4} \exp^{-\frac{(n-t-1)c_2\sqrt{2c_1 - c_1^2}}{4\sqrt{\kappa\widetilde{\kappa}}}} \qquad (87)\end{aligned}$$

Next, let us consider $\mathcal{E}_4$:

$$\begin{aligned}&\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \mathcal{E}_4 \right\rangle \\ &= -\frac{1}{(n-t)^2} \left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, (\mathcal{I} + (\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^{\top})^{-1}\mathcal{A}_{\mathcal{R}}^{\top})(\mathcal{I}-\mathcal{B})^{-2}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\widehat{\mathbf{\Sigma}} \right\rangle \\ &= -\frac{1}{(n-t)^2} \left\langle (\mathbf{I}-\mathbf{A}^{\top})^{-1} \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix} (\mathbf{I}-\mathbf{A})^{-1}, (\mathcal{I}-\mathcal{D})(\mathcal{I}-\mathcal{B})^{-2}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\widehat{\mathbf{\Sigma}} \right\rangle \\ &= -\frac{1}{(q-c\delta)^2(n-t)^2} \left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix} \right), (\mathcal{I}-\mathcal{B}+\mathcal{R})(\mathcal{I}-\mathcal{B})^{-2}(\mathcal{B}^{t+1} - \mathcal{B}^{n+1})\widehat{\mathbf{\Sigma}} \right\rangle \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(using lemma 8)} \\ &\leq \frac{1}{(q-c\delta)^2(n-t)^2} \left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix} \right), (\mathcal{I}-\mathcal{B}+\mathcal{R})(\mathcal{I}-\mathcal{B})^{-2}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \\ &\leq \frac{1}{(q-c\delta)^2(n-t)^2} \cdot \Bigg(\left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix} \right), (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \\ &\qquad\qquad + \left\langle \mathcal{R}^{\top} \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix} \right), (\mathcal{I}-\mathcal{B})^{-2}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \Bigg) \\ &= \frac{1}{(q-c\delta)^2(n-t)^2} \cdot \Bigg(\left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix} \right), (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \\ &\qquad\qquad + \left\langle \otimes_2 \begin{bmatrix}\delta \\ q\end{bmatrix} \otimes \Big(\mathcal{M} - \mathcal{H}_{\mathcal{L}}\mathcal{H}_{\mathcal{R}} \Big) (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1}(\mathbf{I}-\delta\mathbf{H}), (\mathcal{I}-\mathcal{B})^{-2}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \Bigg) \\ &\leq \frac{1}{(q-c\delta)^2(n-t)^2} \cdot \Bigg(\left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix} \right), (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \\ &\qquad\qquad\qquad\qquad + \widetilde{\kappa} \cdot \left\langle (\widehat{\mathbf{\Sigma}}/\sigma^2), (\mathcal{I}-\mathcal{B})^{-2}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \Bigg) \qquad (88)\end{aligned}$$

To bound $\| \otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2}\end{bmatrix} \|$, we will consider a bound along each eigendirection and accumulate the results:

$$\| \otimes_2 \begin{bmatrix} -(c-q\lambda_j)\lambda_j^{-1/2} \\ (1-\delta\lambda_j)\lambda_j^{-1/2}\end{bmatrix} \| \leq \frac{(c-q\lambda_j)^2 + (1-\delta\lambda_j)^2}{\lambda_j}$$

$$\leq 2 \cdot \frac{(1+c^2) + (q^2+\delta^2)\lambda_j^2}{\lambda_j}$$
$$\leq 2 \cdot \frac{2+5\delta^2\lambda_j^2}{\lambda_j} \leq \frac{14}{\lambda_j}$$
$$\implies \| \otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2} \end{bmatrix} \| \leq \frac{14}{\mu}$$

