Proceedings of Machine Learning Research vol 178:1–65, 2022 35th Annual Conference on Learning Theory

Accelerated SGD for Non-Strongly-Convex Least Squares

Aditya Varre ADITYA.VARRE@EPFL.CH
Nicolas Flammarion NICOLAS.FLAMMARION@EPFL.CH
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider stochastic approximation for the least squares regression problem in the non-strongly convex setting. We present the first practical algorithm that achieves the optimal prediction error rates in terms of dependence on the noise of the problem, as $O(d/t)$ while accelerating the forgetting of the initial conditions to $O(d/t^2)$. Our new algorithm is based on a simple modification of the accelerated gradient descent. We provide convergence results for both the averaged and the last iterate of the algorithm. In order to describe the tightness of these new bounds, we present a matching lower bound in the noiseless setting and thus show the optimality of our algorithm.

Keywords: momentum, acceleration, least squares, stochastic gradients, non-strongly convex

1. Introduction

When it comes to large scale machine learning, the stochastic gradient descent (SGD) of Robbins and Monro (1951) is the practitioners' algorithm of choice. Both its practical efficiency and its theoretical performance make it the driving force of modern machine learning (Bottou and Bousquet, 2008). On a practical level, its updates are cheap to compute thanks to stochastic gradients. On a theoretical level, it achieves the optimal rate of convergence with statistically-optimal asymptotic variance for convex problems.

However, the recent successes of deep neural networks brought a new paradigm to the classical learning setting (Ma et al., 2018). In many applications, the variance of gradient noise is not the limiting factor in the optimization anymore; rather it is the distance separating the initialization of the algorithm and the problem solution. Unfortunately, the bias of the stochastic gradient descent, which characterizes how fast the initial conditions are "forgotten", is suboptimal. In this respect, fast gradient methods (including momentum (Polyak, 1964) or accelerated methods (Nesterov, 1983)) are optimal, but have the drawback of being sensitive to noise (d'Aspremont, 2008; Devolder et al., 2014).

This naturally raises the question of whether we can accelerate the bias convergence while still relying on computationally cheap gradient estimates. This question has been partially answered for the elementary problem of least squares regression in a seminal line of research (Dieuleveut et al., 2017; Jain et al., 2018b). Theoretically their methods enjoy the best of both worlds—they converge at the fast rate of accelerated methods while being robust to noise in the gradient. However their investigations are still inconclusive. On the one hand, Jain et al. (2018b) assume the least squares problem to be strongly convex, an assumption which is rarely satisfied in practice but which enables to efficiently stabilise the algorithm. On the other hand, Dieuleveut et al. (2017) makes a simplifying assumption on the gradient oracle they consider and their results do not apply to the cheaply-computed stochastic gradient used in practice. Therefore, even for this simple quadratic problem which is one of the main primitive of machine learning, the question is still open.

In this work, we propose a novel algorithm which accelerates the convergence of the bias term while maintaining the optimal variance for non-strongly convex least squares regression. Our algorithm only requires access to the stream of observations and is easily implementable. It rests on a simple modification of the Nesterov accelerated gradient descent. Following the linear coupling view of Allen-Zhu and Orecchia (2017), acceleration can be obtained by coupling gradient descent and another update with aggressive stepsize. Consequently one simply has to scale down the stepsize in the aggressive update to make it robust to the gradient noise. With this modification, the average of the iterates converges at rate $O(\frac{d\|\boldsymbol{x}_0-\boldsymbol{x}_*\|^2}{t^2}+\frac{\sigma^2 d}{t})$ after t iterations, where $\boldsymbol{x}_0, \boldsymbol{x}_* \in \mathbb{R}^d$ are the starting point and the problem solution, and σ^2 is the noise variance of the linear regression model. In practice, the last iterate is often favored. We show for this latter a convergence of $O(\frac{d\|\boldsymbol{x}_0-\boldsymbol{x}_*\|^2}{t^2}+\sigma^2)$ which is relevant in applications where σ is small. We also investigate the extra dimensional factor compared to the truly accelerated rate. This slowdown comes from the step-size reduction and is shown to be inevitable.

Contributions. In this paper, we make the following contributions:

- In Section 2, we propose a novel stochastic accelerated algorithm AcSGD which rests on a simple modification of the Nesterov accelerated algorithm: scaling down one of its step size makes it provably robust to noise in the gradient.
- In Section 3, we show that the weighted average of the iterates of AcSGD converges at rate $O(\frac{d}{t^2}+\frac{\sigma^2 d}{t})$, thus attaining the optimal rate for the variance and accelerating the bias term.
- In Section 4, we show that the final iterate of AcSGD achieves a convergence rate $O(\frac{d}{t^2}+\sigma^2)$. In particular for noiseless problems, the final iterate converges to the solution at the accelerated rate $O(\frac{d}{t^2})$.
- In Section 5, we show that the dimension dependency in the accelerated rate is necessary for certain distributions and therefore the rates we obtain are optimal.
- The algorithm is simple to implement and practically efficient as we illustrate with simulations on synthetic examples in Section 7.

1.1. Related Work

Our work lies at the intersection of two classical themes - noise stability of accelerated gradient methods and stochastic approximation for least squares.

Accelerated methods and their noise stability. Fast gradient methods refer to first order algorithms which converge at a faster rate than the classical gradient descent—the most famous among them being the accelerated gradient descent of Nesterov (1983). First initiated by Nemirovskij and Yudin (1983), these methods are inspired by algorithms dedicated to the optimization of quadratic functions, i.e., the Heavy ball algorithm (Polyak, 1964) and the conjugate gradient (Hestenes and Stiefel, 1952). For smooth convex problems, these algorithms accelerate the convergence rate of gradient descent from $O(1/t)$ to $O(1/t^2)$, a rate which is optimal among first-order techniques.

These algorithms are however sensitive to noise in the gradients as shown for Heavy ball (Polyak, 1987), conjugate gradient (Greenbaum, 1989), accelerated gradient descent (d'Aspremont, 2008; Devolder et al., 2014) and momentum gradient descent (Yuan et al., 2016). Positive results for accelerated gradient descent were nevertheless obt-ained when the gradients are perturbed with zero-mean finite variance random noise (Lan, 2012; Hu et al., 2009; Xiao, 2009). Convergence rates $O(\frac{L\|\boldsymbol{x}_0-\boldsymbol{x}_*\|^2}{t^2}+\frac{\sigma\|\boldsymbol{x}_0-\boldsymbol{x}_*\|}{\sqrt{t}})$ were proved for L-smooth convex functions with minimum $\boldsymbol{x}_*$, starting point $\boldsymbol{x}_0$ and when the variance of the noisy gradient is bounded by σ^2. Accelerated rates for strongly convex problems were also derived (Ghadimi and Lan, 2012, 2013). For the stochastic Heavy ball, almost sure convergence has been proved (Gadat et al., 2018; Sebbouh et al., 2021) but without improvement over gradient descent.

Stochastic Approximation for Least Squares. Stochastic approximation dates back to Robbins and Monro (1951) and their seminal work on SGD which has then spurred a surge of research. In the convex regime, a complete complexity theory has been derived, with matching upper and lower bounds on the convergence rates (Nemirovski et al., 2008; Bach and Moulines, 2011; Nemirovskij and Yudin, 1983; Agarwal et al., 2012). For smooth problems, averaging techniques (Ruppert, 1988; Polyak, 1990) which consist in replacing the iterates by their average, have had an important theoretical impact. Indeed, Polyak and Juditsky (1992) observed that averaging the SGD iterates along the optimization path provably reduces the impact of gradient noise and makes the estimation rates statistically optimal. The least squares regression problem has been given particular attention (Bach and Moulines, 2013; Dieuleveut and Bach, 2015; Jain et al., 2018a; Flammarion and Bach, 2017; Zou et al., 2021). Bach and Moulines (2013) showed that averaged SGD achieves the non-asymptotic rate of $O(1/t)$ even in the non-strongly convex case. For this problem, the performance of the algorithms can be decomposed as the sum of a bias term, characteristic of the initial-condition forgetting, and a variance term, characteristic of the effect of the noise in the linear statistical model. While averaged SGD obtains the statistically optimal variance term $O(\sigma^2 d/t)$ (Tsybakov, 2003), its bias term converges at a suboptimal rate $O(1/t)$.

Accelerated Stochastic Methods for least squares. Acceleration and stochastic approximation have been reconciled in the setting of least-squares regression by Flammarion and Bach (2015); Dieuleveut et al. (2017); Jain et al. (2018b). Assuming an additive bounded-variance noise oracle, Dieuleveut et al. (2017) designed an algorithm simultaneously achieving optimal prediction error rates, both in terms of forgetting of initial conditions and noise dependence. However this oracle requires the knowledge of the covariance of the features and their algorithm is therefore not applicable in practice. Jain et al. (2018b), relaxed this latter condition and proposed an algorithm using the regular SGD oracle which obtains an accelerated linear rate for strongly convex objectives. However the strong convexity assumption is often too restrictive for machine learning problems where the variables are in large dimension and highly correlated. Thus the strong convexity constant is often insignificant and bounds derived using this assumption are vacuous. We finally note that in the offline setting when multiple passes over the data are possible, accelerated version of variance reduced algorithms have been developed (Frostig et al., 2015; Allen-Zhu, 2017). In the same multipass setting, Paquette and Paquette (2021) studied the convergence of stochastic momentum algorithm and derived asymptotic accelerated rates with a dimension dependent scaling of learning rates similar to ours. The focus of the offline setting is however different and no generalization results are given.

2. Setup: Stochastic Nesterov acceleration for Least squares

We consider the classical problem of least squares regression in the finite dimensional Euclidean space $\mathbb{R}^d$. We observe a stream of samples $(a_n, b_n) \in (\mathbb{R}^d, \mathbb{R})$, for $n \geq 1$, independent and identically sampled from an unknown distribution ρ, such that $\mathbb{E}\|a_n\|^2$ and $\mathbb{E}[b_n^2]$ are finite. The objective is to minimize the population risk

$$\mathcal{R}(x) = \frac{1}{2}\mathbb{E}_\rho\left(\langle x, a\rangle - b\right)^2, \quad \text{where } (a, b) \sim \rho.$$

Covariance. We denote by $\mathbf{H} \stackrel{\text{def}}{=} \mathbb{E}\left[a \otimes a\right]$, the covariance matrix which is also the Hessian of the function $\mathcal{R}$. Without loss of generality, we assume that $\mathbf{H}$ is invertible (by reducing $\mathbb{R}^d$ to a minimal subspace where all $(a_n)_{n\geq 1}$ lie almost surely). The function $\mathcal{R}$ admits then a unique global minimum, we denote by $\boldsymbol{x}_*$, i.e. $\boldsymbol{x}_* = \operatorname{argmin}_{x\in\mathbb{R}^d} \mathcal{R}(x)$. Even if this assumption implies that the eigenvalues of $\mathbf{H}$ are strictly positive, they can still be arbitrarily small. In addition, we do not assume any knowledge of lower bound on the smallest eigenvalue. The smoothness constant of risk $\mathcal{R}$, say L, is the largest eigenvalue of $\mathbf{H}$.

We make the following assumptions on the joint distribution of (a_n, b_n) which are standard in the analysis of stochastic algorithms for the least squares problem.

Assumption 1 (Fourth Moment) *There exists a finite constant R such that*

$$\mathbb{E}\left[\|a\|^2 a \otimes a\right] \preccurlyeq R^2\mathbf{H}. \tag{1}$$

Assumption 2 (Noise Level) *There exists a finite constant σ such that*

$$\mathbb{E}\left[(b - \langle \boldsymbol{x}_*, a\rangle)^2 a \otimes a\right] \preccurlyeq \sigma^2\mathbf{H}. \tag{2}$$

Assumption 3 (Statistical Condition Number) *There exists a finite constant $\tilde{\kappa}$ such that*

$$\mathbb{E}\left[\|a\|^2_{\mathbf{H}^{-1}} a \otimes a\right] \preccurlyeq \tilde{\kappa}\mathbf{H}. \tag{3}$$

Discussion of assumptions. Assumptions 1 and 2 on the fourth moment and the noise level are classical to the analysis of stochastic gradient methods in least squares setting (Bach and Moulines, 2013; Jain et al., 2018a). Assumption 1 holds if the features are bounded, i.e., $\|a\|^2 \leq R^2$, ρ_a almost surely. It also holds, more generally, for features with infinite support such as sub-Gaussian features. Assumption 2 states that the covariance of the gradient at optimum $\boldsymbol{x}_*$ is bounded by $\sigma^2\mathbf{H}$. In the case of homoscedastic/well-specified model i.e. $b = \langle \boldsymbol{x}_*, a\rangle + \epsilon$ where ϵ is independent of a, the above assumption holds with $\sigma^2 = \mathbb{E}\left[\epsilon^2\right]$.

The statistical condition number defined in Assumption 3 is specific to acceleration of SGD. It was introduced by Jain et al. (2018b) in the context of acceleration for strongly convex least squares. It was also used by Even et al. (2021) for the analysis of continuized Nesterov acceleration on non-strongly convex least squares. The statistical condition number is always larger than the dimension, i.e., $\tilde{\kappa} \geq d$, see A.2 in appendix for more details. If $\mathbf{H}$ is not strictly positive definite, a general version of assumption 3 can be stated with pseudoinverse of $\mathbf{H}$ in place of $\mathbf{H}^{-1}$ and in this case the lowerbound on $\tilde{\kappa}$ will be $rank(\mathbf{H})$ in place of d. For sub-Gaussian distribution, $\tilde{\kappa}$ is $O(d)$. However, for one-hot basis distribution, i.e., $a = e_i$ with probability p_i, it is equal to $\tilde{\kappa} = p_{\text{min}}^{-1}$ and thus can be arbitrarily large. In the case of uniform distribution over n data points, Assumption 3 holds with $\tilde{\kappa} = n$, check A.2 for proof.

Nesterov Acceleration. We consider the following algorithm (AcSGD) which starts with the initial values $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\boldsymbol{z}_0 = \boldsymbol{x}_0$ and update for $t \geq 0$

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \beta \nabla_t \mathcal{R}\left(\boldsymbol{x}_t\right), \tag{4a}$$

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \alpha(t+1) \nabla_t \mathcal{R}\left(\boldsymbol{x}_t\right), \tag{4b}$$

$$(t+2)\boldsymbol{x}_{t+1} = (t+1)\boldsymbol{y}_{t+1} + \boldsymbol{z}_{t+1}, \tag{4c}$$

with step sizes $\alpha, \beta > 0$ and where $\nabla_t \mathcal{R}\left(\boldsymbol{x}_t\right)$ is an unbiased estimate of the gradient of $\nabla \mathcal{R}(\boldsymbol{x}_t)$.

This algorithm is similar to the standard three-sequences formulation of the Nesterov accelerated gradient descent (Nesterov, 2005) but with two different learning rates α and β in the gradient steps Eq.(4a), and Eq.(4b). As noted by (Flammarion and Bach, 2015), this formulation captures various algorithms. With exact gradients, for different α, β for e.g. when $\alpha = 0$, we recover averaged gradient descent while with $\beta = 0$ we recover a version of the Heavy ball algorithm and with $\alpha = \beta = 1/L$ we recover the Nesterov acceleration algorithm.

We especially consider the weighted averages of the iterates defined after T iterations by

$$\overline{\boldsymbol{x}}_T \stackrel{\text{def}}{=} \frac{\sum_{t=0}^{T} (t+1)\boldsymbol{x}_t}{\sum_{t=0}^{T} (t+1)}. \tag{5}$$

In contrast to the classical average considered by Polyak and Juditsky (1992), Eq.(5) uses weighted average which gives more importance to the last iterates and is therefore related to tail-averaging.

Stochastic Oracles. Let $(a_t, b_t) \in \left(\mathbb{R}^d, \mathbb{R}\right)$ be the sample at iteration t, we consider the stochastic gradient of $\mathcal{R}$ at $\boldsymbol{x}_t$

$$\nabla_t \mathcal{R}\left(\boldsymbol{x}_t\right) = a_t\left(\langle a_t, \boldsymbol{x}_t\rangle - b_t\right). \tag{6}$$

Note that this is a true stochastic gradient oracle unlike Dieuleveut et al. (2017), where a simpler oracle which assumes the knowledge of the covariance $\mathbf{H}$ is considered. As explained in App. A.1, this oracle combines an additive noise independent of the iterate $\boldsymbol{x}_t$ and a multiplicative noise which scales with $\boldsymbol{x}_t$. Dealing with the multiplicative part of the oracle is the main challenge of our analysis.

3. Convergence of the Averaged Iterates

In this section, we present our main result on the decay rate of the excess error of our estimate. We extend the results of Dieuleveut et al. (2017) to the general stochastic gradient oracle in the following theorem.

Theorem 1 *Consider Algorithm 4 under Assumptions 1, 2, 3 and step sizes satisfying* $(\alpha+2\beta)R^2 \leqslant 1$ *and* $\alpha \leqslant \frac{\beta}{2\tilde{\kappa}}$. *In expectation, the excess risk of estimator* $\overline{\boldsymbol{x}}_T$ *after* T *iterations is bounded as*

$$\mathbb{E}\left[\mathcal{R}\left(\overline{\boldsymbol{x}}_T\right)\right] - \mathcal{R}\left(\boldsymbol{x}_*\right) \leqslant \min\left\{\frac{12}{\alpha T^2}, \frac{48}{\beta T}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \frac{72\,\sigma^2 d}{T}.$$

The constants in the bounds are partially artifacts of the proof technique. The proof can be found in App. B.1. In order to give a clear picture of how the rate depends on the constants $R^2, \tilde{\kappa}$, we give a corollary below for a specific choice of step-sizes.

Corollary 2 *Under the same conditions as Theorem 1 and with the step sizes* $\beta = \frac{1}{3R^2}$, *and* $\alpha = \frac{1}{6\tilde{\kappa}R^2}$. *In expectation, the excess error of estimator* $\overline{\boldsymbol{x}}_T$ *after* T *iterations is bounded as*

$$\mathbb{E}\left[\mathcal{R}\left(\overline{\boldsymbol{x}}_T\right)\right] - \mathcal{R}\left(\boldsymbol{x}_*\right) \leqslant \min\left\{\frac{72\tilde{\kappa}R^2}{T^2}, \frac{144R^2}{T}\right\}\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \frac{72\,\sigma^2 d}{T}.$$

We make the following comments on Theorem 1 and Corollary 2

Optimality of the convergence rate. The convergence rate is composed of two terms: (a) a bias term which describes how fast initial conditions are forgotten and corresponds to the noiseless problem ($\sigma = 0$). (b) A variance term which indicates the effect of the noise in the statistical model, independently of the starting point. It corresponds to the problem where the initialization is the solution $\boldsymbol{x}_*$.

The algorithm recovers the fast rate of $O\left({}^{1}/(\alpha T^2)\right)$ of accelerated gradient descent for the bias term. This is the optimal convergence rate for minimizing quadratic functions with a first-order method. However to make the algorithm robust to the stochastic-gradient noise, the learning rate α has to be scaled with regards to the statistical condition number $\tilde{\kappa}$. For $T \leq \tilde{\kappa}$, the bias of the algorithm decays as that of averaged SGD, i.e, the second component of the bias governs the rate. However the acceleration comes in for $T \geq \tilde{\kappa}$ and we observe an accelerated rate of $O\left({}^{\tilde{\kappa}}/{}_{T^2}\right)$ afterward. This $\tilde{\kappa}$-dependence is the consequence of using computationally cheap rank-one stochastic gradients $a_i\left\langle a_i, \boldsymbol{x}_t - \boldsymbol{x}_*\right\rangle$ instead of the full-rank update $\mathbf{H}(\boldsymbol{x}_t - \boldsymbol{x}_*)$ as in gradient descent. The tightness of the rate with respect to $\tilde{\kappa}$ and consequently on the dimension d is of particular importance and is discussed in Section 5.

The algorithm also recovers the optimal rate $O\left({}^{\sigma^2 d}/{}_{T}\right)$ for the variance term (Tsybakov, 2003). Hence it retains the optimal rate of variance error while improving the rate of bias error over standard SGD.

Stochastic Gradients and Error Accumulation. When true gradients are replaced by stochastic gradients, algorithms accumulate the noisy gradient induced errors as they progress. In order to still converge, the algorithms need to be modified to adapt accordingly. In the case of linear regression, when comparing SGD with GD, the error accumulation due to the multiplicative noise is controlled by scaling the step size from $O({}^{1}/{}_{L})$ to $O({}^{1}/{}_{R^2})$. The error due to the additive noise is controlled by averaging the iterates. In the case of accelerated gradient descent, the scaling of the step sizes becomes intuitive if we consider the *linear coupling* interpretation of Allen-Zhu and Orecchia (2017). In this view of Algorithm 4, a gradient step (on $\boldsymbol{y}_t$) and an *aggressive* gradient step (on $\boldsymbol{z}_t$) are elegantly coupled to achieve acceleration. The aggressive step is more sensitive to noise since it is of scale $O(t)$. Therefore, the step-size α needs to be appropriately scaled down to control the error accumulation of the $\boldsymbol{z}_t$-gradient step. Strikingly, this scaling is proportional to the statistical condition number and therefore to the dimension of the features.

Comparison with Jain et al. (2018b). Note that as both algorithms have the optimal rate for the variance, we only compare the rate for the bias error here. The accelerated stochastic algorithm for strongly convex objectives of Jain et al. (2018b) converges with linear rate $O\left(poly(\mu^{-1}) \cdot e^{-\left(t/\sqrt{\mu^{-1}\tilde{\kappa}}\right)}\right)$ where μ is the smallest eigenvalue of $\mathbf{H}$. We note that (a) this rate is vacuous for finite time horizon smaller that $\sqrt{\tilde{\kappa}/\mu}$ and (b) the algorithm requires the knowledge of the constant μ which is unknown in practice. In comparison, our algorithm converges at rate $O(\tilde{\kappa}/t^2)$ for any arbitrarily small μ and therefore is faster for reasonable finite horizon. Hence, assuming $\mathbf{H}$ invertible does not make the problem strongly convex, emphasizing the relevance of the non-strongly convex setting for least squares problems. The algorithm of Jain et al. (2018b) can also be coupled with an appropriate regularization (Allen-Zhu and Hazan, 2016) to be directly used on non-strongly convex functions. The resulting algorithm achieves a target error ϵ in $O(\sqrt{\tilde{\kappa}/\epsilon}\log \epsilon^{-1})$ iterations. In comparison, our algorithm requires $O(\sqrt{\tilde{\kappa}/\epsilon})$ iterations. Besides the additional logarithmic factor, algorithms with the aforementioned regularization are not truly online, since the target accuracy has to be known and the total number of iterations set in advance. In contrast, our algorithm shows that acceleration can be made robust to stochastic gradients without additional regularization or strong-convexity assumption.