Next, we bound $\|\mathcal{B}^k(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}\|$ (as a consequence of lemma 13 with $\mathbf{Q} = \widehat{\mathbf{\Sigma}}$):

$$\begin{aligned}
\|\mathcal{B}^k(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}\| &\leq \frac{1}{\lambda_{\min}(\mathbf{G})}\|\mathbf{G}^\top\mathcal{B}^k(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}\| \\
&\leq \frac{1}{\lambda_{\min}(\mathbf{G})}\sum_{l=k}^{\infty}\|\mathbf{G}^\top\mathcal{B}^k\widehat{\mathbf{\Sigma}}\| \\
&\leq \frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2c_3\sqrt{2c_1-c_1^2}}\kappa(\mathbf{G})\exp(-k\frac{c_2c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}})\|\widehat{\mathbf{\Sigma}}\| \\
&\leq \frac{4\sigma^2\kappa}{\sqrt{1-\alpha^2}} \cdot \frac{\sqrt{\kappa\widetilde{\kappa}}}{c_1c_3^2}\exp(-k\frac{c_1c_3^2}{\sqrt{\kappa\widetilde{\kappa}}}) \cdot 9\delta^2\|\mathbf{H}\|_2 \\
&\leq \frac{36\sigma^2\kappa}{\sqrt{1-\alpha^2}} \cdot \frac{\sqrt{\kappa\widetilde{\kappa}}}{c_1c_3^2}\exp(-k\frac{c_1c_3^2}{\sqrt{\kappa\widetilde{\kappa}}}) \cdot \delta
\end{aligned}$$

This implies,

$$\left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2} \end{bmatrix} \right), (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \leq$$
$$504 \cdot \frac{\kappa}{\sqrt{1-\alpha^2}} \cdot \frac{\sqrt{\kappa\widetilde{\kappa}}}{c_1c_3^2}\exp\left(-(n+1)\frac{c_1c_3^2}{\sqrt{\kappa\widetilde{\kappa}}}\right) \cdot \frac{\delta}{\mu} \cdot \sigma^2 d$$

Furthermore,

$$\begin{aligned}
\frac{\widetilde{\kappa}}{\sigma^2}\left\langle \widehat{\mathbf{\Sigma}}, (\mathcal{I}-\mathcal{B})^{-2}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle &= \frac{\widetilde{\kappa}}{\sigma^2}\left\langle (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{(n+1)/2}\widehat{\mathbf{\Sigma}}, (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{(n+1)/2}\widehat{\mathbf{\Sigma}} \right\rangle \\
&\leq \frac{\widetilde{\kappa}}{\sigma^2}\|(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{(n+1)/2}\widehat{\mathbf{\Sigma}}\|^2 \cdot d \\
&\leq 1296\frac{\sigma^2 d}{1-\alpha^2}\left(\kappa\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_1c_3^2}\right)^2\delta^2\widetilde{\kappa}\exp(-(n+1)\frac{c_1c_3^2}{\sqrt{\kappa\widetilde{\kappa}}})
\end{aligned}$$

This implies that,

$$\begin{aligned}
&\left\langle \left(\otimes_2 \begin{bmatrix} -(c\mathbf{I}-q\mathbf{H})\mathbf{H}^{-1/2} \\ (\mathbf{I}-\delta\mathbf{H})\mathbf{H}^{-1/2} \end{bmatrix} \right), (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle + \frac{\widetilde{\kappa}}{\sigma^2}\left\langle \widehat{\mathbf{\Sigma}}, (\mathcal{I}-\mathcal{B})^{-2}\mathcal{B}^{n+1}\widehat{\mathbf{\Sigma}} \right\rangle \\
&\leq 2592\frac{\sigma^2 d}{1-\alpha^2}\left(\kappa\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_1c_3^2}\right)^2\delta^2\widetilde{\kappa}\exp(-(n+1)\frac{c_1c_3^2}{\sqrt{\kappa\widetilde{\kappa}}}) \\
&\leq 2592 \cdot \sigma^2 d \cdot \left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_1c_3^2}\right)^3\exp(-(n+1)\frac{c_1c_3^2}{\sqrt{\kappa\widetilde{\kappa}}}) \cdot \delta^2\kappa^2\widetilde{\kappa} \qquad (89)
\end{aligned}$$

Finally, we also note the following:

$$\frac{1}{(q-c\delta)} \leq \frac{1}{(\gamma(1-\alpha))} \leq \frac{\mu}{(1-\alpha)^2} \leq \frac{4\widetilde{\kappa}}{\delta c_4}$$