Finite sum minimization of regularized ERM. We investigate here the competitiveness of our method when compared to direct minimization of the regularized ERM objective[1]. The ERM problem can be efficiently minimized using variance reduced algorithms (Johnson and Zhang, 2013; Schmidt et al., 2017) and in particular their accelerated variants (Frostig et al., 2015; Allen-Zhu, 2017). To achieve a target error of ϵ, these methods required $O\left({\sigma^2 d}/{\epsilon} + \left\|\boldsymbol{x}_*\right\|\sqrt{L\,\sigma^2 d}/{\epsilon}\right)$ basic vector computations. The number of vector computation of our algorithm is comparatively $O\left({\sigma^2 d}/{\epsilon} + \left\|\boldsymbol{x}_*\right\|\sqrt{R^2 d}/\sqrt{\epsilon}\right)$, taking $\tilde{\kappa} = O(d)$ for simplicity. Therefore, our method needs fewer computations for small target errors $\epsilon \leq \frac{L\sigma^2}{R^2}$. In addition, accelerated SVRG needs a $O(dn)$ memory, where n is number of samples in ERM while our single pass method only uses a $O(d)$ space.

Mini-Batch Accelerated SGD. We also consider the mini-batch stochastic gradient oracle which queries the gradient oracle several times and returns the average of these stochastic gradients given by the observations $(a_{t,i}, b_{t,i})_{i \leq b}$:

$$\nabla_t \mathcal{R}\left(\boldsymbol{x}_t\right) = \frac{1}{b}\sum_{i=1}^{b} a_{t,i}\left(\langle a_{t,i}, \boldsymbol{x}_t\rangle - b_{t,i}\right). \tag{7}$$

Mini-batching enables to reduce the variance of the gradient estimate and to parallelize the computations. When we implement Algorithm 4 with the mini-batch stochastic gradient oracle defined in Eq.(7), Theorem 1 becomes valid for learning rates satisfying $(\alpha + 2\beta)R^2 \leqslant b$, $\alpha \leq \frac{b\beta}{2\tilde{\kappa}}$ and $\alpha, \beta \leq \frac{1}{L}$. For batch size $b \lesssim {R^2}/{L}$, the rate of convergence becomes $O\left(\min\left\{\frac{\tilde{\kappa}}{b^2 t^2}, \frac{1}{bt}\right\} + \frac{\sigma^2 d}{bt}\right)$. Even if it does not improve the overall sample complexity, using mini-batch is interesting from a practical point of view: the algorithm can be used with larger step size (α scales with b^2), which speeds up the accelerated phase. Indeed the algorithm is accelerated only after ${\tilde{\kappa}}/{b}$ iterations. For larger batch sizes $b \geq {R^2}/{L}$, β cannot be scaled with b due to the condition $\beta \leq 1/L$. The learning rate α can nevertheless be scaled linearly with b, if $b \leq \tilde{\kappa}$. Thus, increasing the batch size still leads to fast rate for Algorithm 4, in accordance with the findings of Cotter et al. (2011) for accelerated gradient methods. This behavior is in contrast to SGD—where the linear speedup is lost for batch size larger than a certain threshold (Jain et al., 2018a). Finally, we note that when the batch size is $O(\tilde{\kappa})$, the performance of the algorithm matches the one of Nesterov accelerated gradient descent. This fact is consistent with the observation of Hsu et al. (2012) that the empirical covariance of $\tilde{\kappa}$ samples is spectrally close to $\mathbf{H}$.

1. Generalization is not guaranteed without regularization (Györfi et al., 2006).

Overparameterized linear regression. The result also applies to training overparameterized linear models where the number of samples (n) can be orders of magnitude smaller than the dimension of the data (d). Our model naturally includes this setting by replacing the population risk $\mathcal{R}(\theta)$ with the training loss, whose finite sum structure can be rewritten as $\mathbb{E}_{\hat{\rho}}\left(\langle x, a\rangle - b\right)^2$ where $\hat{\rho}$ is the empirical distribution over training data. Due to overparameterization, there exists a perfect interpolator and the noise at the optimum is zero, i.e., $\sigma = 0$. Note that in this case Assumption 3 holds with $\tilde{\kappa} = n$, hence $\tilde{\kappa} < d$ and high dimensionality is not a hindrance like before (see A.2 for details). Hence using Algorithm 4 with stochastic gradients sampled with replacement and choosing stepsizes as specified by Theorem 1, we get an accelerated rate of n/t^2 after t iterations. In the terms of epochs, after k epochs, the convergence rate of the training loss is $1/nk^2$, obtaining a quadratic improvement over the $1/nk$ rate of SGD. It is interesting to further extend these results to the setting of sampling without replacement.

4. Last Iterate

In this section, we study the dynamics of the last iterate of Algorithm 4. The latter is often preferred to the averaged iterate in practice. In general, the noise in the gradient prevents the last iterate to converge. When used with constant step sizes, only a convergence in a $O(\sigma^2)$-neighborhood of the solution can be obtained. Therefore variance reduction techniques (including averaging and decaying step sizes) are required. However in the case of noiseless model, i.e., $b = \langle a, \boldsymbol{x}_*\rangle$ ρ-almost surely, last iterate convergence is possible. In such cases, the algorithms are inherently robust to the noise in the stochastic gradients. This setting is particularly relevant to the training of over-parameterized models in the interpolation setting (Varre et al., 2021).

When studying the behavior of the last iterate, we need to make an additional 4-th order assumption on the distribution of the features.

Assumption 4 (Uniform Kurtosis) *There exists a finite constant κ such that for any positive semidefinite matrix M*

$$\mathbb{E}\left[\langle a, Ma\rangle \; a \otimes a\right] \preccurlyeq \kappa \mathrm{Tr}(M\mathbf{H})\mathbf{H}. \tag{8}$$

The above assumption holds for the Gaussian distribution with $\kappa = 3$ and is also satisfied when $\mathbf{H}^{-\frac{1}{2}}a$ has sub-Gaussian tails (Zou et al., 2021). Therefore Assumption 4 is not too restrictive and is often made when analysing SGD for least squares (Dieuleveut et al., 2017; Flammarion and Bach, 2017). It is nevertheless stronger than Assumption 1. For the one-hot-basis distribution, it only holds for $\kappa = 1/p_{\min}$ which can be arbitrarily large. It also directly implies Assumption 3 with a statistical condition number satisfying $\tilde{\kappa} \leq \kappa d$. Yet, the previous inequality is not tight as the example of the one-hot-basis distribution shows.

Under this assumption, we extend the previous results of Flammarion and Bach (2015) to the general stochastic gradient oracle.

Theorem 3 *Consider Algorithm 4 under Assumptions 2, 4 and step sizes satisfying $\kappa(\alpha + 2\beta)\mathrm{Tr}\mathbf{H} \leq 1, \alpha \leq \frac{\beta}{2\kappa d}$. In expectation, the excess risk of the last iterate $\boldsymbol{x}_t$ after t iterations is bounded as*

$$\mathbb{E}\left[\mathcal{R}\left(\boldsymbol{x}_t\right)\right] \leqslant \min\left\{\frac{3}{\alpha t^2}, \frac{24}{\beta t}\right\}\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + 2\left(\left(\alpha + 2\beta\right)\mathrm{Tr}H + \frac{2\alpha d}{\beta}\right)\sigma^2$$

Let us make some comments on the convergence of last iterate. The proof can be found in App. B.2.

- When the step-sizes are set to $\beta = 1/(3\kappa \mathrm{Tr}\mathbf{H})$ and $\alpha = 1/(6d\kappa^2 \mathrm{Tr}\mathbf{H})$ the upper bound on the excess risk becomes $\min\left\{\frac{18\kappa^2 d \mathrm{Tr} H}{t^2}, \frac{144\kappa \mathrm{Tr}\mathbf{H}}{t}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \frac{4}{\kappa}\sigma^2$.

- For constant step sizes, the excess error of the last iterate does not go to zero in the presence of noise in the model. At infinity, it converges to a neighbourhood of $O(\sigma^2)$ and the constant scales with the learning rate. This neighbourhood shrinks as the step size decreases, as long as the step size of the aggressive step α should decrease at a faster rate compared to β. In comparison, Nesterov accelerated gradient descent ($\alpha = \beta$) is diverging.

- For noiseless least squares where $\sigma = 0$, we get an accelerated rate $O(\kappa d/t^2)$, which has to be compared to the $O(1/t)$-rate of SGD. Even et al. (2021); Vaswani et al. (2019) also study acceleration in the context of noiseless models. The rates of Even et al. (2021) depend on $\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|_{\mathbf{H}^{-1}}$ which can be arbitrarily large for ill-conditioned problem. In contrast, our rates are independent of the conditioning. The strong growth condition of Vaswani et al. (2019) is too stringent, indeed for linear regression their constant ρ is the condition number which can be arbitrarily large.

- Following Berthier et al. (2020), a similar result can be obtained on the minimum of the excess risk $\min\limits_{0 \le k \le t} \mathbb{E}\left[\mathcal{R}\left(\boldsymbol{x}_k\right)\right]$ by only assuming the less stringent Assumption 1.

5. Lowerbound and open questions

In this section, we address the tightness of our result with respect to the statistical condition number $\tilde{\kappa}$. In particular we study the impact of the distribution generating the stream of inputs. We start by defining the class of stochastic first-order algorithms for least-squares we consider.

Definition 4 (Stochastic First Order Algorithm for Least Squares) *Given an initial point $\boldsymbol{x}_0$, and a distribution ρ, a stochastic first order algorithm generates a sequence of iterates $\boldsymbol{x}_k$ such that*

$$\boldsymbol{x}_k \in \boldsymbol{x}_0 + \mathit{span}\left\{\nabla_0 f(\boldsymbol{x}_0), \nabla_1 f(\boldsymbol{x}_1), \cdots, \nabla_{k-1} f(\boldsymbol{x}_{k-1})\right\} \quad \mathit{for}\ k \ge 1, \tag{9}$$

where $\nabla_i f$ are the stochastic gradients at the iteration i defined in Eq.(6).

This definition extends the definition of first order algorithms considered by Nesterov (2004) to the stochastic setting. This class of algorithm defined is fairly general and includes SGD and Algorithm 4. By definition of the stochastic oracle, the condition 9 is equivalent to $\boldsymbol{x}_k - \boldsymbol{x}_0$ belonging to the linear span of the features $\{a_1, \cdots, a_k\}$. It is therefore not possible to control the excess error for iterations $t = O(d)$ since the optimum is then likely to be in the span of more than d features. However it is still possible to lowerbound the excess error in the initial stage of the process. This is the object of the following lemma which provides a lower bound for noiseless problems.

Lemma 5 *For all starting point $\boldsymbol{x}_0$, there exists a distribution ρ over $\mathbb{R}^d \times \mathbb{R}$ satisfying Assumption 1 with $R^2 = 1$, Assumption 2 with $\sigma = 0$, Assumption 3 with $\tilde{\kappa} = d$ and an optimum $\boldsymbol{x}'_*$ verifying $\left\|\boldsymbol{x}'_* - \boldsymbol{x}_0\right\|^2 = 1$, such that the expected excess risk of any stochastic first order algorithm is lower bounded as*

$$\mathbb{E}\left[\mathcal{R}\left(\boldsymbol{x}_{\lfloor d/2 \rfloor}\right)\right] = \Omega\left(\frac{1}{d}\right).$$

Check App. B.3 for the proof of the lemma. The excess risk cannot be decreased by more than a factor d in less than d iterations. Fully accelerated rates $O(R^2\|\boldsymbol{x}'_* - \boldsymbol{x}_0\|^2/t^2)$ are thus proscribed. Indeed, they correspond to a decrease $O(1/d^2)$ for the above problem, contradicting the lower bound. Hence, accelerated rates should be scaled with a factor of dimension d. The rate $O(dR^2\|\boldsymbol{x}'_* - \boldsymbol{x}_0\|^2/t^2)$ of Theorem 1 is therefore optimal at the beginning of the optimization process. On the other side, the SGD algorithm achieves a rate of $O(R^2\|\boldsymbol{x}'_* - \boldsymbol{x}_0\|^2/t)$ on noiseless linear regression. For the regression problem described in Lemma 5, this rate is $O(1/d)$ and also optimal.

The proof of the lemma follow the lines of Jain et al. (2018b) and considers the one-hot basis distribution. It is worth noting that the covariance matrix of the worst-case distribution can be fixed beforehand, i.e., for any covariance matrix, there exists a matching distribution such that direct acceleration is impossible (see details in Lemma 17). Therefore the lower bound does not rely on the construction of a particular Hessian, in contrast to the proof of Nesterov (2004) for the deterministic setting. However, the proof strongly leverages the orthogonality of the features output by the oracle. It is still an open question to study similar complexity result for more general, e.g., Gaussian, features.

A different approach is to consider constraints on the computational resources used by the algorithm. Dagan et al. (2019); Sharan et al. (2019) investigate this question from the angle of memory constraint and derive memory/samples tradeoffs for the problem of regression with Gaussian design. Although their results do no have direct implications on the convergence rate of gradient based methods, we observe some interesting phenomena when increasing the memory resource of the algorithms. The stochastic gradient oracle $(\langle a_i, \boldsymbol{x}_t\rangle - b_i)\, a_i$ uses a memory $O(d)$. If we increase the available memory to $O(d^2)$ and consider instead the running average $\frac{1}{t+1}\sum_{i=0}^{t}(\langle a_i, \boldsymbol{x}_t\rangle - b_i)\, a_i$ as the gradient estimate, α no longer needs to be scaled with d and we empirically observe $O(1/t^2)$ convergence (see Figure 2 in App. B.3). This empirical finding suggests that algorithms using a subquadratic amount of memory may provably converge slower than algorithms without memory constraints. Investigating such speed/memory tradeoff is outside of the scope of this paper, but is a promising direction for further research.

6. Proof technique

For the least squares problem, the analysis of stochastic gradient methods is well studied and techniques have been thoroughly refined. Our analysis follows the common underlying scheme. First, the iterates are rescaled to obtain a time invariant linear system. Second, the estimation error is decomposed as the sum of a bias and variance error term which are studied separately. Finally, the rate is obtained using the bias-variance decomposition. However there are significant gaps yet to be filled for this particular problem. The first of many is that the *existing Lyapunov techniques* for either strongly convex functions or classical SGD are not applicable (see App. C.1, for more details). The study of the variance error comes with a different set of challenges.

Time Rescaling. Using the approach of Flammarion and Bach (2015), we first reformulate the algorithm using the following scaled iterates

$$\boldsymbol{u}_t := (t+1)(\boldsymbol{x}_t - \boldsymbol{x}_*) \qquad \boldsymbol{v}_t := t(\boldsymbol{y}_t - \boldsymbol{x}_*) \qquad \boldsymbol{w}_t := \boldsymbol{z}_t - \boldsymbol{x}_*. \tag{10}$$

Using such time rescaling, we can write Algorithm 4 with stochastic gradient oracle as a time-independent linear recursion (with random coefficients depending only on the observations)

$$\boldsymbol{\theta}_{t+1} = \mathcal{J}_t\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_{t+1}, \tag{11}$$

where $\boldsymbol{\theta}_t \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_t \end{bmatrix}$, $\mathcal{J}_t \overset{\text{def}}{=} \begin{bmatrix} \mathbf{I} - \beta a_t a_t^\top & \mathbf{I} - \beta a_t a_t^\top \\ -\alpha a_t a_t^\top & \mathbf{I} - \alpha a_t a_t^\top \end{bmatrix}$ and $\boldsymbol{\epsilon_{t+1}} \overset{\text{def}}{=} (t+1)\left(b_t - \langle \boldsymbol{x}_*, a_t \rangle\right) \begin{bmatrix} \beta a_t \\ \alpha a_t \end{bmatrix}$. The expected excess risk of the averaged iterate $\overline{\boldsymbol{x}}_T$ can then be simply written as

$$\mathbb{E}\left[\mathcal{R}\left(\overline{\boldsymbol{x}}_T\right)\right] - \mathbb{E}\left[\mathcal{R}\left(\boldsymbol{x}_*\right)\right] = \frac{1}{2}\left(\sum_{t=1}^{T+1} t\right)^{-2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_T \otimes \overline{\boldsymbol{\theta}}_T\right] \right\rangle,$$

where we define $\overline{\boldsymbol{\theta}}_T = \sum_{t=0}^{T} \boldsymbol{\theta}_t$ the sum of the rescaled iterates. All that remains to do is to upper-bound the covariance $\mathbb{E}\left[\overline{\boldsymbol{\theta}}_T \otimes \overline{\boldsymbol{\theta}}_T\right]$. It now becomes clear why we consider the averaging scheme in Eq.(5) instead of the classical average of Polyak and Juditsky (1992): it integrates well with our time re-scaling and makes the analysis simpler.

Bias-Variance Decomposition. To upper bound the covariance of our estimator $\overline{\boldsymbol{\theta}}_t$ we form two independent sub-problems:

- **Bias recursion:** the least squares problem is assumed to be noiseless, i.e, $\boldsymbol{\epsilon_t} = 0$ for all $t \geq 0$. It amounts to the studying the following recursion

$$\boldsymbol{\theta}_{b,t+1} = \mathcal{J}_t\, \boldsymbol{\theta}_{b,t} \text{ started from } \boldsymbol{\theta}_{b,0} = \boldsymbol{\theta}_0. \tag{12}$$

- **Variance recursion:** the recursion starts at the optimum ($\boldsymbol{x}_*$) and the noise $\boldsymbol{\epsilon_t}$ drive the dynamics. It is equivalent to the following recursion

$$\boldsymbol{\theta}_{v,t+1} = \mathcal{J}_t \boldsymbol{\theta}_{v,t} + \boldsymbol{\epsilon_{t+1}} \text{ started from } \boldsymbol{\theta}_{v,0} = \mathbf{0}. \tag{13}$$

The bias-variance decomposition (see Lemma 13 in App. A.4) consists of upperbounding the covariance of the iterates as

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_T \otimes \overline{\boldsymbol{\theta}}_T\right] \preccurlyeq 2\left(\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{b,T} \otimes \overline{\boldsymbol{\theta}}_{b,T}\right] + \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right]\right), \tag{14}$$

where $\overline{\boldsymbol{\theta}}_{b,T} \overset{\text{def}}{=} \sum_{t=0}^{T} \boldsymbol{\theta}_{b,T}$ and $\overline{\boldsymbol{\theta}}_{v,T} \overset{\text{def}}{=} \sum_{t=0}^{T} \boldsymbol{\theta}_{v,T}$. The bias error and the variance error can then be studied separately.

The bias error is directly given by the following lemma which controls the finite sum of the excess bias risk. In contrast with the strongly convex case, no simple Lyapunov function exists and we overcome this by a sharp characterization for the covariance of bias in the lemma below. To prove the lemma, we relate the sum of the expected covariances of the iterates Algorithm 4 with stochastic gradients to the sum of the covariance of iterates of Algorithm 4 with exact gradients. For detailed proof, see Lemma 20.

Lemma 6 (Potential for Bias) *Under Assumptions 1,3 and the step-sizes satisfying the conditions of Theorem 1. For $T \geq 0$,*

$$\sum_{t=0}^{T} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right] \right\rangle \leq \min\left\{\frac{3(T+1)}{\alpha}, \frac{12(T+1)(T+2)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.$$

In order to bound the variance error, we first carefully expand the covariance of $\overline{\boldsymbol{\theta}}_{v,t}$ and relate it to the covariances of the $\boldsymbol{\theta}_{v,t}$ (see Lemma 24 in App. C.4). We then control each of these covariances using the following lemma which shows that they are of order $O(t^2)$. See Lemma 25 in App. C.4, Lemma 30 in App. D for proof.

Lemma 7 *For any $t \geq 0$ and step-sizes satisfying condition of Theorem 1, the covariance is characterized by*

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] \preccurlyeq t^2\sigma^2 \begin{bmatrix} 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} & \alpha\beta^{-1}(2\beta-\alpha)\mathbf{I} \\ \alpha\beta^{-1}(2\beta-\alpha)\mathbf{I} & 2\alpha^2\beta^{-1}\mathbf{I} \end{bmatrix}.$$

The lemma is proved by studying $\mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right]/t^2$ in the limit of $t \to \infty$.

Last iterate convergence. The proof for the last iterate follows the same lines and still uses the bias variance decomposition. The main challenge is to bound the bias error. Following Varre et al. (2021), we show a closed recursion where the excess risk at time T can be related to the excess risk of the previous iterates through a discrete Volterra integral as stated in the following lemma.

Lemma 8 (Final Iterate Risk) *Under Assumption 4 and the step-sizes satisfying $\alpha \leq \beta \leq 1/L$. For $T \geq 0$, the last iterate excess error can be determined by the following discrete Volterra integral*

$$\mathsf{f}\left(\boldsymbol{\theta}_{b,T}\right) \leqslant \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \sum_{t=0}^{T-1}\sum_{i=1}^{d} \mathsf{g}\left(\mathbf{H}, t-k\right)\mathsf{f}\left(\boldsymbol{\theta}_{b,t}\right),$$

where $\mathsf{f}\left(\boldsymbol{\theta}_{b,t}\right) \stackrel{\text{def}}{=} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right] \right\rangle$ and the kernel $\mathsf{g}\left(\mathbf{H}, t\right)$ is defined in Eq.(39) in App. C.3.

We recognize here a new bias-variance decomposition. The decrease of the function value $\mathsf{f}\left(\boldsymbol{\theta}_{b,t}\right)$ is controlled by the sum of a term characterizing how fast the initial conditions are forgotten, and a term characterizing how the gradient noise reverberates through the iterates. The final result is then obtained by a simple induction. For the proof, check Lemma 21 in App. C.3 .

7. Experiments

In this section, we illustrate our theoretical findings of Theorems 1, 3 on synthetic data. For $d = 50$, we consider Gaussian distributed inputs a_n with a random covariance $\mathbf{H}$ whose eigenvalues scales as $1/i^4$, for $1 \leq i \leq d$ and optimum $\boldsymbol{x}_*$ which projects equally on the eigenvectors of the covariance. The outputs b_t are generated through $b_t = \langle a_t, \boldsymbol{x}_* \rangle + \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The step-sizes are chosen as $\beta = 1/3\mathrm{Tr}\mathbf{H}$, $\alpha = 1/(3d\ \mathrm{Tr}\mathbf{H})$ for our algorithm; and $\gamma = 1/3\mathrm{Tr}\mathbf{H}$ for SGD. The parameters of ASGD are chosen following Jain et al. (2018b). All results are averaged over 10 repetitions.

Last Iterate. The left plot in Figure 1 corresponds to the convergence of the excess risk of the last iterate on a synthetic noiseless regression, i.e., $\sigma = 0$. We compare Algorithm 4 (AcSGD) with the algorithm of Jain et al. (2018b) (ASGD) and SGD. Note that for the first $O(d)$ iterations, our algorithm matches the performance of SGD. For $t > O(d)$, the acceleration starts and we observe a rate $O(d/t^2)$. Finally, strong convexity takes effect only after a large number of iterations. ASGD decays with a linear rate thereafter.

Averaging. The right plot in Figure 1 corresponds to the performance of the averaged iterate on a noisy least squares problem with $\sigma = 0.02$. We compare our AcSGD with averaging defined in Eq.(5) (AvAcSGD), ASGD with tail averaging (tail-ASGD) and SGD (AvgSGD) with Polyak-Ruppert averaging. For $O(d)$ iterations, our algorithm matches the rate of SGD with averaging, then exhibits an accelerated rate of $O(d/t^2)$. Finally, it decays with the optimal asymptotic rate $\sigma^2 d/t$.

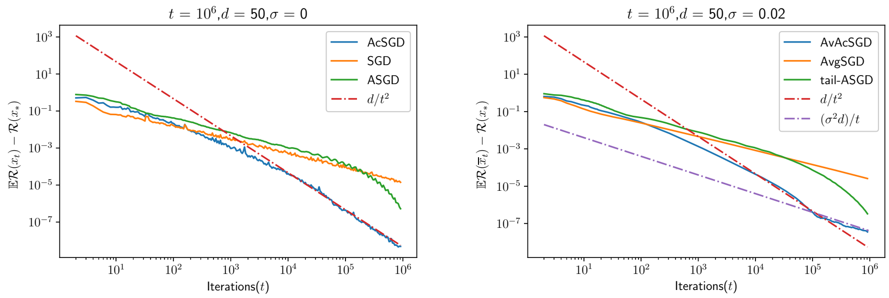

Figure 1: Least-squares regression. Left: Last iterate convergence on a noiseless problem. The plot exhibits a rate $O(d/t^2)$ given by Thm. 3. Right: Averaged iterate convergence on a noisy problem. The plot first exhibits a rate $O(d/t^2)$ and then the optimal rate $O(\sigma^2 d/t)$ as predicted by Thm. 1.

8. Conclusion

In this paper, we show that stochastic accelerated gradient descent can be made robust to gradient noise in the case of least-squares regression. Our new algorithm, based on a simple step-size modification of the celebrated Nesterov accelerated gradient is the first stochastic algorithm which provably accelerates the convergence of the bias while maintaining the optimal convergence of the variance for non-strongly-convex least-squares. There are a number of further direction worth pursuing. Our current analysis is limited to quadratic functions defined in a Euclidean space. An extension of our analysis to all smooth or self-concordant like functions would broaden the applicability of our algorithm. Finally an extension to Hilbert spaces and kernel-based least-squares regression with estimation rates under the usual non-parametric capacity and source conditions would be an interesting development of this work.

Acknowledgments

The authors thank Loucas Pillaud-Vivien and Keivan Rezaei for valuable discussions.

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inform. Theory*, 58(5): 3235–3249, 2012.

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

Z. Allen-Zhu and E. Hazan. Optimal black-box reductions between optimization objectives. *Advances in Neural Information Processing Systems*, 29:1614–1622, 2016.

Z. Allen-Zhu and L. Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS '17, 2017.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems*, 2013.

R. Berthier, F. Bach, and P. Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. In *Advances in Neural Information Processing Systems*, volume 33, pages 2576–2586, 2020.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, 2008.

A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, page 1647–1655, 2011.

Y. Dagan, G. Kur, and O. Shamir. Space lower bounds for linear prediction in the streaming model. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 929–954, 2019.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3):1171–1183, 2008.

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

A. Dieuleveut and F. Bach. Non-parametric stochastic approximation with large step sizes. *Ann. Statist.*, 44(4):1363–1399, 2015.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *J. Mach. Learn. Res.*, 18:101:1–101:51, 2017.

M. Even, R. Berthier, F. Bach, N. Flammarion, H. Hendrikx, P. Gaillard, L. Massoulié, and A. Taylor. Continuized accelerations of deterministic and stochastic gradient descents, and of gossip algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695. PMLR, 2015.

N. Flammarion and F. Bach. Stochastic composite least-squares regression with convergence rate $o(1/n)$. *Conference on Learning Theory*, pages 831–875, 2017.

R. Frostig, R. Ge, S. Kakade, and A. Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2540–2548. PMLR, 07–09 Jul 2015.

S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. *Electron. J. Stat.*, 12(1):461–529, 2018.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM J. Optim.*, 22(4):1469–1492, 2012.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms. *SIAM J. Optim.*, 23(4): 2061–2089, 2013.

A. Greenbaum. Behavior of slightly perturbed lanczos and conjugate-gradient recurrences. *Linear Algebra and its Applications*, 113:7–63, 1989. ISSN 0024-3795.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2006.

M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49:409–436, 1952.

D. Hsu, S. M Kakade, and T. Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

C. Hu, W. Pan, and J. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018a.

P. Jain, S. M Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018b.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133 (1):365–397, 2012.

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

S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.

A. Nemirovskij and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. In *Doklady an ussr*, volume 269, pages 543–547, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic, 2004. A basic course.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005.

C. Paquette and E. Paquette. Dynamics of stochastic momentum methods on large-scale, quadratic models. In *Advances in Neural Information Processing Systems*, 2021.

B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.

B. T. Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika*, 51(7): 98–107, 1990.

B. T Polyak and A. B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

O. Sebbouh, R. M Gower, and A. Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, 2021.

V. Sharan, A. Sidford, and G. Valiant. Memory-sample tradeoffs for linear regression with small error. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing*, STOC, page 890–901, 2019.

A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2003.

A. Varre, L. Pillaud-Vivien, and N. Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. In *Advances in Neural Information Processing Systems*, 2021.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

K. Yuan, B. Ying, and A. H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016.

D. Zou, J. Wu, V. Braverman, Q. Gu, and S. M. Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4633–4635, 2021.

Appendix A. Further Setup and Preliminaries

Organization. The appendix is organized as follows,

- In Section A, we extend the setup of the problem in App. A.1 and introduce operators in App. A.4 to study the recursions of covariance of the estimator. In the later subsection, we give the proof for bias variance decomposition.
- In Section B, we give the proofs for Theorem 1, Theorem 3 and Lemma 5.
- In Section C, we study the recursions on expected covariance of the bias and the variance processes.
- In Section D, we investigate the properties of the operators. In particular, we are interested in inverting few operators.
- In Section E, we study the summations of geometric series of a particular 2×2 matrix by considering its eigendecomposition.

A.1. Preliminaries

Notations. We denote the stream of i.i.d samples by $(a_i, b_i)_{i \geq 1}$. We use $\otimes$ to denote the tensor product and $\otimes_k$ to denote the Kronecker product. Let $\boldsymbol{\mathcal{F}}_t$ denote the filtration generated by the samples $\left\{(a_i, b_i)_{i=1}^t\right\}$.

Additive and Multiplicative Noise. Define for $t \geq 1$,

$$\eta_t = b_t - \langle \boldsymbol{x}_*, a_t \rangle \,, \tag{15}$$

since $\boldsymbol{x}_*$ is the optimum, from the first order optimality of $\boldsymbol{x}_*$,

$$\mathbb{E}\left[\eta_t a_t\right] = 0. \tag{16}$$

In context of least squares, SGD oracle can be written as follows. Let (a_t, b_t) be the sample at iteration t, the gradient at $\boldsymbol{x}_t$ is

$$\begin{aligned}\nabla_t \mathcal{R}\left(\boldsymbol{x}_t\right) &= a_t\left(\langle a_t, \boldsymbol{x}_t\rangle - b_t\right),\\ &= a_t\left(\langle a_t, \boldsymbol{x}_t\rangle - \left(\langle a_t, \boldsymbol{x}_*\rangle + \eta_t\right)\right) = a_t a_t^\top\left(\boldsymbol{x}_t - \boldsymbol{x}_*\right) - \eta_t a_t.\end{aligned}$$

From this, the stochastic gradient can be written as

$$\nabla_t \mathcal{R}\left(\boldsymbol{x}_t\right) = a_t a_t^\top\left(\boldsymbol{x}_t - \boldsymbol{x}_*\right) - \eta_t a_t. \tag{17}$$

As the exact gradient will be $\mathbf{H}\left(\boldsymbol{x}_t - \boldsymbol{x}_*\right)$ the noise in the oracle is

$$\left(\mathbf{H}\left(\boldsymbol{x}_t - \boldsymbol{x}_*\right)\right) - \left(a_t a_t^\top\left(\boldsymbol{x}_t - \boldsymbol{x}_*\right) - \eta_t a_t\right) = \left(\mathbf{H} - a_t a_t^\top\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_*\right) + \eta_t a_t.$$

Note that the zero mean noise $(\mathbf{H} - a_t a_t^\top)$ is multiplicative in nature and hence called multiplicative noise. The zero mean noise $\eta_t a_t$ is called additive noise. In the work of Dieuleveut et al. (2017), stochastic gradients of form $\mathbf{H}\left(\boldsymbol{x}_t - \boldsymbol{x}_*\right) + \epsilon$ for some bounded-variance random variable ϵ are considered. Hence, the results holds only in the case of stochastic oracles with additive noise.

A.2. Discussion on $\tilde{\kappa}$

Lowerbound on $\tilde{\kappa}$. From, assumption 3, we have

$$\mathbb{E}\Big[\left\|a\right\|_{\mathbf{H}^{-1}}^2 \, a \otimes a\Big] \preccurlyeq \tilde{\kappa}\mathbf{H}.$$

Multiplying both sides with $\mathbf{H}^{-1}$ and taking the trace preserves the inequality and we get,

$$\begin{aligned}
&\mathrm{Tr}\Big(\mathbb{E}\Big[\left\|a\right\|_{\mathbf{H}^{-1}}^2 \, \mathbf{H}^{-1}[a \otimes a]\Big]\Big) \leq \ \tilde{\kappa}\mathrm{Tr}\big(\mathbf{H}^{-1}\mathbf{H}\big) \leq \tilde{\kappa}d. \\
&\mathrm{Tr}\Big(\mathbb{E}\Big[\left\|a\right\|_{\mathbf{H}^{-1}}^2 \, \mathbf{H}^{-1}[a \otimes a]\Big]\Big) = \mathbb{E}\Big[\left\|a\right\|_{\mathbf{H}^{-1}}^2 \, \mathrm{Tr}\big[\mathbf{H}^{-1}[a \otimes a]\big]\Big] = \mathbb{E}\left[\left[\left\|a\right\|_{\mathbf{H}^{-1}}^2\right]^2\right]
\end{aligned}$$

Using the fact $\mathbb{E}\left[X^2\right] > \mathbb{E}\left[X\right]^2$, for any random variable X, we get the following,

$$\begin{aligned}
\mathbb{E}\left[\left[\left\|a\right\|_{\mathbf{H}^{-1}}^2\right]^2\right] &\geq \mathbb{E}\left[\left[\left\|a\right\|_{\mathbf{H}^{-1}}^2\right]\right]^2 = \mathrm{Tr}\Big(\mathbf{H}^{-1}\mathbb{E}\left[aa^\top\right]\Big)^2 \qquad \text{, using } \mathbb{E}\left[aa^\top\right] = \mathbf{H}, \\
&= \mathrm{Tr}\big(\mathbf{H}^{-1}\mathbf{H}\big)^2 = d^2
\end{aligned}$$

Combining these, we get,

$$\tilde{\kappa}d \geq d^2 \implies \tilde{\kappa} \geq d.$$

Note that the above calculation only holds when $\mathbf{H}$ is full rank otherwise we can do the same calculation with pseudoinverse in place of $\mathbf{H}^{-1}$ and get a lower bound of $rank(\mathbf{H})$ i.e. number of non-zero eigen values on $\tilde{\kappa}$.

Overparameterized Linear Regression. In this subsection, we focus on $\tilde{\kappa}$ in training overparameterized linear regressions. Note that the finite sum structure can be rewritten $\mathbb{E}_{\hat{\rho}}\left(\langle x, a\rangle - b\right)^2$ where $\hat{\rho}$ is the empirical distribution which is uniform over training data. Let $\{a_j\}_{j=1}^n, a_j \in \mathbb{R}^d$ be the set of training data and $\hat{\rho}_x$, which is a marginal of $\hat{\rho}$ will be uniform distribution over a_j's. Let $A \in R^{n\times d}$ be the the matrix where j^{th} row is $a_j^\top$. Since the regression is overparameterized $n < d$. The covariance in this case is

$$\mathbf{H} = \frac{1}{n}\sum_{j=1}^{n} a_j a_j^\top = \frac{1}{n}(A^\top A)$$

The expression on the left hand side of Asmp. 3 is

$$\mathbb{E}_{a\sim\hat{\rho}}\left[\left\|a\right\|_{\mathbf{H}^\dagger}^2 \ aa^\top\right],$$

where $\mathbf{H}^{\dagger}$ is the pseudo inverse. This is a general version of the assumption 3 with $\mathbf{H}^{\dagger}$ in place of $\mathbf{H}^{-1}$. In this case, we show that asmp. 3 holds with $\tilde{\kappa} = n$. The proof rests entirely on Lemma A.1. in Liu and Wright (2016), we just restate here for completeness. For any vector $x \in \mathbb{R}^d$,

$$\begin{aligned} x^\top \mathbb{E}_{a\sim\hat{\rho}}\left[\left\|a\right\|^2_{\mathbf{H}^{\dagger}} \ aa^\top\right] x &= \frac{1}{n}\sum_{j=1}^{n}\left\|a_j\right\|^2_{\mathbf{H}^{\dagger}} \ x^\top a_j a_j^\top x = \frac{1}{n}\sum_{j=1}^{n}\left\|a_j\right\|^2_{\mathbf{H}^{\dagger}} \ (x^\top a_j)^2, \\ &= \frac{1}{n}\sum_{j=1}^{n}\mathrm{Tr}\Big(\mathbf{H}^{\dagger} a_j a_j^\top (x^\top a_j)^2\Big), \\ &= \mathrm{Tr}\left(\mathbf{H}^{\dagger}\ \frac{1}{n}\sum_{j=1}^{n} a_j a_j^\top (x^\top a_j)^2\right), \end{aligned}$$

Since $\mathbf{H} = \frac{1}{n}(A^\top A)$, we have $\mathbf{H}^{\dagger} = n(A^\top A)^{\dagger}$. Using the notation of diagonal matrix $D = (Diag(Ax))^2$ and the other expression can be compactly written as follows

$$= \mathrm{Tr}\Big((A^\top A)^{\dagger} A^\top D A\Big).$$

Using same approach as Liu and Wright (2016), consider the singular value decomposition of $A = U\Sigma V^T$. We consider the compact singular value decomposition where $U^\top U = I, V^\top V = I$ where I is identity matrix of dimension same as the rank of A. Let $U^\top = [u_1, u_2, \cdots, u_n]$, it is easy to show that $\left\|u_i\right\| \leq 1$, for $i = 1, \ldots, n$. Using this we get,

$$\mathrm{Tr}\Big((A^\top A)^{\dagger} A^\top D A\Big) = \mathrm{Tr}\Big(V\Sigma^{-2}V^\top \ (V\Sigma U^\top) D (U\Sigma V)\Big),$$

using the cyclic property of trace we get,

$$\begin{aligned} \mathrm{Tr}\Big(V\Sigma^{-2}V^\top \ (V\Sigma U^\top) D (U\Sigma V)\Big) = \mathrm{Tr}\Big(U^\top D U\Big) &= \sum_{j=1}^{n}(a_j^\top x)^2\left\|u_i\right\|^2, \\ &\leq \sum_{j=1}^{n}(a_j^\top x)^2 = x^\top(\sum_{j=1}^{n} a_j a_j^\top)x = n\, x^\top \mathbf{H} x. \end{aligned}$$

As this holds for any x, we have

$$\mathbb{E}_{a\sim\hat{\rho}}\left[\left\|a\right\|^2_{\mathbf{H}^{\dagger}} \ aa^\top\right] \preccurlyeq n\mathbf{H}.$$

Hence, for the case of linear regression in overparameterized setting, assumption 3 holds for $\tilde{\kappa} = n$.

A.3. Recursion after Rescaling

Using Eq.(17), we can write Algorithm 4 as follows,

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \beta a_t a_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_*) + \beta \eta_t a_t, \tag{18}$$

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \alpha(t+1) a_t a_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_*) + \alpha(t+1)\eta_t a_t, \tag{19}$$

$$(t+2)\boldsymbol{x}_{t+1} = (t+1)\boldsymbol{y}_{t+1} + \boldsymbol{z}_{t+1}. \tag{20}$$

Recalling the time rescaling of the iterates

$$\boldsymbol{u}_t = (t+1)(\boldsymbol{x}_t - \boldsymbol{x}_*), \tag{21}$$
$$\boldsymbol{v}_t = t(\boldsymbol{y}_t - \boldsymbol{x}_*), \tag{22}$$
$$\boldsymbol{w}_t = \boldsymbol{z}_t - \boldsymbol{x}_*. \tag{23}$$

Now we rewrite the recursion using these rescaled iterates. Multiplying Eq.(18) by $t+1$, and using Eq.(21) and Eq.(22), we get,

$$\begin{aligned}
& & \boldsymbol{v}_{t+1} &= \boldsymbol{u}_t - \beta a_t a_t^\top \boldsymbol{u}_t + \beta \eta_t a_t (t+1), \\
&\text{using Eq.(19) and Eq.(21),} & \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \alpha a_t a_t^\top (\boldsymbol{u}_t) + \alpha (t+1) \eta_t a_t, \\
&\text{from Eq.(20),} & \boldsymbol{u}_t &= \boldsymbol{v}_t + \boldsymbol{w}_t, \\
& & \boldsymbol{v}_{t+1} &= \left(\mathbf{I} - \beta a_t a_t^\top\right)(\boldsymbol{v}_t + \boldsymbol{w}_t) + \beta \eta_t a_t (t+1), \\
& & \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \alpha a_t a_t^\top (\boldsymbol{v}_t + \boldsymbol{w}_t) + \alpha (t+1) \eta_t a_t.
\end{aligned}$$

Writing these updates compactly in form of a matrix recursion gives,

$$\begin{bmatrix} \boldsymbol{v}_{t+1} \\ \boldsymbol{w}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \beta a_t a_t^\top & \mathbf{I} - \beta a_t a_t^\top \\ -\alpha a_t a_t^\top & \mathbf{I} - \alpha a_t a_t^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_t \end{bmatrix} + (t+1)\,\eta_t \begin{bmatrix} \beta a_t \\ \alpha a_t \end{bmatrix}.$$

The above recursion can be written as follows,

$$\boldsymbol{\theta}_{t+1} = \mathcal{J}_t \boldsymbol{\theta}_t + \boldsymbol{\epsilon_{t+1}}, \tag{24}$$

where we defined $\boldsymbol{\theta}_t \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_t \end{bmatrix}$, the random matrix $\mathcal{J}_t \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I} - \beta a_t a_t^\top & \mathbf{I} - \beta a_t a_t^\top \\ -\alpha a_t a_t^\top & \mathbf{I} - \alpha a_t a_t^\top \end{bmatrix}$ and the random noise vector $\boldsymbol{\epsilon_{t+1}} \stackrel{\text{def}}{=} (t+1)\,\eta_t \begin{bmatrix} \beta a_t \\ \alpha a_t \end{bmatrix}$.

Excess Risk of the estimator. The excess risk of any estimate $\boldsymbol{x}$ can be written as

$$\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{x_*}) = \frac{1}{2} \left\langle \boldsymbol{x} - \boldsymbol{x}_*, \mathbf{H}(\boldsymbol{x} - \boldsymbol{x}_*) \right\rangle. \tag{25}$$

Our estimator is defined in Eq.(5) as a time-weighted averaged. Recall,

$$\begin{aligned}
\overline{\boldsymbol{x}}_T &= \frac{\sum_{t=0}^T (t+1) \boldsymbol{x}_t}{\sum_{t=0}^T (t+1)}, \\
\overline{\boldsymbol{x}}_T - \boldsymbol{x}_* &= \frac{\sum_{t=0}^T (t+1)(\boldsymbol{x}_t - \boldsymbol{x}_*)}{\sum_{t=0}^T (t+1)} = \frac{\sum_{t=0}^T \boldsymbol{u}_t}{\sum_{t=0}^T (t+1)}.
\end{aligned}$$

Using the above formulation of excess risk, we have

$$\left(\sum_{t=0}^T (t+1) \right)^2 \cdot \left(\mathcal{R}(\overline{\boldsymbol{x}}_T) - \mathcal{R}(\boldsymbol{x_*}) \right) = \frac{1}{2} \left\langle \sum_{t=0}^T \boldsymbol{u}_t, \mathbf{H} \left(\sum_{t=0}^T \boldsymbol{u}_t \right) \right\rangle.$$

We relate this to the covariance of $\overline{\boldsymbol{\theta}}_T$ in the following way,

$$\overline{\boldsymbol{\theta}}_T = \sum_{t=0}^{T} \boldsymbol{\theta}_t = \sum_{t=0}^{T} \begin{bmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_t \end{bmatrix} = \begin{bmatrix} \sum_{t=0}^{T} \boldsymbol{v}_t \\ \sum_{t=0}^{T} \boldsymbol{w}_t \end{bmatrix}.$$

From Eq.(20) we have the fact that $\boldsymbol{u}_t = \boldsymbol{v}_t + \boldsymbol{w}_t$, for $t \geq 1$. Using this,

$$\sum_{t=0}^{T} \boldsymbol{u}_t = \sum_{t=0}^{T} \boldsymbol{v}_t + \sum_{t=0}^{T} \boldsymbol{w}_t.$$

From the above formulations, with some simple algebra we get,

$$\left\langle \sum_{t=0}^{T} \boldsymbol{u}_t, \mathbf{H} \left(\sum_{t=0}^{T} \boldsymbol{u}_t \right) \right\rangle = \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \overline{\boldsymbol{\theta}}_T \otimes \overline{\boldsymbol{\theta}}_T \right\rangle.$$

Hence, by taking expectation, the excess risk can be related to the covariance of $\overline{\boldsymbol{\theta}}_T$ as,

$$\mathbb{E}\left[\mathcal{R}\left(\overline{\boldsymbol{x}}_T\right)\right] - \mathcal{R}\left(\boldsymbol{x}_*\right) = \frac{1}{2} \left(\sum_{t=0}^{T} (t+1) \right)^{-2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_T \otimes \overline{\boldsymbol{\theta}}_T\right] \right\rangle. \tag{26}$$

Step sizes. We use the following conditions on the step sizes α, β

$$(\alpha + 2\beta) R^2 \leqslant 1, \quad \alpha \leqslant \frac{\beta}{2\tilde{\kappa}}. \tag{27}$$

These conditions are a direct result of our analysis.