Plugging equation 89 into equation 88, we get:

$$\begin{aligned}
\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathcal{E}_4 \right\rangle &= 2592 \cdot \frac{\sigma^2 d}{(n-t)^2 (q-c\delta)^2} \cdot \left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_1 c_3^2} \right)^3 \exp(-(n+1)\frac{c_1 c_3^2}{\sqrt{\kappa\widetilde{\kappa}}}) \cdot \delta^2 \kappa^2 \widetilde{\kappa} \\
&\leq 41472 \cdot \frac{\sigma^2 d}{(n-t)^2} \cdot \frac{1}{c_4^2} \cdot \left(\frac{\sqrt{\kappa\widetilde{\kappa}}}{c_1 c_3^2} \right)^3 \exp(-(n+1)\frac{c_1 c_3^2}{\sqrt{\kappa\widetilde{\kappa}}}) \cdot \kappa^2 \widetilde{\kappa}^3 \\
&= 41472 \cdot \frac{\sigma^2 d}{(n-t)^2} \cdot \frac{1}{c_4^2 (c_1 c_3^2)^3} \cdot \exp\left(-(n+1)\frac{c_1 c_3^2}{\sqrt{\kappa\widetilde{\kappa}}} \right) \cdot (\kappa\widetilde{\kappa})^{7/2} \widetilde{\kappa} \\
&\leq C \cdot \frac{\sigma^2 d}{(n-t)^2} \cdot \exp\left(-(n+1)\frac{c_1 c_3^2}{\sqrt{\kappa\widetilde{\kappa}}} \right) \cdot (\kappa\widetilde{\kappa})^{7/2} \widetilde{\kappa} \qquad (90)
\end{aligned}$$

Next, we consider $\mathcal{E}_5$:

$$\begin{aligned}
&\left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathcal{E}_5 \right\rangle \\
&= \frac{1}{(n-t)^2} \sum_{j=t+1}^{n} \left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \left((\mathcal{I}-\mathcal{A}_{\mathcal{L}})^{-1}\mathcal{A}_{\mathcal{L}}^{n+1-j} + (\mathcal{I}-\mathcal{A}_{\mathcal{R}}^\top)^{-1}(\mathcal{A}_{\mathcal{R}}^\top)^{n+1-j} \right)(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^j \widehat{\mathbf{\Sigma}} \right\rangle \\
&= \frac{1}{(n-t)^2} \sum_{j=t+1}^{n} \Bigg(\left\langle (\mathcal{I}-\mathbf{A}^\top)^{-1}\mathbf{A}^\top \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{A}^{n-j}(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^j \widehat{\mathbf{\Sigma}} \right\rangle \\
&\qquad\qquad + \left\langle \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{A}(\mathcal{I}-\mathbf{A})^{-1}, (\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^j \widehat{\mathbf{\Sigma}} (\mathbf{A}^\top)^{n-j} \right\rangle \Bigg) \\
&\leq \frac{4d}{(n-t)^2} \sum_{j=t+1}^{n} \| (\mathbf{I}-\mathbf{A}^\top)^{-1}\mathbf{A}^\top \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix} \| \cdot \| \mathbf{A}^{n-j}(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^j \widehat{\mathbf{\Sigma}} \| \qquad (91)
\end{aligned}$$

In a manner similar to bounding $\|\mathbf{A}^{n-t}(\mathcal{I}-\mathcal{B})^{-1}\widehat{\mathbf{\Sigma}}\|$ as in equation 85, we can bound $\|\mathbf{A}^{n-j}(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^j\widehat{\mathbf{\Sigma}}\|$ as:

$$\|\mathbf{A}^{n-j}(\mathcal{I}-\mathcal{B})^{-1}\mathcal{B}^j\widehat{\mathbf{\Sigma}}\| \leq \frac{108\sqrt{2}\sigma^2}{\sqrt{1-\alpha^2}}\kappa(n-j)\alpha^{(n-j-1)/2} \cdot \frac{\sqrt{\kappa\widetilde{\kappa}}}{c_2 c_3 \sqrt{2c_1 - c_1^2}} \cdot \exp^{-(\frac{j c_2 c_3 \sqrt{2c_1 - c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}})} \cdot \delta^2 \|\mathbf{H}\|$$

Furthermore, we will consider the bound $\|(\mathbf{I}-\mathbf{A}^\top)^{-1}\mathbf{A}^\top \begin{bmatrix} \mathbf{H} & 0 \\ 0 & 0 \end{bmatrix}\|$ along one eigen direction (by employing equation 26) and collect the results:

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{A}_j^\top)^{-1}\mathbf{A}_j^\top \begin{bmatrix} \lambda_j & 0 \\ 0 & 0 \end{bmatrix}\| &\leq \frac{1+c^2}{q-c\delta} \leq \frac{2}{q-c\delta} \\
&\leq \frac{2}{\gamma(1-\alpha)} \leq \frac{4\widetilde{\kappa}}{\delta c_4}
\end{aligned}$$

$$\implies \|(\mathbf{I}-\mathbf{A}^\top)^{-1}\mathbf{A}^\top \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}\| \leq \frac{4\widetilde{\kappa}}{\delta c_4}$$

Plugging this into equation 91, and upper bounding the sum by $(n-t)$ times the largest term of the series:

$$\left\langle \begin{bmatrix}\mathbf{H} & 0\\ 0 & 0\end{bmatrix}, \mathcal{E}_5 \right\rangle \leq 6912\cdot\sigma^2 d\cdot \frac{(\kappa\widetilde{\kappa})^{7/4}}{c_3 c_4 (c_1 c_3)^{3/2}} \exp^{-(n+1)\cdot\frac{c_2 c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}}$$
$$\leq C\cdot\sigma^2 d\cdot(\kappa\widetilde{\kappa})^{7/4}\cdot\exp^{-(n+1)\cdot\frac{c_2 c_3\sqrt{2c_1-c_1^2}}{\sqrt{\kappa\widetilde{\kappa}}}} \quad (92)$$

Summing up equations 82, 83, 87, 90, 92, the statement of the lemma follows. ■

Appendix F. Proof of Theorem 1

Proof [Proof of Theorem 1] The proof of the theorem follows through various lemmas that have been proven in the appendix:

- Section B provides the bias-variance decomposition and provides an exact tensor expression governing the covariance of the bias error (through lemma 3)and the variance error (lemma 5).
- Section D provides a scalar bound of the bias error through lemma 16. The technical contribution of this section (which introduces a new potential function) is in lemma 4.
- Section E provides a scalar bound of the variance error through lemma 20. The key technical contribution of this section is in the introduction of a stochastic process viewpoint of the proposed accelerated stochastic gradient method through lemmas 6, 17. These lemmas provide a tight characterization of the stationary distribution of the covariance of the iterates of the accelerated method. Lemma 19 is necessary to show the sharp burn-in (up to log factors), beyond which the leading order term of the error is up to constants the statistically optimal error rate $\mathcal{O}(\sigma^2 d/n)$.

Combining the results of these lemmas, we obtain the following guarantee of algorithm 1:

$$\mathbb{E}\left[P(\bar{\mathbf{x}}_{t,n})\right] - P(\mathbf{x}^*) \leq C\cdot\frac{(\kappa\widetilde{\kappa})^{9/4}d\kappa}{(n-t)^2}\cdot\exp\left(-\frac{t+1}{9\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\left(P(\mathbf{x}_0)-P(\mathbf{x}^*)\right)$$
$$+ C\cdot(\kappa\widetilde{\kappa})^{5/4}d\kappa\cdot\exp\left(\frac{-n}{9\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot\left(P(\mathbf{x}_0)-P(\mathbf{x}^*)\right) + 5\frac{\sigma^2 d}{n-t}$$
$$+ C\cdot\frac{\sigma^2 d}{(n-t)^2}\sqrt{\kappa\widetilde{\kappa}} + C\cdot\sigma^2 d\cdot(\kappa\widetilde{\kappa})^{7/4}\cdot\exp\left(\frac{-(n+1)}{9\sqrt{\kappa\widetilde{\kappa}}}\right)$$
$$+ C\cdot\frac{\sigma^2 d}{n-t}(\kappa\widetilde{\kappa})^{11/4}\exp\left(-\frac{(n-t-1)}{30\sqrt{\kappa\widetilde{\kappa}}}\right)$$
$$+ C\cdot\frac{\sigma^2 d}{(n-t)^2}\cdot\exp\left(-\frac{(n+1)}{9\sqrt{\kappa\widetilde{\kappa}}}\right)\cdot(\kappa\widetilde{\kappa})^{7/2}\widetilde{\kappa}$$

Where, C is a universal constant.

■