Eigen Decomposition of H. Since the covariance is positive definite, the eigendecomposition of $\mathbf{H}$ is given as follows,

$$\mathbf{H} \stackrel{\text{def}}{=} \sum_{i=1}^{d} \lambda_i e_i e_i^\top, \tag{28}$$

where $\lambda_i > 0$'s are the eigenvalues and e_i's are orthonormal eigenvectors.

A.4. Operators

As seen above, the excess risk in expectation can be related to the expected covariance of the $\overline{\boldsymbol{\theta}}_T$,i.e., $\mathbb{E}\left[\overline{\boldsymbol{\theta}}_T \otimes \overline{\boldsymbol{\theta}}_T\right]$. In order to aid the analysis of the covariance, we introduce different operators. The expected value of $\mathcal{J}_t$, denoted by $\mathcal{A} = \mathbb{E}\left[\mathcal{J}_t\right]$ is given by

$$\mathcal{A} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I} - \beta\mathbf{H} & \mathbf{I} - \beta\mathbf{H} \\ -\alpha\mathbf{H} & \mathbf{I} - \alpha\mathbf{H} \end{bmatrix}. \tag{29}$$

If the feature a is sampled according to the marginal distribution of ρ ,i.e., $(a, b) \sim \rho$, define the random matrices $\mathcal{J}$ as follows

$$\mathcal{J} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I} - \beta a a^\top & \mathbf{I} - \beta a a^\top \\ -\alpha a a^\top & \mathbf{I} - \alpha a a^\top \end{bmatrix}. \tag{30}$$

Note that $\mathcal{J}_t$ from Eq.(24) and $\mathcal{J}$ are identically distributed.

Definition 9 *For any PSD matrix Θ, the operators $\mathcal{T}, \widetilde{\mathcal{T}}, \mathcal{M}$ are defined as follows*

(a) $\mathcal{T} \circ \Theta \overset{\text{def}}{=} \mathbb{E}\left[\mathcal{J}\Theta\mathcal{J}^\top\right]$

(b) $\widetilde{\mathcal{T}} \circ \Theta \overset{\text{def}}{=} \mathcal{A}\Theta\mathcal{A}^\top$

(c) $\mathcal{M} \circ \Theta \overset{\text{def}}{=} \mathbb{E}\left[(\mathcal{J}-\mathcal{A})\,\Theta\,(\mathcal{J}-\mathcal{A})^\top\right]$

We proceed to show a few properties of these operators.

Lemma 10 *For the operators $\mathcal{T}, \widetilde{\mathcal{T}}, \mathcal{M}$, the following properties holds.*

(a) $\mathcal{T}, \widetilde{\mathcal{T}}, \mathcal{M}$ *are symmetric and positive*

(b) $\mathcal{T} = \widetilde{\mathcal{T}} + \mathcal{M}$

The operator $\mathcal{O}$ is defined as positive if for any PSD matrix Θ, $\mathcal{O} \circ \Theta$ is also PSD.

Proof For any vector ν, consider the following scalar product,

$$\left\langle \nu, \mathcal{J}\Theta\mathcal{J}^\top\nu \right\rangle = \left\langle \left(\mathcal{J}^\top\nu\right),\ \Theta\ \left(\mathcal{J}^\top\nu\right) \right\rangle.$$

This quantity is non-negative as Θ is a PSD. Hence $\mathcal{T}$ is positive. Similarly the other two operators are also positive. For the second statement,

$$\mathbb{E}\left[(\mathcal{J}-\mathcal{A})\,\Theta\,(\mathcal{J}-\mathcal{A})^\top\right] = \mathbb{E}\left[\mathcal{J}\Theta\mathcal{J}^\top\right] - \mathbb{E}\left[\mathcal{J}\Theta\mathcal{A}^\top\right] - \mathbb{E}\left[\mathcal{A}\Theta\mathcal{J}^\top\right] + \mathbb{E}\left[\mathcal{A}\Theta\mathcal{A}^\top\right].$$

Using $\mathcal{A} = \mathbb{E}\left[\mathcal{J}\right]$,

$$\mathbb{E}\left[(\mathcal{J}-\mathcal{A})\,\Theta\,(\mathcal{J}-\mathcal{A})^\top\right] = \mathbb{E}\left[\mathcal{J}\Theta\mathcal{J}^\top\right] - \mathbb{E}\left[\mathcal{A}\Theta\mathcal{A}^\top\right].$$

This completes the proof of the lemma. ■

Remark 11 *For any PSD matrix Θ and any operators $\mathcal{T}, \mathcal{M}, \widetilde{\mathcal{T}}$, the transpose is defined as following,*

(a) $\mathcal{T}^\top \circ \Theta \overset{\text{def}}{=} \mathbb{E}\left[\mathcal{J}^\top\Theta\mathcal{J}\right].$

(b) $\widetilde{\mathcal{T}}^\top \circ \Theta \overset{\text{def}}{=} \mathcal{A}^\top\Theta\mathcal{A}.$

(c) $\mathcal{M}^\top \circ \Theta \overset{\text{def}}{=} \mathbb{E}\left[(\mathcal{J}-\mathcal{A})^\top\,\Theta\,(\mathcal{J}-\mathcal{A})\right].$

Having introduced operators, we present a lemma which is central to the analysis, we give a almost eigenvector and eigenvalue of the operators. We call it an almost eigenvector as only an upperbound holds in this case.

Lemma 12 *For step sizes satisfying Condition 27, the following properties hold on the inverse and eigen values of operators $\mathcal{T}, \widetilde{\mathcal{T}}, \mathcal{M}$*

(a) *For stepsizes* $0 < \alpha, \beta < \frac{1}{L}$, $(1 - \widetilde{\mathcal{T}})^{-1}$ *exists.*

(b) $\mathbf{\Xi}$ *is an almost eigen vector of* $\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1}$ *with an eigen value less than 1 and* $(1 - \mathcal{T})^{-1} \circ \mathbf{\Xi}$ *exists,*

$$\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1} \circ \mathbf{\Xi} \preccurlyeq \frac{2}{3}\mathbf{\Xi}.$$

(c) Υ *is an almost eigen vector of* $\left(\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1}\right)$ *with an eigen value less than 1,*

$$\left(\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1}\right) \circ \Upsilon \preccurlyeq \frac{2}{3}\Upsilon.$$

where

$$\Upsilon \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix} \qquad \mathbf{\Xi} \stackrel{\text{def}}{=} \begin{bmatrix} \beta^2 \mathbf{H} & \alpha\beta\mathbf{H} \\ \alpha\beta\mathbf{H} & \alpha^2\mathbf{H} \end{bmatrix}. \tag{31}$$

Proof From the diagonalization of the covariance $\mathbf{H}$ from Eq.(28), note that $\mathcal{A}$ can be diagonalized as follows

$$\mathcal{A} = \sum_{i=1}^{d} \begin{bmatrix} 1 - \beta\lambda_i & 1 - \beta\lambda_i \\ -\alpha\lambda_i & 1 - \alpha\lambda_i \end{bmatrix} \otimes_k e_i e_i^\top.$$

From Property 2, $0 < \alpha, \beta < \frac{1}{L}$ the absolute value of $\mathcal{A}$ eigen values will be less than 1 and for any PSD matrix Θ the inverse can be defined by the sum of geometric series as follows,

$$(1 - \widetilde{\mathcal{T}})^{-1} \circ \Theta = \sum_{t \geq 0} \mathcal{A}^t \Theta \left(\mathcal{A}^\top\right)^t.$$

To compute $(1 - \widetilde{\mathcal{T}})^{-1} \circ \mathbf{\Xi}$, although the calculations are a bit extensive, the underlying scheme remains the same. After formulating the inverse as a sum of geometric series, we use the diagonalization of the $\mathcal{A}$ and $\mathbf{\Xi}$ to compute the geometric series. In the last part to compute $\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1}$, we use Property 1 and Assumptions 1, 3 on distribution ρ along with the conditions on the stepsizes Eq.(27) to get the final bounds. The remaining parts can be proven using Lemmas 30, 33, 34. ■

Bias-Variance Decomposition. Recall the bias recursion Eq.(12), the variance recursion Eq.(13).

$$\begin{aligned} \boldsymbol{\theta}_{b,t+1} &= \mathcal{J}_t \boldsymbol{\theta}_{b,t} \text{ started from } \boldsymbol{\theta}_{b,0} = \boldsymbol{\theta}_0, \\ \boldsymbol{\theta}_{v,t+1} &= \mathcal{J}_t \boldsymbol{\theta}_{v,t} + \boldsymbol{\epsilon_{t+1}} \text{ started from } \boldsymbol{\theta}_{v,0} = \mathbf{0}. \end{aligned}$$

Now we prove a bias-variance decomposition lemma. Similar lemmas have been derived in the works of Bach and Moulines (2013); Jain et al. (2018b). Following the proof in these works, we re-derive it here for the sake of completeness.

Lemma 13 *For $T \geq 0$, the expected covariance of $\overline{\boldsymbol{\theta}}_T$ can be bounded as follows,*

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_T \otimes \overline{\boldsymbol{\theta}}_T\right] \preccurlyeq 2\left(\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{b,T} \otimes \overline{\boldsymbol{\theta}}_{b,T}\right] + \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right]\right). \tag{32}$$

Proof In the first part, using induction we prove that $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{b,t} + \boldsymbol{\theta}_{v,t}$. Note that the hypothesis holds at $k = 0$ because $\boldsymbol{\theta}_{b,0} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_{v,0} = \mathbf{0}$. Assume that $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{b,t} + \boldsymbol{\theta}_{v,t}$ holds for $0 \leq t \leq k-1$. We prove that hypothesis also holds for k. From the recursion on $\boldsymbol{\theta}_t$ in Eq.(24), we get,

$$\begin{aligned}
\boldsymbol{\theta}_k &= \mathcal{J}_{k-1}\boldsymbol{\theta}_{k-1} + \boldsymbol{\epsilon_k}, \\
\boldsymbol{\theta}_k &= \mathcal{J}_{k-1}\left(\boldsymbol{\theta}_{b,k-1} + \boldsymbol{\theta}_{v,k-1}\right) + \boldsymbol{\epsilon_k}, \quad \text{from induction hypothesis,} \\
&= \mathcal{J}_{k-1}\boldsymbol{\theta}_{b,k-1} + \mathcal{J}_{k-1}\boldsymbol{\theta}_{v,k-1} + \boldsymbol{\epsilon_k}.
\end{aligned}$$

Form the recursion of bias and variance Eq.(12), Eq.(13). We show that $\boldsymbol{\theta}_k = \boldsymbol{\theta}_{b,k} + \boldsymbol{\theta}_{v,k}$. From induction, this is true for all $k \geq 0$. Summing these equalities from $k = 0, \ldots, T$, we get,

$$\overline{\boldsymbol{\theta}}_T = \overline{\boldsymbol{\theta}}_{b,T} + \overline{\boldsymbol{\theta}}_{v,T}.$$

Using the Cauchy Schwarz inequality and then taking expectation, we get the statement of the lemma.
■

Recursions on Covariance. In the following lemma, we show how the recursions on the expected covariance of the bias and variance processes are governed by the operators defined above.

Lemma 14 *For $t \geq 0$, the recursion on the covariance satisfies*

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\theta}_{b,t+1} \otimes \boldsymbol{\theta}_{b,t+1}\right] &= \mathcal{T} \circ \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right], \\
\mathbb{E}\left[\boldsymbol{\theta}_{v,t+1} \otimes \boldsymbol{\theta}_{v,t+1}\right] &= \mathcal{T} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_{t+1}} \otimes \boldsymbol{\epsilon_{t+1}}\right].
\end{aligned}$$

Proof From the recursion of the bias process Eq.(12),

$$\boldsymbol{\theta}_{b,t+1} = \mathcal{J}_t\boldsymbol{\theta}_{b,t}.$$

Now the expectation of covariance is

$$\mathbb{E}\left[\boldsymbol{\theta}_{b,t+1} \otimes \boldsymbol{\theta}_{b,t+1}\right] = \mathbb{E}\left[\mathcal{J}_t\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\mathcal{J}_t^\top\right].$$

Note that $\mathcal{J}_t$ is independent of $\boldsymbol{\theta}_{b,t}$. Hence using the definition of operator $\mathcal{T}$ completes the proof of the first part. Now, from the recursion of the variance process Eq.(13),

$$\boldsymbol{\theta}_{v,t+1} = \mathcal{J}_t\boldsymbol{\theta}_{v,t} + \boldsymbol{\epsilon_{t+1}}.$$

As we know that $\boldsymbol{\theta}_{v,0} = \mathbf{0}$ and for $t \geq 1$, $\mathbb{E}\left[\boldsymbol{\epsilon_t}\right] = 0$ from Eq.(16). As $\mathcal{J}_t$ is independent of $\boldsymbol{\theta}_{v,t}$, we get $\mathbb{E}\left[\boldsymbol{\theta}_{v,t+1}\right] = \mathcal{A}\mathbb{E}\left[\boldsymbol{\theta}_{v,t}\right]$. Combining these we have for $t \geq 0$, $\mathbb{E}\left[\boldsymbol{\theta}_{v,t}\right] = \mathbf{0}$. Now, the expectation of the covariance is

$$\begin{aligned}
\boldsymbol{\theta}_{v,t+1} \otimes \boldsymbol{\theta}_{v,t+1} &= \left(\mathcal{J}_t\boldsymbol{\theta}_{v,t} + \boldsymbol{\epsilon_{t+1}}\right) \otimes \left(\mathcal{J}_t\boldsymbol{\theta}_{v,t} + \boldsymbol{\epsilon_{t+1}}\right), \\
\mathbb{E}\left[\boldsymbol{\theta}_{v,t+1} \otimes \boldsymbol{\theta}_{v,t+1}\right] &= \mathbb{E}\left[\left(\mathcal{J}_t\boldsymbol{\theta}_{v,t} + \boldsymbol{\epsilon_{t+1}}\right) \otimes \left(\mathcal{J}_t\boldsymbol{\theta}_{v,t} + \boldsymbol{\epsilon_{t+1}}\right)\right].
\end{aligned}$$

Using the fact that $\mathcal{J}_t, \boldsymbol{\epsilon_{t+1}}$ are independent of $\boldsymbol{\theta}_{v,t}$ and $\mathbb{E}\left[\boldsymbol{\theta}_{v,t}\right] = 0$,

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,t+1} \otimes \boldsymbol{\theta}_{v,t+1}\right] = \mathbb{E}\left[\mathcal{J}_t \;\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right]\; \mathcal{J}_t^{\top}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_{t+1}} \otimes \boldsymbol{\epsilon_{t+1}}\right].$$

Note that $\mathcal{J}_t$ is independent of $\boldsymbol{\theta}_{v,t}$. Hence using the definition of operator $\mathcal{T}$,

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,t+1} \otimes \boldsymbol{\theta}_{v,t+1}\right] = \mathcal{T} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_{t+1}} \otimes \boldsymbol{\epsilon_{t+1}}\right]. \tag{33}$$

■

A.5. Mini-Batching

In this subsection, we discuss how we can use the same proof techniques for the mini-batch stochastic gradient oracles. Recall the mini-batch oracle for some batch size b with samples $(a_{t,i}, b_{t,i}) \sim \rho$, for $1 \le i \le b$,

$$\nabla_t \mathcal{R}\left(\boldsymbol{x}_t\right) = \frac{1}{b}\sum_{i=1}^{b} a_{t,i}\left(\langle a_{t,i}, \boldsymbol{x}_t\rangle - b_{t,i}\right). \tag{34}$$

Following the approach in A.3, we get the time rescaled recursion with

$$\boldsymbol{\theta}_{t+1} = \mathcal{J}_t^{mb}\boldsymbol{\theta}_t + \boldsymbol{\epsilon}^{mb}_{\boldsymbol{t+1}}.$$

where

$$\mathcal{J}_t^{mb} = \frac{1}{b}\sum_{i=1}^{b}\mathcal{J}_{t,i} \quad , \quad \mathcal{J}_{t,i} = \begin{bmatrix} \mathbf{I} - \beta a_{t,i}a_{t,i}^{\top} & \mathbf{I} - \beta a_{t,i}a_{t,i}^{\top} \\ -\alpha a_{t,i}a_{t,i}^{\top} & \mathbf{I} - \alpha a_{t,i}a_{t,i}^{\top} \end{bmatrix},$$

$$\boldsymbol{\epsilon}^{mb}_{\boldsymbol{t+1}} = \frac{1}{b}\sum_{i=1}^{b}\boldsymbol{\epsilon_{t,i}} \quad , \quad \boldsymbol{\epsilon_{t,i}} = \left(b_{t,i} - \langle a_{t,i}, \boldsymbol{x}_*\rangle\right)\begin{bmatrix} \beta a_{t,i} \\ \alpha a_{t,i} \end{bmatrix}.$$

Note that $\mathcal{J}_{t,i}$'s are independent and identically distributed to $\mathcal{J}$ with $\mathbb{E}\left[\mathcal{J}_{t,i}\right] = \mathcal{A}$. Hence, by linearity of expectation $\mathbb{E}\left[\mathcal{J}_t^{mb}\right] = \mathcal{A}$. Now we can define the operators specific to mini-batch oracles. Note that $\widetilde{\mathcal{T}}$ stays the same.

$$\mathcal{T}_{mb} \circ \Theta = \mathbb{E}\left[\mathcal{J}_t^{mb}\Theta(\mathcal{J}_t^{mb})^{\top}\right] \qquad \mathcal{M}_{mb} \circ \Theta = \mathbb{E}\left[\left(\mathcal{J}_t^{mb} - \mathcal{A}\right)\Theta\left(\mathcal{J}_t^{mb} - \mathcal{A}\right)^{\top}\right]. \tag{35}$$

Using the fact that $\mathcal{J}_t^{mb} - \mathcal{A} = \frac{1}{b}\sum_{i=1}^{b}\left(\mathcal{J}_{t,i} - \mathcal{A}\right)$ and $\mathcal{J}_{t,i} - \mathcal{A}$'s are zero mean i.i.d random matrices, it is evident that $\mathcal{M}_{mb} \circ \Theta = \frac{1}{b}\mathcal{M} \circ \Theta$. Hence,

$$\mathcal{M}_{mb} = \frac{1}{b}\mathcal{M}.$$

Using this fact we give a version of Lemma 12 for mini-batch with different step size constraints. Define $\mathcal{M}_{mb}^{\top}$ along the same line as $\mathcal{M}^{\top}$.

Lemma 15 *For step sizes satisfying $0 < \alpha, \beta < \frac{1}{L}$ and $(\alpha + 2\beta)R^2 \leq b, \alpha \leq \frac{\beta b}{2\tilde{\kappa}}$, the following properties hold on the inverse and eigen values of operators $\widetilde{\mathcal{T}}, \mathcal{M}_{mb}$*

a *$\boldsymbol{\Xi}$ is an almost eigen vector of $\mathcal{M}_{mb} \circ (1 - \widetilde{\mathcal{T}})^{-1}$ with an eigen value less than 1,*

$$\mathcal{M}_{mb} \circ (1 - \widetilde{\mathcal{T}})^{-1} \circ \boldsymbol{\Xi} \preccurlyeq \frac{2}{3}\boldsymbol{\Xi}.$$

b *Υ is an almost eigen vector of $\left(\mathcal{M}_{mb}^{\top} \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^{\top}\right)^{-1}\right)$ with an eigen value less than 1,*

$$\left(\mathcal{M}_{mb}^{\top} \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^{\top}\right)^{-1}\right) \circ \Upsilon \preccurlyeq \frac{2}{3}\Upsilon.$$

Proof

Note that $0 < \alpha, \beta < \frac{1}{L}$ is necessary for $(1 - \widetilde{\mathcal{T}})^{-1}$ to exist. The rescaling of the other condition on step-size is due to the fact that $\mathcal{M}_{mb} = \frac{1}{b}\mathcal{M}$. Following the Lemmas 33, 34 with this new operator will give the required condition on the step-sizes. ■

For Theorem 1 with mini-batch oracles, we can follow the original proof of Theorem 1 with stochastic oracle with this new Lemma 15 for the operator $\mathcal{M}_{mb}$.

Appendix B. Proof of the main results

B.1. Proof of Theorem 1

The proof involves three parts. In the first part, we consider the bias recursion and bound the excess risk in the bias process. In the second, we bound the excess error in the variance process. In the last part, we use the bias-variance decomposition and the relation between covariance of $\boldsymbol{\theta}_T$ and excess error of $\overline{\boldsymbol{x}}_T$ from Eq.(26).

Bias Error. For the bias part, we show a relation between the finite sum of covariance of the iterates in case of bias process Eq.(12) with stochastic gradients and the finite sum of covariance of iterates of bias process with exact gradients in Lemma 19. Using the fact that Υ is almost a eigenvector of $\mathcal{M}^{\top} \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^{\top}\right)$ (see Lemma 12) and using the Nesterov Lyapunov techniques to control the sum of the covariance of the iterates of bias process with exact gradients (Lemma 27), we get the sum of excess risk of the bias iterates (see Lemma 20). From here we have,

$$\sum_{t=0}^{T} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right] \right\rangle \leq \min\left\{\frac{3(T+1)}{\alpha}, \frac{12(T+1)(T+2)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.$$

We use the property that $\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \boldsymbol{\theta} \otimes \boldsymbol{\theta} \right\rangle$ is convex in $\boldsymbol{\theta}$. As $\overline{\boldsymbol{\theta}}_{b,T} \overset{\text{def}}{=} \sum_{t=0}^{T} \boldsymbol{\theta}_{b,T}$, applying Jensens inequality,

$$\begin{aligned}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \overline{\boldsymbol{\theta}}_{b,T} \otimes \overline{\boldsymbol{\theta}}_{b,T} \right\rangle &\leq (T+1) \sum_{t=1}^{T} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t} \right\rangle, \\ &\leq \min\left\{\frac{3(T+1)^2}{\alpha}, \frac{12(T+1)^2(T+2)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.\end{aligned}$$

Variance Error We expand the expected covariance of $\overline{\boldsymbol{\theta}}_{v,T}$ in Lemma 24 such that the coefficients of $\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}$ in the formulation are positive and any upper bound on the covariance of iterates $\boldsymbol{\theta}_{v,t}$, for $t \leq T$ would give an upper bound on the expected covariance of $\overline{\boldsymbol{\theta}}_{v,T}$. Then we bound the limiting covariance of the iterates, i.e., $\mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right]/t^2$ in Lemma 25. The fact that $\boldsymbol{\Xi}$ is almost eigen vector of $\mathcal{M} \circ \left(1 - \widetilde{\mathcal{T}}\right)^{-1}$ is used here. Using this upper bound in the above formulation of covariance of $\overline{\boldsymbol{\theta}}_{v,T}$ to give the bound in Lemma 26. From here, we have

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] \right\rangle \leq 18\left(\sigma^2 d\right) T^3.$$

Now using the bias-variance decomposition, we get,

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{T} \otimes \overline{\boldsymbol{\theta}}_{T}\right] \right\rangle \leq 2\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{b,T} \otimes \overline{\boldsymbol{\theta}}_{b,T}\right] \right\rangle + 2\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] \right\rangle.$$

and using the formulation of excess risk of $\overline{\boldsymbol{x}}_T$ with covariance of $\overline{\boldsymbol{\theta}}_T$ from Eq.(26),

$$\mathbb{E}\left[\mathcal{R}\left(\overline{\boldsymbol{x}}_T\right)\right] - \mathcal{R}\left(\boldsymbol{x}_*\right) = \frac{1}{2}\left(\sum_{t=0}^{T}(t+1)\right)^{-2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{T} \otimes \overline{\boldsymbol{\theta}}_{T}\right] \right\rangle.$$

Combining these will prove Theorem 1.

B.2. Proof of Theorem 3

For the last iterate too, we employ the bias-variance decomposition. First the variance part, we use Lemma 7. Note that if Assumption 4 holds with constant κ then Assumption 1 holds with $R^2 = \kappa \mathrm{Tr}\mathbf{H}$. and Assumption 3 holds with $\tilde{\kappa} = \kappa d$. Hence this satisfies the condition on step size required for Lemma 7.

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] \preccurlyeq t^2\sigma^2 \begin{bmatrix} 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} & \alpha\beta^{-1}(2\beta-\alpha)\mathbf{I} \\ \alpha\beta^{-1}(2\beta-\alpha)\mathbf{I} & 2\alpha^2\beta^{-1}\mathbf{I} \end{bmatrix},$$
$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] \right\rangle \leq 2t^2\sigma^2\left((\alpha + 2\beta)\mathrm{Tr}\mathbf{H} + \frac{2\alpha d}{\beta}\right).$$

For the bias, we require uniform kurtosis ,i.e., Assumption 4. Under this assumption, one can related the variance due to stochastic oracle to the excess risk of the iterate (see Lemma 32). Using this we give a closed recursion for the excess risk of the last iterate as a discrete Volterra integral of risk of the previous iterates in Lemma 22. Using simple induction to bound this (note that scaling on step-sizes will be used here) will give,

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t} \right\rangle \leqslant \min\left\{\frac{3}{\alpha}, \frac{24(t+1)}{\beta}\right\}\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2$$

Using the bias-variance decomposition along with the fact that

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right] \right\rangle = 2(t+1)^2 \cdot \left(\mathbb{E}\left[\mathcal{R}\left(\boldsymbol{x}_t\right)\right] - \mathcal{R}\left(\boldsymbol{x}_*\right)\right)$$

This proves Theorem 3.

B.3. Lower Bound

Lemma 16 *For all starting point $\boldsymbol{x}_0$, there exists a distribution ρ over $\mathbb{R}^d \times \mathbb{R}$ satisfying Assumption 1 with $R^2 = 1$, Assumption 2 with $\sigma = 0$, Assumption 3 with $\tilde{\kappa} = d$ and optimum $\boldsymbol{x}'_*$ verifying $\left\|\boldsymbol{x}'_* - \boldsymbol{x}_0\right\|^2 = 1$, such that the expected excess risk of any stochastic first order algorithm is lower bounded as*

$$\mathbb{E}\left[\mathcal{R}\left(\boldsymbol{x}_{\lfloor d/2 \rfloor}\right)\right] = \Omega\left(\frac{1}{d}\right).$$

Proof Let $(e_i)_{i=1}^d$ be a set of orthonormal basis. Define the following

- The optimum

$$\boldsymbol{x}'_* \stackrel{\text{def}}{=} \boldsymbol{x}_0 + \frac{1}{\sqrt{d}} \sum_{i=1}^{d} e_i.$$

  It can be easily verified that $\left\|\boldsymbol{x}'_* - \boldsymbol{x}_0\right\|^2 = 1$.

- The feature distribution ρ where each e_i is sampled with a probability $1/d$. In this case the Hessian $\mathbf{H}' = (d\mathbf{I})^{-1}$. Note that for this distribution $R^2 = 1$. The excess risk at any x is as follows

$$\begin{aligned}
\mathcal{R}(x) &= \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}'_*\right)^\top \mathbf{H}'\left(\boldsymbol{x} - \boldsymbol{x}'_*\right), \\
&= \frac{1}{2d}\sum_{i=1}^{d}\left(\left\langle \boldsymbol{x} - \boldsymbol{x}'_*, e_i\right\rangle\right)^2 = \frac{1}{2d}\sum_{i=1}^{d}\left(\left\langle \boldsymbol{x} - \boldsymbol{x}_0, e_i\right\rangle + \left\langle \boldsymbol{x}_0 - \boldsymbol{x}'_*, e_i\right\rangle\right)^2, \\
&= \frac{1}{2d}\sum_{i=1}^{d}\left(\left\langle \boldsymbol{x} - \boldsymbol{x}_0, e_i\right\rangle + \left\langle \boldsymbol{x}_0 - \boldsymbol{x}'_*, e_i\right\rangle\right)^2, \\
&= \frac{1}{2d}\sum_{i=1}^{d}\left(\left\langle \boldsymbol{x} - \boldsymbol{x}_0, e_i\right\rangle - \frac{1}{\sqrt{d}}\right)^2, \qquad \text{using construction of } \boldsymbol{x}'_*.
\end{aligned}$$

- For $n \geq 1$, $b_n = \langle a_n, \boldsymbol{x}_* \rangle$ where $a_n \sim \rho$. Hence Assumption 2 holds with $\sigma = 0$. From the construction it can be seen that $\tilde{\kappa} = d$.

Consider any stochastic first order algorithm $\mathcal{S}$ for t iterations. Lets say $a_1 = e_{i_1}, \cdots, a_t = e_{i_t}$ be the inputs from the stream till time t. From Definition 4, the estimator $\boldsymbol{x}_t$ after t iterations

$$\boldsymbol{x}_t \in \boldsymbol{x}_0 + \mathit{span}\left\{\nabla_0 f(\boldsymbol{x}_0), \nabla_1 f(\boldsymbol{x}_1), \cdots, \nabla_{k-1} f(\boldsymbol{x}_{k-1})\right\}.$$

Note that for above defined noiseless linear regression the stochastic gradient at time k is $\nabla_k f(\boldsymbol{x}_k) = e_{i_k}\left\langle e_{i_k}, \boldsymbol{x}_k - \boldsymbol{x}'_*\right\rangle$. Using the fact that it is always along the direction of e_{i_k}.

$$e \stackrel{\text{def}}{=} \boldsymbol{x}_{d/2} - \boldsymbol{x}_0 \in \mathit{span}\left\{e_{i_1}, \cdots, e_{i_{d/2}}\right\}.$$

Plugging this in the above expression for excess risk, we get,

$$\mathbb{E}\mathcal{R}\left(\boldsymbol{x}_t\right) = \frac{1}{2d}\sum_{i=1}^{d}\mathbb{E}\left(\langle e, e_i\rangle - \frac{1}{\sqrt{d}}\right)^2.$$

From the construction of ρ', e is in the span of $d/2$ orthogonal features. Hence, the remaining $d/2$ directions contribute to the excess error. In technical terms, let $\mathcal{P}$ be the set $\left\{e_{i_1}, \ldots, e_{i_{d/2}}\right\}$. Note that $|\mathcal{P}| = d/2$. Then,

$$\begin{aligned}\mathbb{E}\mathcal{R}\left(\boldsymbol{x}_t\right) &= \frac{1}{2d}\sum_{i=1}^{d}\mathbb{E}\left(\langle e, e_i\rangle - \frac{1}{\sqrt{d}}\right)^2, \\ &\geq \frac{1}{2d}\sum_{e\notin\mathcal{P}}\frac{1}{d} = \frac{d-|\mathcal{P}|}{2d^2} = \frac{1}{4d}.\end{aligned}$$

■

Lemma 17 *For any initial point $\boldsymbol{x}_0$ and Hessian $\mathbf{H}'$ with $\mathrm{Tr}(\mathbf{H}') = 1$, there exists a distribution ρ' which prevent acceleration.*

Proof Let

$$\mathbf{H}' = \sum_{i=1}^{d} p_i e_i e_i^\top.$$

The excess risk on any noise less problem with $\boldsymbol{x}'_*$ as optimum and $\mathbf{H}'$ as Hessian can be written as,

$$\begin{aligned}\mathcal{R}\left(\boldsymbol{x}\right) &= \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}'_*\right)^\top \mathbf{H}'\left(\boldsymbol{x} - \boldsymbol{x}'_*\right), \\ &= \frac{1}{2}\sum_{i=1}^{d} p_i\left(\langle \boldsymbol{x} - \boldsymbol{x}'_*, e_i\rangle\right)^2 = \frac{1}{2d}\sum_{i=1}^{d}\left(\langle \boldsymbol{x} - \boldsymbol{x}_0, e_i\rangle + \langle \boldsymbol{x}_0 - \boldsymbol{x}'_*, e_i\rangle\right)^2, \\ &= \frac{1}{2}\sum_{i=1}^{d} p_i\left(\langle \boldsymbol{x} - \boldsymbol{x}_0, e_i\rangle + \langle \boldsymbol{x}_0 - \boldsymbol{x}'_*, e_i\rangle\right)^2.\end{aligned}$$

Let ρ' be the one hot basis distribution where e_i is sampled with probability p_i. Consider any stochastic first order algorithm $\mathcal{S}$ for t iterations. Lets say $e_{i_1}, \cdots, e_{i_t}$ be the inputs from the stream till time t. From Definition 4, the estimator $\boldsymbol{x}_t$ after t iterations

$$\boldsymbol{x}_t \in \boldsymbol{x}_0 + \mathit{span}\left\{\nabla_0 f(\boldsymbol{x}_0), \nabla_1 f(\boldsymbol{x}_1), \cdots, \nabla_{k-1} f(\boldsymbol{x}_{k-1})\right\}.$$

Note that for noiseless regression the stochastic gradient at time k is $\nabla_k f(\boldsymbol{x}_k) = e_{i_k}(\langle e_{i_k}, \boldsymbol{x}_k - \boldsymbol{x}_*\rangle$. Using the fact that it is always along the direction of e_{i_k}.

$$e \overset{\text{def}}{=} \boldsymbol{x}_t - \boldsymbol{x}_0 \in \mathit{span}\left\{e_{i_1}, \cdots, e_{i_t}\right\}.$$

Plugging this in the above expression for excess risk, we get,

$$\mathbb{E}\mathcal{R}\left(\boldsymbol{x}_t\right) = \frac{1}{2}\sum_{i=1}^{d} p_i \mathbb{E}\left(\langle e, e_i\rangle + \left\langle \boldsymbol{x}_0 - \boldsymbol{x}'_*, e_i\right\rangle\right)^2.$$

If none of the e_{i_k}'s , for $k \leq t$ are e_i then $\langle e, e_i\rangle = 0$. This event occurs with a probability $(1-p_i)^t$. Hence, with probability $(1-p_i)^t$, $\langle e, e_i\rangle = 0$. Taking this into consideration,

$$\mathbb{E}\left(\langle e, e_i\rangle + \left\langle \boldsymbol{x}_0 - \boldsymbol{x}'_*, e_i\right\rangle\right)^2 \geq (1-p_i)^t \left\langle \boldsymbol{x}_0 - \boldsymbol{x}'_*, e_i\right\rangle^2$$

Hence,

$$\mathbb{E}\mathcal{R}\left(\boldsymbol{x}_t\right) \geq \frac{1}{2}\sum_{i=1}^{d} p_i \left(1-p_i\right)^t \left\langle \boldsymbol{x}_0 - \boldsymbol{x}'_*, e_i\right\rangle^2$$

Noting that the right hand side corresponds to the performance of gradient descent with step size 1 after $t/2$ iterations. In conclusion, performance of gradient descent is better than any stochastic algorithm. Hence, direct acceleration with this oracle defined by ρ' is not possible. ■

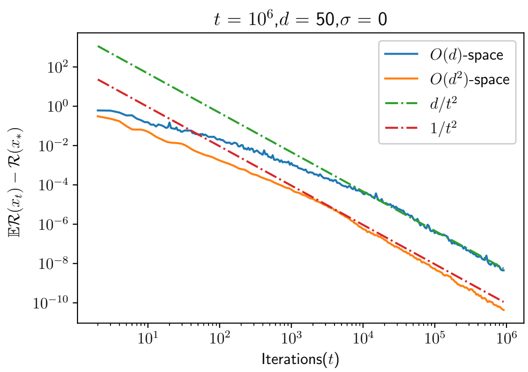

Figure 2: Least-squares regression. Comparison of Alg. 4 with space constrains. Note that the version with $O(d^2)$-space decrease at rate $1/t^2$ where Alg. 4 with $O(d)$ decays at rate d/t^2.

Space Complexity. In Figure 2, we demonstrate that the with additional space the speed of the decay can be improved. Note that the version with $O(d^2)$-space decrease at rate $1/t^2$ where Algorithm 4 with $O(d)$ decays at rate d/t^2. The set up for this experiment is same as the setup of the plot on Last Iterate described in Section 7. The $O(d) - curve$ corresponds to the Algorithm 4 with SGD oracle Eq.(6) with step sizes $\alpha = 1/3\, d\mathrm{Tr}\mathbf{H}$, $\beta = 1/3\mathrm{Tr}\mathbf{H}$ where $\mathbf{H}$ is the covariance of Gaussian data. The $O(d^2) - curve$ corresponds to the Algorithm 4 with running average SGD oracle in Section 5 with step sizes $\alpha = 1/3\mathrm{Tr}\mathbf{H}$, $\beta = 1/3\mathrm{Tr}\mathbf{H}$.

Appendix C. Bias and Variance

In this section, we investigate the recursions of expected covariance of the bias and variance process.

C.1. Our technique for Bias

The work of Jain et al. (2018b) introduces a novel Lyapunov function $c_1\mathbb{E}\left[\left\|\boldsymbol{y}_t-\boldsymbol{x}_*\right\|^2\right]+c_2\mathbb{E}\left[\left\|\boldsymbol{z}_t-\boldsymbol{x}_*\right\|^2_{\mathbf{H}^{-1}}\right]$, for some constants c_1, c_2 for the analysis of bias error to show accelerated SGD rates for strongly convex least squares. Using similar Lyapunov function on the non-strongly convex version of Nesterov acceleration algorithm, in Even et al. (2021), a rate of convergence for $\mathbb{E}\left[\left\|\boldsymbol{x}_t-\boldsymbol{x}_*\right\|^2\right]$,i.e,

$$\mathbb{E}\left[\left\|\boldsymbol{x}_t-\boldsymbol{x}_*\right\|^2\right] \lesssim \frac{\left\|\boldsymbol{x}_0-\boldsymbol{x}_*\right\|^2_{\mathbf{H}^{-1}}}{t^2}.$$

is shown. Even with this result, it is still unclear how to relate excess error $\mathbb{E}\left[\left\|\boldsymbol{x}_t-\boldsymbol{x}_*\right\|^2_{\mathbf{H}}\right]$ and distance of initialization $\left\|\boldsymbol{x}_0-\boldsymbol{x}_*\right\|^2$. Note that for non-strongly convex functions $\left\|\boldsymbol{x}_0-\boldsymbol{x}_*\right\|^2_{\mathbf{H}^{-1}}$ can be arbitrarily large in comparison to $\left\|\boldsymbol{x}_0-\boldsymbol{x}_*\right\|^2$. As there is an absence of direct Lyapunov techniques for bias error, it is needed to introduce a new method.

In recent times, many works (Zou et al., 2021; Varre et al., 2021) have studied the sharp characterization of bias process in SGD to understand the performance of SGD for over-parameterized least squares. In Zou et al. (2021), it is shown that sum of covariance i.e. $\sum_{i=0}^{t}\mathbb{E}\left[\boldsymbol{\theta}_{b,i}\otimes\boldsymbol{\theta}_{b,i}\right]$ of SGD at the limit $t\to\infty$ is used to give sharp bounds for bias excess risk. Even this approach cannot be used to in our case for two reasons (a) the limit in the case of our accelerated methods still depends on $\left\|\boldsymbol{x}_0-\boldsymbol{x}_*\right\|^2_{\mathbf{H}^{-1}}$ which can be arbitrarily large (b) this requires more restricting uniformly bounded kurtosis assumption. In our approach we give sharp estimates for finite sum of covariance and relate them to the sum of covariance for the Algorithm 4 with exact gradients (see Lemma 19). This method gives us the bounds on the excess risk of bias part. Also, note that our approach does not require the assumption of bounded uniform kurtosis and works with standard fourth moment Assumption 1.

C.2. Bias

Recalling the recursion Eq.(12), we have

$$\boldsymbol{\theta}_{b,t+1} = \mathcal{J}_t\,\boldsymbol{\theta}_{b,t}. \tag{36}$$

For all $t\geq 0$, we use the following notation for the sake of brevity

$$\mathcal{C}_t \stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\theta}_{b,t}\otimes\boldsymbol{\theta}_{b,t}\right].$$

Lemma 18 *For $t\geq 0$, the covariance of the bias iterates is determined by the recursion,*

$$\mathcal{C}_{t+1} = \widetilde{\mathcal{T}}^{(t+1)}\circ\mathcal{C}_0 + \sum_{k=0}^{t}\widetilde{\mathcal{T}}^{k}\circ\mathcal{M}\circ\mathcal{C}_{t-k}$$

Proof From the recursion on the bias covariance of Lemma 14, we have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\theta}_{b,t+1}\otimes\boldsymbol{\theta}_{b,t+1}\right] &= \mathcal{T}\circ\mathbb{E}\left[\boldsymbol{\theta}_{b,t}\otimes\boldsymbol{\theta}_{b,t}\right],\\
\mathcal{C}_{t+1} &= \mathcal{T}\circ\mathcal{C}_t = \left(\widetilde{\mathcal{T}}+\mathcal{M}\right)\circ\mathcal{C}_t, \quad \text{from Lemma 10,}\\
&= \widetilde{\mathcal{T}}\circ\mathcal{C}_t+\mathcal{M}\circ\mathcal{C}_t,\\
&= \widetilde{\mathcal{T}}\circ\left(\widetilde{\mathcal{T}}\circ\mathcal{C}_{t-1}+\mathcal{M}\circ\mathcal{C}_{t-1}\right)+\mathcal{M}\circ\mathcal{C}_t = \widetilde{\mathcal{T}}^2\circ\mathcal{C}_{t-1}+\widetilde{\mathcal{T}}\circ\mathcal{M}\circ\mathcal{C}_{t-1}+\mathcal{M}\circ\mathcal{C}_t.
\end{aligned}$$

Expanding this recursively we get the following expression

$$\mathcal{C}_{t+1} = \widetilde{\mathcal{T}}^{(t+1)} \circ \mathcal{C}_0 + \sum_{k=0}^{t} \widetilde{\mathcal{T}}^{k} \circ \mathcal{M} \circ \mathcal{C}_{t-k}.$$

■

Lemma 19 *For $T \geq 0$,*

$$\left(\mathcal{I} - \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right) \circ \sum_{t=0}^{T} \mathcal{C}_t \preccurlyeq \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{(t)}\right) \circ \mathcal{C}_0. \tag{37}$$

Proof From Lemma 18,

$$\mathcal{C}_t = \widetilde{\mathcal{T}}^{(t)} \circ \mathcal{C}_0 + \sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{t-k-1} \circ \mathcal{M} \circ \mathcal{C}_k.$$

Consider the following summation,

$$\begin{aligned}
\sum_{t=0}^{T} \mathcal{C}_t &= \sum_{t} \left[\widetilde{\mathcal{T}}^{(t)} \circ \mathcal{C}_0 + \sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{t-1-k} \circ \mathcal{M} \circ \mathcal{C}_k\right], \\
&= \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{(t)}\right) \circ \mathcal{C}_0 + \sum_{t=0}^{T} \sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{t-1-k} \circ \mathcal{M} \circ \mathcal{C}_k.
\end{aligned}$$

Exchanging the summations for the second part,

$$\begin{aligned}
\sum_{t=0}^{T} \mathcal{C}_t &= \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{(t)}\right) \circ \mathcal{C}_0 + \sum_{k=0}^{T-1} \sum_{t=k+1}^{T} \widetilde{\mathcal{T}}^{t-1-k} \circ \mathcal{M} \circ \mathcal{C}_k, \\
&= \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{(t)}\right) \circ \mathcal{C}_0 + \sum_{k=0}^{T-1} \left[\sum_{t=0}^{T-k-1} \widetilde{\mathcal{T}}^{t}\right] \circ \mathcal{M} \circ \mathcal{C}_k.
\end{aligned}$$

Note that $\mathcal{C}_k$ is PSD, for $k \geq 0$ and $\mathcal{M}$ is positive. Hence, $\mathcal{M} \circ \mathcal{C}_k$ is PSD. Since $\widetilde{\mathcal{T}} \succcurlyeq 0$ (to be precise $\widetilde{\mathcal{T}} \circ \mathcal{M} \circ \mathcal{C}_k \succcurlyeq 0$, but we drop this for simplicity of writing), we can say the following things,

$$\begin{aligned}
\forall t \geq 0,\ \ \widetilde{\mathcal{T}}^{\,t} \succcurlyeq 0, \\
\text{for any } t' \geq 0,\ \ \sum_{t > t'} \widetilde{\mathcal{T}}^{\,t} \succcurlyeq 0, \\
\text{Hence, for any } t' \geq 0,\ \ \sum_{t \geq 0} \widetilde{\mathcal{T}}^{\,t} \succcurlyeq \sum_{t=0}^{t'} \widetilde{\mathcal{T}}^{\,t}.
\end{aligned}$$

$$\sum_{t=0}^{T} \mathcal{C}_t \preccurlyeq \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{(t)}\right) \circ \mathcal{C}_0 + \sum_{k=0}^{T-1} \left[\sum_{t \geq 0} \widetilde{\mathcal{T}}^{\,t}\right] \circ \mathcal{M} \circ \mathcal{C}_k.$$

Here we use the fact that
$$\sum_{t\geq 0} \widetilde{\mathcal{T}}^{\,t} = \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1}.$$

$$\begin{aligned}\sum_{t=0}^{T} \mathcal{C}_t &\preccurlyeq \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{\,(t)}\right) \circ \mathcal{C}_0 + \sum_{k=0}^{T-1} \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M} \circ \mathcal{C}_k, \\ &\preccurlyeq \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{\,(t)}\right) \circ \mathcal{C}_0 + \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M} \circ \sum_{k=0}^{T-1} \mathcal{C}_k.\end{aligned}$$

Using the fact that $\mathcal{C}_T \succcurlyeq 0$, we have
$$\sum_{k=0}^{T-1} \mathcal{C}_k \leq \sum_{k=0}^{T} \mathcal{C}_k$$

Hence,

$$\sum_{t=0}^{T} \mathcal{C}_t \preccurlyeq \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{\,(t)}\right) \circ \mathcal{C}_0 + \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M} \circ \sum_{k=0}^{T} \mathcal{C}_k.$$

From this we can prove the lemma

$$\left(\mathcal{I} - \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right) \circ \sum_{t=0}^{T} \mathcal{C}_t \preccurlyeq \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{\,(t)}\right) \circ \mathcal{C}_0.$$

■

Lemma 20 *With the stepsizes satisfying $(\alpha + 2\beta)R^2 \leq 1, \alpha \leq \frac{\beta}{2\tilde{\kappa}}$ the sum of covariance can be bounded by*

$$\sum_{t=0}^{T} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right] \right\rangle \leq \min\left\{\frac{3(T+1)}{\alpha}, \frac{12(T+1)(T+2)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.$$

Proof From Lemma 19,

$$\left(\mathcal{I} - \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right) \circ \sum_{t=0}^{T} \mathcal{C}_t \preccurlyeq \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{\,(t)}\right) \circ \mathcal{C}_0,$$
$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \left(\mathcal{I} - \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right) \circ \sum_{t=0}^{T} \mathcal{C}_t \right\rangle \leq \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{\,(t)}\right) \circ \mathcal{C}_0 \right\rangle.$$

Using the definition of transpose of operators in Remark 11, we get

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \left(\mathcal{I} - \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right) \circ \sum_{t=0}^{T} \mathcal{C}_t \right\rangle = \left\langle \left(\mathcal{I} - \mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1}\right) \circ \Upsilon, \sum_{t=0}^{T} \mathcal{C}_t \right\rangle \tag{38}$$

As the condition on the stepsize is satisfied, we can use the fact that Υ is almost eigen vector of $\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1}$ from Lemmas 15, 34,

$$\begin{aligned} \mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1} \circ \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix} &\preccurlyeq \frac{2}{3} \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix} \\ \left(\mathcal{I} - \mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1}\right) \circ \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix} &\succcurlyeq \frac{1}{3} \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}. \end{aligned}$$

Combining them we get,

$$\begin{aligned} \frac{1}{3} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \sum_{t=0}^{T} \mathcal{C}_t \right\rangle &\leq \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \left(\sum_{t=0}^{T} \widetilde{\mathcal{T}}^{(t)}\right) \circ \mathcal{C}_0 \right\rangle, \\ \sum_{t=0}^{T} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathcal{C}_t \right\rangle &\leq 3 \sum_{t=0}^{T} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \left(\widetilde{\mathcal{T}}^{(t)}\right) \circ \mathcal{C}_0 \right\rangle. \end{aligned}$$

Note that $\boldsymbol{\theta}_{b,0} = \boldsymbol{\theta}_0$ gives $\mathcal{C}_0 = \boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0$. From Lemma 27, for $0 \leqslant t \leqslant T$,

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \widetilde{\mathcal{T}}^{(t)} \circ \, [\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0] \right\rangle \leq \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2.$$

Summing this for $0 \leq t \leq T$ proves the lemma. ∎

C.3. Bias Last Iterate

Lemma 21 (Final Iterate Risk) *Under Assumptions 4 and the step-sizes satisfying $\alpha \leq \beta \leq 1/L$. For $T \geq 0$, the last iterate excess error can be determined by the following discrete Volterra integral*

$$\mathsf{f}\left(\boldsymbol{\theta}_{b,T}\right) \leqslant \min\left\{\frac{1}{\alpha}, \frac{8(T+1)}{\beta}\right\} \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2 + \sum_{t=0}^{T-1} \mathsf{g}\left(\mathbf{H}, t-k-1\right) \mathsf{f}\left(\boldsymbol{\theta}_{b,t}\right),$$

where $\mathsf{f}\left(\boldsymbol{\theta}_{b,t}\right) \stackrel{\text{def}}{=} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right] \right\rangle$ *and the kernel*

$$\mathsf{g}\left(\mathbf{H}, t\right) \stackrel{\text{def}}{=} \kappa \left\langle \Upsilon, \widetilde{\mathcal{T}}^t \circ \boldsymbol{\Xi} \right\rangle. \tag{39}$$

Proof Invoking Lemma 18 and $\boldsymbol{\theta}_{b,0} = \boldsymbol{\theta}_0$ gives

$$\mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right] = \widetilde{\mathcal{T}}^{(t)} \circ \left[\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0\right] + \sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{t-k-1} \circ \mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right].$$

Using Lemma 32, note Υ is defined at Eq.(31)

$$\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right] \preccurlyeq \kappa \left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right]\right\rangle \boldsymbol{\Xi}.$$

Using this and fact that $\mathcal{T}$ is positive and $\boldsymbol{\Xi}$ is PSD,

$$\mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right] \preccurlyeq \widetilde{\mathcal{T}}^{(t)} \circ \left[\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0\right] + \sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{t-k-1} \circ \kappa \left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right]\right\rangle \boldsymbol{\Xi}.$$

Taking the scalar product with Υ on both sides, gives us

$$\left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\right\rangle \leq \left\langle \Upsilon, \widetilde{\mathcal{T}}^{(t)} \circ \left[\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0\right]\right\rangle + \sum_{k=0}^{t-1} \left\langle \Upsilon, \widetilde{\mathcal{T}}^{t-k-1} \circ \boldsymbol{\Xi}\right\rangle \kappa \left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right]\right\rangle.$$

From Lemma 27, we get,

$$\left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\right\rangle \leq \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \sum_{k=0}^{t-1} \left\langle \Upsilon, \widetilde{\mathcal{T}}^{t-k-1} \circ \boldsymbol{\Xi}\right\rangle \kappa \left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right]\right\rangle.$$

The definition of $\mathsf{f}\left(\boldsymbol{\theta}_{b,t}\right)$ proves the lemma. ■

Lemma 22 *With* $(\alpha + 2\beta) \leq \frac{1}{\kappa \operatorname{Tr}\mathbf{H}}, \alpha \leq \frac{\beta}{2\kappa d}$*, after* t *iterations of Algorithm 4 the bias excess error,*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right\rangle \leqslant \min\left\{\frac{3}{\alpha}, \frac{24(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.$$

Proof From Lemma 21,

$$\left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\right\rangle \leq \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \sum_{k=0}^{t-1} \left\langle \Upsilon, \widetilde{\mathcal{T}}^{t-k-1} \circ \boldsymbol{\Xi}\right\rangle \kappa \left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right]\right\rangle.$$

Now we will use induction to show that $\left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right]\right\rangle$ is bounded.

Induction Hypothesis There exists a constant C, for all $0 \leqslant k \leqslant t-1$, $\left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,k} \otimes \boldsymbol{\theta}_{b,k}\right]\right\rangle \leq C$. Using this,

$$\begin{aligned}
\left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\right\rangle &\leq \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \sum_{k=0}^{t-1} \left\langle \Upsilon, \widetilde{\mathcal{T}}^{t-k-1} \circ \boldsymbol{\Xi}\right\rangle \kappa C, \\
&= \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \left\langle \Upsilon, \sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{t-k-1} \circ \boldsymbol{\Xi}\right\rangle \kappa C.
\end{aligned}$$

As $\mathcal{T}$ is positive and $\Upsilon, \boldsymbol{\Xi}$ is PSD,

$$\left(\sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{k}\right) \circ \boldsymbol{\Xi} \preccurlyeq \left(\sum_{k=0}^{\infty} \widetilde{\mathcal{T}}^{k}\right) \circ \boldsymbol{\Xi} = \left(1 - \widetilde{\mathcal{T}}\right)^{-1} \circ \boldsymbol{\Xi}$$

Using this upperbound,

$$\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\rangle \leqslant \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \kappa C \left\langle \Upsilon, \left(1 - \widetilde{\mathcal{T}}\right)^{-1} \circ \boldsymbol{\Xi} \right\rangle. \tag{40}$$

As the step sizes are chosen accordingly, using Lemma 29,

$$\kappa C \left\langle \Upsilon, \left(1 - \widetilde{\mathcal{T}}\right)^{-1} \circ \boldsymbol{\Xi} \right\rangle \leq \frac{2C}{3}.$$

Substituting these back in Eq.(40),

$$\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\rangle \leqslant \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 + \frac{2C}{3}.$$

If we choose C such that,

$$\begin{aligned} C &= \min\left\{\frac{3}{\alpha}, \frac{24(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2, \\ \frac{2C}{3} &= \min\left\{\frac{2}{\alpha}, \frac{16(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2, \\ \langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\rangle &\leqslant \min\left\{\frac{3}{\alpha}, \frac{24(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 = C. \end{aligned}$$

Hence we have shown that $\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\rangle \leq C$. From induction we can say that for all $T > 0$,

$$\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_{b,t} \otimes \boldsymbol{\theta}_{b,t}\right]\rangle \leqslant C \quad \text{where} \quad C = \min\left\{\frac{3}{\alpha}, \frac{24(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2 \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.$$

■

C.4. Variance

We start by extending the the definition of the random matrix $\mathcal{J}_t$,

Definition 23 *For every $0 \leq i \leq j$, define the random linear operator $\mathcal{J}(i, j)$ as follows*

$$\mathcal{J}(j, i) = \prod_{k=i}^{j-1} \mathcal{J}_k \quad \textit{and} \quad \mathcal{J}(i, i) = \mathcal{I}. \tag{41}$$

Recalling the variance subproblem Eq.(13) and using the above definition,

$$\boldsymbol{\theta}_{v,0} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \qquad \boldsymbol{\theta}_{v,t} = \mathcal{J}(t, t-1)\boldsymbol{\theta}_{v,t-1} + \boldsymbol{\epsilon_t}. \tag{42}$$

Using this recursion for $t - 1$ and expanding it, we will get the following

$$\boldsymbol{\theta}_{v,t} = \mathcal{J}(t, t-1)\left(\mathcal{J}(t-1, t-2)\boldsymbol{\theta}_{v,t-2} + \boldsymbol{\epsilon_{t-1}}\right) + \boldsymbol{\epsilon_t}.$$

Using the definition of $\mathcal{J}(i,j)$ Def. 23,

$$\boldsymbol{\theta}_{v,t} = \mathcal{J}(t,t-2)\boldsymbol{\theta}_{v,t-2} + \mathcal{J}(t,t-1)\boldsymbol{\epsilon_{t-1}} + \boldsymbol{\epsilon_t}.$$

Expanding it further for any $0 \le i \le t$, we have the following expression

$$\boldsymbol{\theta}_{v,t} = \mathcal{J}(t,i)\boldsymbol{\theta}_{v,i} + \sum_{k=i+1}^{t} \mathcal{J}(t,k)\boldsymbol{\epsilon_k}. \tag{43}$$

Lemma 24 *With the recursion defined by Eq.(42) and the expected covariance of the* $\overline{\boldsymbol{\theta}}_{v,t}$

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] = \sum_{i=1}^{T} \left(\sum_{j\ge i}^{T} \mathcal{A}^{j-i}\right) \{\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right]\} \left(\sum_{j\ge i}^{T} \mathcal{A}^{j-i}\right)^{\top}.$$

Proof Recall that

$$\overline{\boldsymbol{\theta}}_{v,T} = \sum_{t=0}^{T} \boldsymbol{\theta}_{v,t}.$$

Considering the covariance of $\overline{\boldsymbol{\theta}}_{v,T}$,

$$\begin{aligned}
\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T} &= \left(\sum_{i=0}^{T} \boldsymbol{\theta}_{v,i}\right) \otimes \left(\sum_{j=0}^{T} \boldsymbol{\theta}_{v,j}\right), \\
&= \sum_{i} \left(\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i} + \sum_{j>i} \left(\boldsymbol{\theta}_{v,j} \otimes \boldsymbol{\theta}_{v,i} + \boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,j}\right)\right).
\end{aligned}$$

Taking expectation and using the linearity of expectation we get,

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] = \sum_{i=1}^{T} \left[\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] + \sum_{j>i}^{T} \left(\mathbb{E}\left[\boldsymbol{\theta}_{v,j} \otimes \boldsymbol{\theta}_{v,i}\right] + \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,j}\right]\right)\right].$$

Note that from Eq.(43), we can write $\boldsymbol{\theta}_{v,j} \otimes \boldsymbol{\theta}_{v,i}$ for $j > i$ as follows

$$\boldsymbol{\theta}_{v,j} \otimes \boldsymbol{\theta}_{v,i} = \mathcal{J}(t,i)\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i} + \sum_{k=i+1}^{j} \mathcal{J}(t,k)\boldsymbol{\epsilon_k} \otimes \boldsymbol{\theta}_{v,i}.$$

Now taking the expectation,

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,j} \otimes \boldsymbol{\theta}_{v,i}\right] = \mathbb{E}\left[\mathcal{J}(t,i)\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] + \sum_{k=i+1}^{j} \mathbb{E}\left[\mathcal{J}(t,k)\boldsymbol{\epsilon_k} \otimes \boldsymbol{\theta}_{v,i}\right].$$

For all $k > i$, $\mathcal{J}(t,k)$, $\boldsymbol{\epsilon_k}$ is independent of $\boldsymbol{\theta}_{v,i}$, $\mathcal{J}(t,k)$, $\boldsymbol{\epsilon_k}$ are also independent from their definition. We have $\mathbb{E}\left[\mathcal{J}(t,k)\right] = \mathcal{A}^{t-k}$, and $\mathbb{E}\left[\boldsymbol{\epsilon_k}\right] = 0$. Using these,

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,j} \otimes \boldsymbol{\theta}_{v,i}\right] = \mathcal{A}^{j-i}\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right].$$

With the same reasoning, for $j \geq i$,

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,j}\right] = \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\mathcal{A}^\top\right)^{j-i}.$$

Substituting the above here gives,

$$\begin{aligned}
\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] &= \sum_{i=1}^{T} \left\{ \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] + \sum_{j>i}^{T} \left(\mathbb{E}\left[\boldsymbol{\theta}_{v,j} \otimes \boldsymbol{\theta}_{v,i}\right] + \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,j}\right]\right) \right\}, \\
&= \sum_{i=1}^{T} \left\{ \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] + \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right) \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] + \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right)^\top \right\}.
\end{aligned}$$

Note that from here, a upper bound on $\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right]$ doesnot translate to an upperbound on the $\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right]$ as the matrix $\mathcal{A}$ is not positive unlike the case of SGD. Using the following identity for any two matrices S and a vector ϕ,

$$\phi \otimes \phi + S \cdot \phi \otimes \phi + \phi \otimes \phi \cdot S^\top = (I + S) \cdot \phi \otimes \phi \cdot (I + S)^\top - S \cdot \phi \otimes \phi \cdot S^\top.$$

$$\begin{aligned}
\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] = \sum_{i=1}^{T} \left(\mathcal{I} + \sum_{j>i}^{T} \mathcal{A}^{j-i}\right) \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\mathcal{I} + \sum_{j>i}^{T} \mathcal{A}^{j-i}\right)^\top \\
- \sum_{i=1}^{T} \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right) \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right)^\top.
\end{aligned}$$

Now the first term can be written as follows,

$$\sum_{i=1}^{T} \left(\mathcal{I} + \sum_{j>i}^{T} \mathcal{A}^{j-i}\right) \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\mathcal{I} + \sum_{j>i}^{T} \mathcal{A}^{j-i}\right)^\top = \sum_{i=1}^{T} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right)^\top.$$

For the second term note that at $i = T$ the summation will be 0. So we directly consider the summation till $T - 1$.

$$\begin{aligned}
\sum_{i=1}^{T} \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right) \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right)^\top &= \sum_{i=1}^{T-1} \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right) \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right)^\top, \\
&= \sum_{i=1}^{T-1} \left(\sum_{j\geq i+1}^{T} \mathcal{A}^{j-i-1}\right) \mathcal{A}\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \mathcal{A}^\top \left(\sum_{j\geq i+1}^{T} \mathcal{A}^{j-i-1}\right)^\top.
\end{aligned}$$

By definition of $\widetilde{\mathcal{T}}$ and change of variable '$i + 1 \to i$' gives

$$\sum_{i=1}^{T} \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right) \mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] \left(\sum_{j>i}^{T} \mathcal{A}^{j-i}\right)^\top = \sum_{i=2}^{T} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \widetilde{\mathcal{T}} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right)^\top.$$

Combining both parts and noting that $\boldsymbol{\theta}_{v,0} = 0$ we get,

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] = \sum_{i=1}^{T} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \left\{\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] - \widetilde{\mathcal{T}} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right]\right\} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right)^{\top}.$$

From Lemma 14,

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] &= \mathcal{T} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right], \\
\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] &= \left(\widetilde{\mathcal{T}} + \mathcal{M}\right) \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right], \\
\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] - \widetilde{\mathcal{T}} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] &= \mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right].
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] &= \sum_{i=1}^{T} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \left\{\mathbb{E}\left[\boldsymbol{\theta}_{v,i} \otimes \boldsymbol{\theta}_{v,i}\right] - \widetilde{\mathcal{T}} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right]\right\} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right)^{\top}, \\
&= \sum_{i=1}^{T} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \left\{\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right]\right\} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right)^{\top}.
\end{aligned}$$

This proves the lemma. ■

Lemma 25 *With the recursion defined by Eq.(42) and step sizes satisfying Condition 27, for $t \geq 0$,*

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] \preccurlyeq t^2\sigma^2 \left(\mathcal{I} - \mathcal{T}\right)^{-1} \circ \boldsymbol{\Xi}.$$

Proof From Lemma 14, we have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] &= \mathcal{T} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,t-1} \otimes \boldsymbol{\theta}_{v,t-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_t} \otimes \boldsymbol{\epsilon_t}\right], \\
&= \mathcal{T}^2 \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,t-2} \otimes \boldsymbol{\theta}_{v,t-2}\right] + \mathcal{T} \circ \mathbb{E}\left[\boldsymbol{\epsilon_{t-1}} \otimes \boldsymbol{\epsilon_{t-1}}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_t} \otimes \boldsymbol{\epsilon_t}\right], \\
&= \sum_{k=0}^{t-1} \mathcal{T}^k \circ \mathbb{E}\left[\boldsymbol{\epsilon_{t-k}} \otimes \boldsymbol{\epsilon_{t-k}}\right].
\end{aligned}$$

Recalling from the definition of $\boldsymbol{\epsilon_k}$ and its covariance,

$$\begin{aligned}
\boldsymbol{\epsilon_k} &= k\eta_k \begin{bmatrix} \beta a_t \\ \alpha a_t \end{bmatrix}, \\
\boldsymbol{\epsilon_k} \otimes \boldsymbol{\epsilon_k} &= k^2\eta_k^2 \begin{bmatrix} \beta a_t \\ \alpha a_t \end{bmatrix} \otimes \begin{bmatrix} \beta a_t \\ \alpha a_t \end{bmatrix}, \\
&= k^2 \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \left[\eta_k^2 \ a_k \otimes a_k\right].
\end{aligned}$$

where $\otimes_k$ is the kronecker product. Taking the expectation, we have

$$\mathbb{E}\left[\boldsymbol{\epsilon_k} \otimes \boldsymbol{\epsilon_k}\right] = k^2 \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \mathbb{E}\left[\eta_k^2 \ a_k \otimes a_k\right].$$

From the Assumption 2, we have

$$\mathbb{E}\left[\eta_k^2 \;\; a_k \otimes a_k\right] = \mathbb{E}\left[\left(b_k - \langle \boldsymbol{x}_*, a_k\rangle\right)^2 a_k \otimes a_k\right] \preccurlyeq \sigma^2 \mathbf{H}.$$

Using the fact that kronecker product of two PSD matrices is a PSD and recalling $\boldsymbol{\Xi}$ from Eq.(31), we get

$$\begin{aligned} \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \left(\sigma^2 \mathbf{H} - \mathbb{E}\left[\eta_k^2 \;\; a_k \otimes a_k\right]\right) &\succcurlyeq 0, \\ \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \left(\mathbb{E}\left[\eta_k^2 \;\; a_k \otimes a_k\right]\right) &\preccurlyeq \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \left(\sigma^2 \mathbf{H}\right), \\ &\preccurlyeq \sigma^2 \begin{bmatrix} \beta^2 \mathbf{H} & \alpha\beta \mathbf{H} \\ \alpha\beta \mathbf{H} & \alpha^2 \mathbf{H} \end{bmatrix} = \sigma^2 \boldsymbol{\Xi}. \end{aligned}$$

Combining these we get the following,

$$\mathbb{E}\left[\boldsymbol{\epsilon_k} \otimes \boldsymbol{\epsilon_k}\right] \preccurlyeq \sigma^2 k^2 \cdot \boldsymbol{\Xi}.$$

Using this upper bound in the expansion of $\mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right]$,

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] &= \sum_{k=0}^{t-1} \mathcal{T}^k \circ \mathbb{E}\left[\boldsymbol{\epsilon_{t-k}} \otimes \boldsymbol{\epsilon_{t-k}}\right], \\ &\preccurlyeq \sigma^2 \sum_{k=0}^{t-1} \mathcal{T}^k \circ (t-k)^2 \, \boldsymbol{\Xi}. \end{aligned}$$

For $0 \leqslant k \leqslant T$, we have $(t-k)^2 \leqslant t^2$ and using the fact that $\mathcal{T}, \boldsymbol{\Xi}$ are positive,

$$\begin{aligned} \sum_{k=0}^{t-1} \mathcal{T}^k \circ (t-k)^2 \, \boldsymbol{\Xi} &\preccurlyeq t^2 \sum_{k=0}^{t-1} \mathcal{T}^k \circ \boldsymbol{\Xi}, \\ &\preccurlyeq t^2 \sum_{k=0}^{\infty} \mathcal{T}^k \circ \boldsymbol{\Xi} \; = \; t^2 \left(\mathcal{I} - \mathcal{T}\right)^{-1} \circ \boldsymbol{\Xi}. \end{aligned}$$

Hence, we have

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,t} \otimes \boldsymbol{\theta}_{v,t}\right] \preccurlyeq \sigma^2 t^2 \left(\mathcal{I} - \mathcal{T}\right)^{-1} \circ \boldsymbol{\Xi}$$

This completes the proof of the lemma. ■

Lemma 26 *With α and β satisfying Condition 27, the excess error after T iterations of the variance process,*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] \right\rangle \leq 18 \left(\sigma^2 d\right) T^3$$

Proof From Lemma 24,

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] = \sum_{i=1}^{T} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \left\{\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right]\right\} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right)^{\top}.$$

First lets upperbound $\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right]$. We have the following

- Invoking Lemma 25,

$$\mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] \preccurlyeq (i-1)^2\sigma^2 \left(\mathcal{I} - \mathcal{T}\right)^{-1} \circ \boldsymbol{\Xi}.$$

- For the choice of stepsizes from Lemma 30,

$$\left(\mathcal{I} - \mathcal{T}\right)^{-1} \circ \boldsymbol{\Xi} \preccurlyeq 3\left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \boldsymbol{\Xi}.$$

- Combining these to get

$$\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] \preccurlyeq 3(i-1)^2\sigma^2\mathcal{M} \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \boldsymbol{\Xi}.$$

  the step sizes chosen allows us to invoke Lemma 33. Hence,

$$\mathcal{M} \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \boldsymbol{\Xi} \preccurlyeq \frac{2}{3}\boldsymbol{\Xi},$$
$$3\sigma^2(i-1)^2\mathcal{M} \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \boldsymbol{\Xi} \preccurlyeq 2\sigma^2(i-1)^2\boldsymbol{\Xi}.$$

- The remaining $\mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right]$ can be upperbounded by $\sigma^2 i^2\boldsymbol{\Xi}$.

Combining the above gives

$$\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right] \preccurlyeq 2\sigma^2(i-1)^2\boldsymbol{\Xi} + \sigma^2 i^2\boldsymbol{\Xi}.$$

For $0 \leq i \leq T$ this can be bounded as follows.

$$\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right] \preccurlyeq 3\sigma^2 T^2\boldsymbol{\Xi}.$$

Note that this can be used in Lemma 24 to bound $\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right]$ because for any matrix P, $P(.)P^{\top}$ is a positive operator. Hence

$$\left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \left\{\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_{v,i-1} \otimes \boldsymbol{\theta}_{v,i-1}\right] + \mathbb{E}\left[\boldsymbol{\epsilon_i} \otimes \boldsymbol{\epsilon_i}\right]\right\} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right)^{\top} \preccurlyeq 3\sigma^2 T^2 \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \cdot \boldsymbol{\Xi} \cdot \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right)^{\top}.$$

Adding this and using Lemma 24,

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] \preccurlyeq 3\sigma^2 T^2 \cdot \sum_{i=1}^{T} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right) \boldsymbol{\Xi} \left(\sum_{j\geq i}^{T} \mathcal{A}^{j-i}\right)^{\top}.$$

Using the following identity,

$$\left(\sum_{j\geq i}^{T}\mathcal{A}^{j-i}\right)=(\mathcal{I}-\mathcal{A})^{-1}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right),$$

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T}\otimes\overline{\boldsymbol{\theta}}_{v,T}\right]\preccurlyeq 3\sigma^2T^2\cdot(\mathcal{I}-\mathcal{A})^{-1}\left[\sum_{i=1}^{T}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right)\boldsymbol{\Xi}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right)^{\top}\right]\left(\mathcal{I}-\mathcal{A}^{\top}\right)^{-1}.$$

Note that we are interested in $\left\langle\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix},\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T}\otimes\overline{\boldsymbol{\theta}}_{v,T}\right]\right\rangle$.

$$\begin{aligned}
&\left\langle\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix},\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T}\otimes\overline{\boldsymbol{\theta}}_{v,T}\right]\right\rangle\\
&\qquad\leq 3\sigma^2T^2\left\langle\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix},(\mathcal{I}-\mathcal{A})^{-1}\left[\sum_{i=1}^{T}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right)\boldsymbol{\Xi}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right)^{\top}\right]\left(\mathcal{I}-\mathcal{A}^{\top}\right)^{-1}\right\rangle,\\
&\qquad= 3\sigma^2T^2\left\langle\left(\mathcal{I}-\mathcal{A}^{\top}\right)^{-1}\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix}(\mathcal{I}-\mathcal{A})^{-1},\sum_{i=1}^{T}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right)\boldsymbol{\Xi}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right)^{\top}\right\rangle.
\end{aligned}$$

Note that

$$\begin{aligned}
(\mathcal{I}-\mathcal{A})^{-1}&=\begin{bmatrix}\mathbf{I}&(\alpha\mathbf{H})^{-1}(\mathbf{I}-\beta\mathbf{H})\\-\mathbf{I}&(\alpha\mathbf{H})^{-1}(\beta\mathbf{H})\end{bmatrix},\\
\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix}(\mathcal{I}-\mathcal{A})^{-1}&=\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix}\begin{bmatrix}\mathbf{I}&(\alpha\mathbf{H})^{-1}(\mathbf{I}-\beta\mathbf{H})\\-\mathbf{I}&(\alpha\mathbf{H})^{-1}(\beta\mathbf{H})\end{bmatrix},\\
&=\begin{bmatrix}0&\alpha^{-1}\mathbf{I}\\0&\alpha^{-1}\mathbf{I}\end{bmatrix},\\
\left(\mathcal{I}-\mathcal{A}^{\top}\right)^{-1}\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix}(\mathcal{I}-\mathcal{A})^{-1}&=\begin{bmatrix}\mathbf{I}&-\mathbf{I}\\(\alpha\mathbf{H})^{-1}(\mathbf{I}-\beta\mathbf{H})&(\alpha\mathbf{H})^{-1}(\beta\mathbf{H})\end{bmatrix}\begin{bmatrix}0&\alpha^{-1}\mathbf{I}\\0&\alpha^{-1}\mathbf{I}\end{bmatrix}=\begin{bmatrix}0&0\\0&\alpha^{-2}\mathbf{H}^{-1}\end{bmatrix}.
\end{aligned}$$

Substituting this,

$$\begin{aligned}
\left\langle\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix},\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T}\otimes\overline{\boldsymbol{\theta}}_{v,T}\right]\right\rangle&\leq 3\sigma^2T^2\left\langle\begin{bmatrix}0&0\\0&\alpha^{-2}\mathbf{H}^{-1}\end{bmatrix},\sum_{i=1}^{T}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right)\boldsymbol{\Xi}\left(\mathcal{I}-\mathcal{A}^{(T-i+1)}\right)^{\top}\right\rangle,\\
&=3\sigma^2T^2\sum_{i=1}^{T}\left\langle\begin{bmatrix}0&0\\0&\alpha^{-2}\mathbf{H}^{-1}\end{bmatrix},\left(\mathcal{I}-\mathcal{A}^{(i)}\right)\boldsymbol{\Xi}\left(\mathcal{I}-\mathcal{A}^{(i)}\right)^{\top}\right\rangle.
\end{aligned}$$

From Cauchy Schwarz, we know

$$\left(\mathcal{I}-\mathcal{A}^{(i)}\right)\boldsymbol{\Xi}\left(\mathcal{I}-\mathcal{A}^{(i)}\right)^{\top}\preccurlyeq 2\boldsymbol{\Xi}+2\mathcal{A}^{(i)}\boldsymbol{\Xi}\left(\mathcal{A}^{\top}\right)^{(i)}=2\boldsymbol{\Xi}+2\widetilde{\mathcal{T}}^{i}\circ\boldsymbol{\Xi},$$

Using this,

$$\left\langle\begin{bmatrix}\mathbf{H}&\mathbf{H}\\\mathbf{H}&\mathbf{H}\end{bmatrix},\mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T}\otimes\overline{\boldsymbol{\theta}}_{v,T}\right]\right\rangle\leq 3\sigma^2T^2\sum_{i=1}^{T}\left\langle\begin{bmatrix}0&0\\0&\alpha^{-2}\mathbf{H}^{-1}\end{bmatrix},2\boldsymbol{\Xi}+2\widetilde{\mathcal{T}}^{i}\circ\boldsymbol{\Xi}\right\rangle.$$

Using Lemma 28 for the right part, we get the following,

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \mathbb{E}\left[\overline{\boldsymbol{\theta}}_{v,T} \otimes \overline{\boldsymbol{\theta}}_{v,T}\right] \right\rangle \leq 3\sigma^2 T^2 \sum_{i=1}^{T} 6d = 18\left(\sigma^2 d\right) T^3.$$

■

C.5. Potentials for Nesterov Method

In this section, we will use the potential functions used in the proof of nesterov accelerated method to bounding the terms in our recursion. Consider the Algorithm 4 with exact gradients in that setting,

$$\boldsymbol{y}'_{t+1} = \boldsymbol{x}'_t - \beta \mathbf{H}\left(\boldsymbol{x}'_t - \boldsymbol{x}_*\right), \tag{44a}$$

$$\boldsymbol{z}'_{t+1} = \boldsymbol{z}'_t - \alpha(t+1)\mathbf{H}\left(\boldsymbol{x}'_t - \boldsymbol{x}_*\right), \tag{44b}$$

$$(t+2)\boldsymbol{x}'_{t+1} = (t+1)\boldsymbol{y}'_{t+1} + \boldsymbol{z}'_{t+1}. \tag{44c}$$

By similar rescaling and the definition of the operator, it can be seen that

$$\begin{bmatrix} t(\boldsymbol{y}'_t - \boldsymbol{x}_*) \\ \boldsymbol{z}'_t - \boldsymbol{x}_* \end{bmatrix} = \mathcal{A}^t \circ \begin{bmatrix} 0 \\ \boldsymbol{x}_0 - \boldsymbol{x}_*. \end{bmatrix} \tag{45}$$

Lemma 27 *For the step sizes satisfying* $0 < \alpha \leq \beta \leq 1/L$*,*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \widetilde{\mathcal{T}}^{(t)} \circ [\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0] \right\rangle \leq \min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\} \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.$$

Proof From the above equivalence Eq.(45), we can see that

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}, \widetilde{\mathcal{T}}^{(t)} \circ [\boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}_0] \right\rangle = (t+1)^2 \left\|\boldsymbol{x}'_t - \boldsymbol{x}_0\right\|^2.$$

Now using the potential function, $V_t = (t)(t+1)\left\|\boldsymbol{y}'_t - \boldsymbol{x}_*\right\|^2_{\mathbf{H}} + \frac{1}{\alpha}\left\|\boldsymbol{z}'_t - \boldsymbol{x}_*\right\|^2$, for $\alpha \leq \beta$ we can see that $V_t \leq V_{t-1} \leq \ldots V_0$. Using this

$$t^2\left\|\boldsymbol{y}'_t - \boldsymbol{x}_*\right\|^2 \leq \frac{1}{\alpha}\left\|\boldsymbol{z}'_0 - \boldsymbol{x}_*\right\|^2 = \frac{1}{\alpha}\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2,$$

$$\left\|\boldsymbol{z}'_t - \boldsymbol{x}_*\right\|^2 \leq \left\|\boldsymbol{z}'_0 - \boldsymbol{x}_*\right\|^2 = \left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.$$

Noting that $(t+1)(\boldsymbol{x}'_t - \boldsymbol{x}_*) = t(\boldsymbol{y}'_t - \boldsymbol{x}_*) + (\boldsymbol{z}'_t - \boldsymbol{x}_*)$ and using Cauchy-Schwarz inequality,

$$(t+1)^2\left\|\boldsymbol{x}'_t - \boldsymbol{x}_0\right\|^2 \leq 2t^2\left\|\boldsymbol{y}'_t - \boldsymbol{x}_*\right\|^2 + 2\left\|\boldsymbol{z}'_t - \boldsymbol{x}_*\right\|^2 = \left(\frac{2}{\alpha} + 2\right)\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2.$$

But doing exact computations gives better bounds. The above algorithm is exactly equivalent to the algorithm considered in Flammarion and Bach (2015) as seen below

$$\eta_{t+1} = (\mathbf{I} - \alpha\mathbf{H})\eta_t + (\mathbf{I} - \beta\mathbf{H})(\eta_t - \eta_{t-1}).$$

where $\eta_t = (t+1)(\boldsymbol{x}'_t - \boldsymbol{x}_*)$. Hence we can apply their results giving the bound $\min\left\{\frac{1}{\alpha}, \frac{8(t+1)}{\beta}\right\}\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2$.
■

Lemma 28

$$\left\langle \begin{bmatrix} 0 & 0 \\ 0 & \alpha^{-2}\mathbf{H}^{-1} \end{bmatrix}, \widetilde{\mathcal{T}}^{(t)} \circ \mathbf{\Xi} \right\rangle \leq 2.$$

Proof Both $\mathcal{A}$ and Ξ are diagonizable wrt to the eigen basis of $\mathbf{H}$. We will now project these block matrices onto their eigen basis and compute the summation of each component individually. Note that,

$$\mathcal{A} = \sum_{i=1}^{d} \mathcal{A}_i \otimes_k e_i e_i^\top, \qquad \mathbf{\Xi} = \sum_{i=1}^{d} \mathbf{\Xi}_i \otimes_k e_i e_i^\top. \tag{46}$$

where $\mathcal{A}_i$ and $\mathbf{\Xi}_i$ are

$$\mathcal{A}_i = \begin{bmatrix} 1-\beta\lambda_i & 1-\beta\lambda_i \\ -\alpha\lambda_i & 1-\alpha\lambda_i \end{bmatrix}, \qquad \mathbf{\Xi}_i = \begin{bmatrix} \beta^2\lambda_i & \beta\alpha\lambda_i \\ \beta\alpha\lambda_i & \alpha^2\lambda_i \end{bmatrix}. \tag{47}$$

Now, the scalar product using the properties of Kronecker product,

$$\left\langle \begin{bmatrix} 0 & 0 \\ 0 & \alpha^{-2}\mathbf{H}^{-1} \end{bmatrix}, \widetilde{\mathcal{T}}^{(t)} \circ \mathbf{\Xi} \right\rangle = \sum_{i=1}^{d} \left\langle \begin{bmatrix} 0 & 0 \\ 0 & \alpha^{-2}\lambda_i^{-1} \end{bmatrix}, \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top \right\rangle.$$

To compute $\left\langle \begin{bmatrix} 0 & 0 \\ 0 & \alpha^{-2}\lambda_i^{-1} \end{bmatrix}, \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top \right\rangle$ we invoke Lemma 67, with

$$\begin{aligned} \Gamma &= \mathcal{A}_i, & \aleph &= \lambda_i \mathbf{\Xi}_i, \\ b &= \beta\lambda_i, & a &= \alpha\lambda_i. \end{aligned}$$

which gives,

$$\left\langle \begin{bmatrix} 0 & 0 \\ 0 & \alpha^{-2}\lambda_i^{-1} \end{bmatrix}, \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top \right\rangle \leq 1 + \frac{\alpha\lambda_i}{(1-\beta\lambda_i)^2}.$$

Note from condition on step sizes Eq.(27) that $\alpha \leq \beta/2, \beta\lambda_i \leq \beta L \leq \beta R^2 \leq 1/2$,

$$\left\langle \begin{bmatrix} 0 & 0 \\ 0 & \alpha^{-2}\lambda_i^{-1} \end{bmatrix}, \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top \right\rangle \leq 1 + \frac{\alpha\lambda_i}{(1-\beta\lambda_i)^2} \leq 2.$$

Computing the sum across dimension, we get the desired bound. ■

Appendix D. Inverting operators

In this section, we give proof for the almost eigenvalues of the operators $\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1}$, $\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1}$. As described earlier, although the calculations are a bit extensive, the underlying scheme remains the same. To compute $(1 - \widetilde{\mathcal{T}})^{-1}$, $\left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1}$, we formulate inverse as a summation of geometric series. Then we use the diagonalization of the $\mathbf{H}$ and compute the geometric series. In the last part, we use Property 1 and Assumptions 1, 3 on the data features to get the final bounds.

Property 1 *Using* $\mathbb{E}\left[aa^\top\right] = \mathbf{H}$*, the following property holds for any PSD matrix* $(\cdot)$*,*

$$\mathbb{E}\left[\left(\mathbf{H} - aa^\top\right)(\cdot)\left(\mathbf{H} - aa^\top\right)\right] = \mathbb{E}\left[aa^\top(\cdot)aa^\top\right] - \mathbb{E}\left[\mathbf{H}(\cdot)\mathbf{H}\right] \preccurlyeq \mathbb{E}\left[aa^\top(\cdot)aa^\top\right].$$

Lemma 29 *With* $0 < \alpha, \beta < 1/L$ *and* $(\alpha + 2\beta)L < 1$*,*

$$\left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi} \preccurlyeq \frac{1}{3}\begin{bmatrix} 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} & \alpha\beta^{-1}(2\beta - \alpha)\mathbf{I} \\ \alpha\beta^{-1}(2\beta - \alpha)\mathbf{I} & 2\alpha^2\beta^{-1}\mathbf{I} \end{bmatrix}. \tag{48}$$

Proof We will compute the inverse by evaluating the summation of the following infinite series,

$$\left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi} = \sum_{t=0}^{\infty} \widetilde{\mathcal{T}}^t \circ \mathbf{\Xi} = \sum_{t=0}^{\infty} \mathcal{A}^t \cdot \mathbf{\Xi} \cdot (\mathcal{A}^t)^\top.$$

Both $\mathcal{A}$ and $\mathbf{\Xi}$ are diagonizable wrt to the eigen basis of $\mathbf{H}$. We will now project these block matrices onto their eigen basis and compute the summation of each component individually. Note that,

$$\mathcal{A} = \sum_i \mathcal{A}_i \otimes_k e_i e_i^\top, \qquad \mathbf{\Xi} = \sum_i \mathbf{\Xi}_i \otimes_k e_i e_i^\top. \tag{49}$$

where $\mathcal{A}_i$ and $\mathbf{\Xi}_i$ are

$$\mathcal{A}_i = \begin{bmatrix} 1 - \beta\lambda_i & 1 - \beta\lambda_i \\ -\alpha\lambda_i & 1 - \alpha\lambda_i \end{bmatrix}, \qquad \mathbf{\Xi}_i = \begin{bmatrix} \beta^2\lambda_i & \beta\alpha\lambda_i \\ \beta\alpha\lambda_i & \alpha^2\lambda_i \end{bmatrix}. \tag{50}$$

Using these projections,

$$\sum_{t=0}^{\infty} \mathcal{A}^t \cdot \mathbf{\Xi} \cdot (\mathcal{A}^t)^\top = \sum_{t=0}^{\infty}\sum_i \left(\mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top\right) \otimes_k e_i e_i^\top, \tag{51}$$

$$= \sum_i \left[\sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top\right] \otimes_k e_i e_i^\top. \tag{52}$$

We invoke Lemma 37 with

$$\Gamma = \mathcal{A}_i, \qquad \aleph = \lambda_i \mathbf{\Xi}_i,$$
$$b = \beta\lambda_i, \qquad a = \alpha\lambda_i.$$

$$\sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \lambda_i \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top = \frac{1}{\beta\lambda_i(4 - (\alpha + 2\beta)\lambda_i)}\begin{bmatrix} 2\alpha\lambda_i + \beta\lambda_i(2\beta\lambda_i - 3\alpha\lambda_i) & \alpha\lambda_i(2\beta\lambda_i - \alpha\lambda_i) \\ \alpha\lambda_i(2\beta\lambda_i - \alpha\lambda_i) & 2(\alpha\lambda_i)^2 \end{bmatrix},$$
$$\sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top = \frac{1}{(4 - (\alpha + 2\beta)\lambda_i)}\begin{bmatrix} 2\alpha(\beta\lambda_i)^{-1} + (2\beta - 3\alpha) & \alpha\beta^{-1}(2\beta - \alpha) \\ \alpha\beta^{-1}(2\beta - \alpha) & 2\alpha^2\beta^{-1} \end{bmatrix}.$$

We have

$$(\alpha + 2\beta)\lambda_i \leq (\alpha + 2\beta)L \leq 1,$$
$$\text{Hence, } 4 - ((\alpha + 2\beta)\lambda_i) \geq 3.$$

Also, $\sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top$ is PSD as $\widetilde{\mathcal{T}}, \mathbf{\Xi}$ are positive. Hence, the following holds

$$\begin{aligned} 3\sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top &\preccurlyeq 4 - ((\alpha+2\beta)\lambda_i) \sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top, \\ &= \begin{bmatrix} 2\alpha(\beta\lambda_i)^{-1} + (2\beta - 3\alpha) & \alpha\beta^{-1}(2\beta-\alpha) \\ \alpha\beta^{-1}(2\beta-\alpha) & 2\alpha^2\beta^{-1} \end{bmatrix}. \end{aligned}$$

This given the following

$$\sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top \preccurlyeq \frac{1}{3} \begin{bmatrix} 2\alpha(\beta\lambda_i)^{-1} + (2\beta - 3\alpha) & \alpha\beta^{-1}(2\beta-\alpha) \\ \alpha\beta^{-1}(2\beta-\alpha) & 2\alpha^2\beta^{-1} \end{bmatrix}. \tag{53}$$

Using the fact that kronecker product of two PSD matrices is positive,

$$\sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top \otimes_k e_i e_i^\top \preccurlyeq \frac{1}{3} \begin{bmatrix} 2\alpha(\beta\lambda_i)^{-1} + (2\beta - 3\alpha) & \alpha\beta^{-1}(2\beta-\alpha) \\ \alpha\beta^{-1}(2\beta-\alpha) & 2\alpha^2\beta^{-1} \end{bmatrix} \otimes_k e_i e_i^\top.$$

Now adding this result along all directions we get

$$\begin{aligned} \sum_i \sum_{t=0}^{\infty} \mathcal{A}_i^t \cdot \mathbf{\Xi}_i \cdot \left(\mathcal{A}_i^t\right)^\top \otimes_k e_i e_i^\top &\preccurlyeq \frac{1}{3} \sum_i \begin{bmatrix} 2\alpha(\beta\lambda_i)^{-1} + (2\beta - 3\alpha) & \alpha\beta^{-1}(2\beta-\alpha) \\ \alpha\beta^{-1}(2\beta-\alpha) & 2\alpha^2\beta^{-1} \end{bmatrix} \otimes_k e_i e_i^\top, \\ \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi} &\preccurlyeq \frac{1}{3} \begin{bmatrix} 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} & \alpha\beta^{-1}(2\beta-\alpha)\mathbf{I} \\ \alpha\beta^{-1}(2\beta-\alpha)\mathbf{I} & 2\alpha^2\beta^{-1}\mathbf{I} \end{bmatrix}, \end{aligned}$$

where the last inequation comes from the facts

$$\mathbf{I} = \sum_i e_i e_i^\top \qquad H^{-1} = \sum_i \lambda_i^{-1} e_i e_i^\top.$$

■

Lemma 30 *With* $(\alpha + 2\beta)R^2 \leq 1, \alpha \leq \frac{\beta}{2\widetilde{\kappa}}$,

$$\mathbf{\Phi}_\infty \overset{\text{def}}{=} (\mathcal{I} - \mathcal{T})^{-1} \circ \mathbf{\Xi} \preccurlyeq 3 \cdot \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi} = \begin{bmatrix} 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} & \alpha\beta^{-1}(2\beta-\alpha)\mathbf{I} \\ \alpha\beta^{-1}(2\beta-\alpha)\mathbf{I} & 2\alpha^2\beta^{-1}\mathbf{I} \end{bmatrix}. \tag{54}$$

Proof Writing the inverse as a sum of exponential series gives us

$$\mathbf{\Phi}_\infty \overset{\text{def}}{=} (\mathcal{I} - \mathcal{T})^{-1} \circ \mathbf{\Xi} = \sum_{t=0}^{\infty} \mathcal{T}^t \circ \mathbf{\Xi}.$$

Recursion for $\mathcal{T}^t \circ \mathbf{\Xi}$ will be as follows,

$$\begin{aligned}\mathcal{T}^t \circ \mathbf{\Xi} &= \widetilde{\mathcal{T}} \circ \mathcal{T}^{t-1} \circ \mathbf{\Xi} + \mathcal{M} \circ \mathcal{T}^{t-1} \circ \mathbf{\Xi},\\ &= \widetilde{\mathcal{T}}^2 \circ \mathcal{T}^{t-2} \circ \mathbf{\Xi} + \widetilde{\mathcal{T}} \circ \mathcal{M} \circ \mathcal{T}^{t-2} \circ \mathbf{\Xi} + \mathcal{M} \circ \mathcal{T}^{t-1} \circ \mathbf{\Xi},\\ &= \widetilde{\mathcal{T}}^t \circ \mathbf{\Xi} + \sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{t-k-1} \circ \mathcal{M} \circ \mathcal{T}^k \circ \mathbf{\Xi}.\end{aligned}$$

Taking the sum of these terms from 0 to ∞

$$\sum_{t=0}^{\infty} \mathcal{T}^t \circ \mathbf{\Xi} = \sum_{t=0}^{\infty} \widetilde{\mathcal{T}}^t \circ \mathbf{\Xi} + \sum_{t=0}^{\infty} \sum_{k=0}^{t-1} \widetilde{\mathcal{T}}^{t-k-1} \circ \mathcal{M} \circ \mathcal{T}^k \circ \mathbf{\Xi}.$$

Interchanging the summations in the second part,

$$\sum_{t=0}^{\infty} \mathcal{T}^t \circ \mathbf{\Xi} = \left(\sum_{t=0}^{\infty} \widetilde{\mathcal{T}}^t \right) \circ \mathbf{\Xi} + \sum_{k=0}^{\infty} \left(\sum_{t=k+1}^{\infty} \widetilde{\mathcal{T}}^{t-k-1} \right) \circ \mathcal{M} \circ \mathcal{T}^k \circ \mathbf{\Xi}$$

Using $\sum_{t=k+1}^{\infty} \widetilde{\mathcal{T}}^{t-k-1} = \sum_{t=0}^{\infty} \widetilde{\mathcal{T}}^t = (\mathcal{I} - \widetilde{\mathcal{T}})^{-1}$,

$$\begin{aligned}\sum_{t=0}^{\infty} \mathcal{T}^t \circ \mathbf{\Xi} &= \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi} + \sum_{k=0}^{\infty} \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M} \circ \mathcal{T}^k \circ \mathbf{\Xi},\\ &= \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi} + \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M} \circ \sum_{k=0}^{\infty} \mathcal{T}^k \circ \mathbf{\Xi},\\ \mathbf{\Phi}_\infty &= \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi} + \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M} \circ \mathbf{\Phi}_\infty.\end{aligned}$$

From this we have,

$$\begin{aligned}\mathbf{\Phi}_\infty - \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M} \circ \mathbf{\Phi}_\infty &= \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi},\\ \left(\mathcal{I} - \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right) \circ \mathbf{\Phi}_\infty &= \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi},\\ \mathbf{\Phi}_\infty &= \left(\mathcal{I} - \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right)^{-1} \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi}.\end{aligned}$$

Writing the inverse as a sum of exponential series gives us

$$\mathbf{\Phi}_\infty = \sum_{t=0}^{\infty} \left(\left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M} \right)^t \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}\right)^{-1} \circ \mathbf{\Xi}, \tag{55}$$

Note $(\alpha + 2\beta)R^2 \leq 1 \implies (\alpha + 2\beta)\lambda_{\max} \leq 1$. Hence we can invoke Lemma 33 here.

$$\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1} \circ \mathbf{\Xi} \preccurlyeq \left[\frac{2\alpha\tilde{\kappa}}{3\beta} + \frac{(\alpha + 2\beta)R^2}{3} \right] \mathbf{\Xi}$$

Using

$$(\alpha+2\beta)R^2 \le 1, \alpha \le \frac{\beta}{2\tilde{\kappa}},$$

$$\mathcal{M} \circ (1-\widetilde{\mathcal{T}})^{-1} \circ \boldsymbol{\Xi} \preccurlyeq \left[\frac{1}{3}+\frac{1}{3}\right] \boldsymbol{\Xi} \preccurlyeq \frac{2}{3}\boldsymbol{\Xi}.$$

Using this in Eq.(55),

$$\boldsymbol{\Phi}_\infty = \sum_{t=0}^{\infty} \left(\left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right)^t \circ \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1} \circ \boldsymbol{\Xi},$$

$$\text{Use } \left(\left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1} \circ \mathcal{M}\right)^t \circ \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1} = \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1} \circ \left(\mathcal{M} \circ \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1}\right)^t,$$

$$\boldsymbol{\Phi}_\infty = \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1} \circ \sum_{t=0}^{\infty} \left(\mathcal{M} \circ \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1}\right)^t \circ \boldsymbol{\Xi},$$

$$\text{Using } \left(\mathcal{M} \circ \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1}\right)^t \preccurlyeq \left[\frac{2}{3}\right]^t \boldsymbol{\Xi},$$

$$\boldsymbol{\Phi}_\infty \preccurlyeq \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1} \circ \sum_{t=0}^{\infty} \left[\frac{2}{3}\right]^t \cdot \boldsymbol{\Xi},$$

$$\preccurlyeq \left(\mathcal{I}-\widetilde{\mathcal{T}}\right)^{-1} \circ 3 \cdot \boldsymbol{\Xi}.$$

This completes the proof. ■

Lemma 31 *For any block matrix* $\begin{bmatrix} P & Q \\ R & S \end{bmatrix}$,

$$\mathcal{M} \circ \begin{bmatrix} P & Q \\ R & S \end{bmatrix} = \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \mathbb{E}\left[\left(\mathbf{H} - aa^\top\right)(P+Q+R+S)\left(\mathbf{H} - aa^\top\right)\right]$$

Proof

$$\mathcal{M} \circ \begin{bmatrix} P & Q \\ R & S \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} \beta\mathbf{H}_a & \beta\mathbf{H}_a \\ \alpha\mathbf{H}_a & \alpha\mathbf{H}_a \end{bmatrix}\begin{bmatrix} P & Q \\ R & S \end{bmatrix}\begin{bmatrix} \beta\mathbf{H}_a & \alpha\mathbf{H}_a \\ \beta\mathbf{H}_a & \alpha\mathbf{H}_a \end{bmatrix}\right]$$

where $\mathbf{H}_a = \left(\mathbf{H} - aa^\top\right)$

$$\begin{bmatrix} \beta\mathbf{H}_a & \beta\mathbf{H}_a \\ \alpha\mathbf{H}_a & \alpha\mathbf{H}_a \end{bmatrix}\begin{bmatrix} P & Q \\ R & S \end{bmatrix} = \begin{bmatrix} \beta\mathbf{H}_a(P+R) & \beta\mathbf{H}_a(Q+S) \\ \alpha\mathbf{H}_a(P+R) & \alpha\mathbf{H}_a(Q+S) \end{bmatrix},$$

$$\begin{bmatrix} \beta\mathbf{H}_a(P+R) & \beta\mathbf{H}_a(Q+S) \\ \alpha\mathbf{H}_a(P+R) & \alpha\mathbf{H}_a(Q+S) \end{bmatrix}\begin{bmatrix} \beta\mathbf{H}_a & \alpha\mathbf{H}_a \\ \beta\mathbf{H}_a & \alpha\mathbf{H}_a \end{bmatrix} = \begin{bmatrix} \beta^2\mathbf{H}_a(P+Q+R+S)\mathbf{H}_a & \alpha\beta\mathbf{H}_a(P+Q+R+S)\mathbf{H}_a \\ \alpha\beta\mathbf{H}_a(P+Q+R+S)\mathbf{H}_a & \alpha^2\mathbf{H}_a(P+Q+R+S)\mathbf{H}_a \end{bmatrix},$$

$$= \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \mathbf{H}_a(P+Q+R+S)\mathbf{H}_a.$$

Taking expectation,

$$\mathcal{M} \circ \begin{bmatrix} P & Q \\ R & S \end{bmatrix} = \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \mathbb{E}\left[\mathbf{H}_a (P+Q+R+S) \mathbf{H}_a\right].$$

This completes the proof. ∎

Lemma 32 *Under Assumption 4, for any $t \geq 0$,*

$$\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right] \preccurlyeq \kappa \left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right] \right\rangle \boldsymbol{\Xi}.$$

Proof

$$\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right] = \mathbb{E}\left[\mathcal{J} \mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right] \mathcal{J}^\top\right],$$

$$\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t = \begin{bmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_t \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_t \boldsymbol{v}_t^\top & \boldsymbol{v}_t \boldsymbol{w}_t^\top \\ \boldsymbol{w}_t \boldsymbol{v}_t^\top & \boldsymbol{w}_t \boldsymbol{w}_t^\top \end{bmatrix},$$

$$\mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right] = \begin{bmatrix} \mathbb{E}\left[\boldsymbol{v}_t \boldsymbol{v}_t^\top\right] & \mathbb{E}\left[\boldsymbol{v}_t \boldsymbol{w}_t^\top\right] \\ \mathbb{E}\left[\boldsymbol{w}_t \boldsymbol{v}_t^\top\right] & \mathbb{E}\left[\boldsymbol{w}_t \boldsymbol{w}_t^\top\right] \end{bmatrix}.$$

As J and $\mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right]$ are independent, invoking Lemma 31 with

$$\begin{bmatrix} P & Q \\ R & S \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[\boldsymbol{v}_t \boldsymbol{v}_t^\top\right] & \mathbb{E}\left[\boldsymbol{v}_t \boldsymbol{w}_t^\top\right] \\ \mathbb{E}\left[\boldsymbol{w}_t \boldsymbol{v}_t^\top\right] & \mathbb{E}\left[\boldsymbol{w}_t \boldsymbol{w}_t^\top\right] \end{bmatrix}.$$

Now $P+Q+R+S$ in our case will be

$$\begin{aligned} P+Q+R+S &= \mathbb{E}\left[\boldsymbol{v}_t \boldsymbol{v}_t^\top + \boldsymbol{v}_t \boldsymbol{w}_t^\top + \boldsymbol{w}_t \boldsymbol{v}_t^\top + \boldsymbol{w}_t \boldsymbol{w}_t^\top\right], \\ &= \mathbb{E}\left[\left(\boldsymbol{v}_t + \boldsymbol{w}_t\right)\left(\boldsymbol{v}_t + \boldsymbol{w}_t\right)^\top\right] = \mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\top\right], \text{ from Eq.(4c)}. \end{aligned}$$

Using this we get

$$\mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right] = \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \mathbb{E}\left[\left(\mathbf{H} - aa^\top\right) \mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\top\right] \left(\mathbf{H} - aa^\top\right)\right].$$

Using Property 1,

$$\mathbb{E}\left[\left(\mathbf{H} - aa^\top\right) \mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\top\right] \left(\mathbf{H} - aa^\top\right)\right] \preccurlyeq \mathbb{E}\left[aa^\top \mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\top\right] aa^\top\right] = \mathbb{E}\left[\left\langle a, \mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\top\right] a \right\rangle aa^\top\right].$$

Using Assumption 4 with $M = \mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\top\right]$,

$$\mathbb{E}\left[\left\langle a, \mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\top\right] a \right\rangle aa^\top\right] \preccurlyeq \kappa \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{H} \boldsymbol{u}_t \boldsymbol{u}_t^\top\right]\right) \mathbf{H}.$$

As kronecker product of two PSD matrices is positive,

$$\begin{aligned} \mathcal{M} \circ \mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right] &\preccurlyeq \begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \kappa \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{H} \boldsymbol{u}_t \boldsymbol{u}_t^\top\right]\right) \mathbf{H}, \\ &= \kappa \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{H} \boldsymbol{u}_t \boldsymbol{u}_t^\top\right]\right) \begin{bmatrix} \beta^2 \mathbf{H} & \alpha\beta \mathbf{H} \\ \alpha\beta \mathbf{H} & \alpha^2 \mathbf{H} \end{bmatrix} = \kappa \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{H} \boldsymbol{u}_t \boldsymbol{u}_t^\top\right]\right) \boldsymbol{\Xi}. \end{aligned}$$

Noting that $\left\langle \Upsilon, \mathbb{E}\left[\boldsymbol{\theta}_t \otimes \boldsymbol{\theta}_t\right] \right\rangle = \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{H} \boldsymbol{u}_t \boldsymbol{u}_t^\top\right]\right)$ completes the proof. ∎

Lemma 33 *With* $(\alpha + 2\beta)R^2 \leq 1, \alpha \leq \frac{\beta}{2\tilde{\kappa}}$,

$$\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1} \circ \boldsymbol{\Xi} \preccurlyeq \frac{2}{3}\boldsymbol{\Xi}$$

Proof Note $(\alpha + 2\beta)R^2 \leq 1 \implies (\alpha + 2\beta)\lambda_{\max} \leq 1$. Hence we can invoke Lemma 29 here.

$$\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1} \circ \boldsymbol{\Xi} \preccurlyeq \mathcal{M} \circ \frac{1}{3}\begin{bmatrix} 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} & \alpha\beta^{-1}(2\beta - \alpha)\mathbf{I} \\ \alpha\beta^{-1}(2\beta - \alpha)\mathbf{I} & 2\alpha^2\beta^{-1}\mathbf{I} \end{bmatrix}.$$

Invoking the Lemma 31 for a block matrix,

$$\begin{bmatrix} P & Q \\ R & S \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} & \alpha\beta^{-1}(2\beta - \alpha)\mathbf{I} \\ \alpha\beta^{-1}(2\beta - \alpha)\mathbf{I} & 2\alpha^2\beta^{-1}\mathbf{I} \end{bmatrix}.$$

Now $P + Q + R + S$ in our case is

$$\begin{aligned} 3 * (P + Q + R + S) &= 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} + 2\alpha\beta^{-1}(2\beta - \alpha)\mathbf{I} + 2\alpha^2\beta^{-1}\mathbf{I}, \\ &= 2\alpha(\beta\mathbf{H})^{-1} + (2\beta - 3\alpha)\mathbf{I} + 4\alpha\mathbf{I} - 2\alpha^2\beta^{-1}\mathbf{I} + +2\alpha^2\beta^{-1}\mathbf{I}, \\ &= 2\alpha(\beta\mathbf{H})^{-1} + (\alpha + 2\beta)\mathbf{I}. \end{aligned}$$

$$\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1} \circ \boldsymbol{\Xi} = \frac{1}{3}\begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \mathbb{E}\left[\left(\mathbf{H} - aa^\top\right)\left[2\alpha(\beta\mathbf{H})^{-1} + (\alpha + 2\beta)\mathbf{I}\right]\left(\mathbf{H} - aa^\top\right)\right]. \tag{56}$$

Using Property 1,

$$\begin{aligned} \mathbb{E}\left[\left(\mathbf{H} - aa^\top\right)\left[2\alpha(\beta\mathbf{H})^{-1} + (\alpha + 2\beta)\mathbf{I}\right]\left(\mathbf{H} - aa^\top\right)\right] &\preccurlyeq \mathbb{E}\left[aa^\top\left[2\alpha(\beta\mathbf{H})^{-1} + (\alpha + 2\beta)\mathbf{I}\right]aa^\top\right], \\ &= 2\frac{\alpha}{\beta}\mathbb{E}\left[\|a\|^2_{\mathbf{H}^{-1}}\, aa^\top\right] + (\alpha + 2\beta)\mathbb{E}\left[\|a\|^2\, aa^\top\right]. \end{aligned}$$

Using the Assumptions 1, 3 of the feature distribution, we have

$$\mathbb{E}\left[\left(\mathbf{H} - aa^\top\right)\left[2\alpha(\beta\mathbf{H})^{-1} + (\alpha + 2\beta)\mathbf{I}\right]\left(\mathbf{H} - aa^\top\right)\right] \preccurlyeq \frac{2\alpha\tilde{\kappa}}{\beta}\mathbf{H} + (\alpha + 2\beta)R^2\mathbf{H},$$

Using the above in Eq.(56) and that fact that kronecker product of two PSD matrices is positive we get,

$$\begin{aligned} \mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1} \circ \boldsymbol{\Xi} &\preccurlyeq \frac{1}{3}\begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \left[\frac{2\alpha\tilde{\kappa}}{\beta}\mathbf{H} + (\alpha + 2\beta)R^2\mathbf{H}\right], \\ &= \left[\frac{2\alpha\tilde{\kappa}}{3\beta} + \frac{(\alpha + 2\beta)R^2}{3}\right]\left(\begin{bmatrix} \beta^2 & \alpha\beta \\ \alpha\beta & \alpha^2 \end{bmatrix} \otimes_k \mathbf{H}\right), \\ &= \left[\frac{2\alpha\tilde{\kappa}}{3\beta} + \frac{(\alpha + 2\beta)R^2}{3}\right]\boldsymbol{\Xi}. \end{aligned}$$

where the last step is from the definition of $\boldsymbol{\Xi}$. Using

$$(\alpha + 2\beta)R^2 \leq 1, \alpha \leq \frac{\beta}{2\tilde{\kappa}},$$

$$\mathcal{M} \circ (1 - \widetilde{\mathcal{T}})^{-1} \circ \mathbf{\Xi} \preccurlyeq \left[\frac{1}{3} + \frac{1}{3}\right] \mathbf{\Xi} \preccurlyeq \frac{2}{3}\mathbf{\Xi}.$$

■

Lemma 34 *With $(\alpha + 2\beta)R^2 \leq 1, \alpha \leq \frac{\beta}{2\tilde{\kappa}}$, and Υ from Eq.(31), we have,*

$$\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1} \circ \Upsilon \preccurlyeq \frac{2}{3}\Upsilon. \tag{57}$$

Proof Compute the inverse by evaluating the summation of the following infinite series,

$$\left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1} \circ \Upsilon = \sum_{t=0}^{\infty} \left(\widetilde{\mathcal{T}}^t\right)^\top \circ \mathbf{\Xi} = \sum_{t=0}^{\infty} (\mathcal{A}^t)^\top \cdot \Upsilon \cdot \mathcal{A}^t.$$

From this it follows that,

$$\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1} = \mathbb{E}\left[\mathcal{J}^\top \cdot \left((\mathcal{A}^t)^\top \cdot \Upsilon \cdot \mathcal{A}^t\right) \cdot \mathcal{J}\right].$$

- Both $\mathcal{A}$ and Υ are diagonizable wrt to the eigen basis of $\mathbf{H}$. We will now project these block matrices onto their eigen basis and compute the summation of each component individually. Note that,

$$\mathcal{A} = \sum_i \mathcal{A}_i \otimes_k e_i e_i^\top, \qquad \Upsilon = \sum_i \Upsilon_i \otimes_k e_i e_i^\top. \tag{58}$$

  where $\mathcal{A}_i$ and Υ_i are

$$\mathcal{A}_i = \begin{bmatrix} 1 - \beta\lambda_i & 1 - \beta\lambda_i \\ -\alpha\lambda_i & 1 - \alpha\lambda_i \end{bmatrix} \qquad \Upsilon_i = \begin{bmatrix} \lambda_i & \lambda_i \\ \lambda_i & \lambda_i \end{bmatrix}.$$

  Using these projections,

$$(\mathcal{A}^t)^\top \cdot \Upsilon \cdot (\mathcal{A}^t) = \sum_i \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \otimes_k e_i e_i^\top.$$

- Now the random matrix

$$\mathcal{J} = \begin{bmatrix} \beta\left(H - aa^\top\right) & \beta\left(H - aa^\top\right) \\ \alpha\left(H - aa^\top\right) & \alpha\left(H - aa^\top\right) \end{bmatrix} = \begin{bmatrix} \beta & \beta \\ \alpha & \alpha \end{bmatrix} \otimes_k \left(H - aa^\top\right).$$

  From the mixed product property of kronecker product i.e for any matrices of appropriate dimension P, Q, R, S

$$(P \otimes_k Q)\,(R \otimes_k S) = PR \otimes_k QS.$$

$$\mathbb{E}\left[\mathcal{J}^\top \cdot \left[\left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \otimes_k e_i e_i^\top\right] \cdot \mathcal{J}\right] = \begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \begin{bmatrix} \beta & \beta \\ \alpha & \alpha \end{bmatrix} \otimes_k \mathbb{E}\left[\left(H - aa^\top\right) e_i e_i^\top \left(H - aa^\top\right)\right]. \tag{59}$$

Using the above observations,

$$\sum_{t=0}^{\infty}(\mathcal{A}^t)^\top \cdot \Upsilon \cdot (\mathcal{A}^t) = \sum_{t=0}^{\infty}\sum_i \left(\left(\mathcal{A}_i^t\right)^\top \cdot \Upsilon_i \cdot \left(\mathcal{A}_i^t\right)\right) \otimes_k e_i e_i^\top,$$

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\mathcal{J}^\top(\mathcal{A}^t)^\top \cdot \Upsilon \cdot (\mathcal{A}^t)\mathcal{J}\right] = \sum_{t=0}^{\infty}\sum_i \mathbb{E}\left[\mathcal{J}^\top \cdot \left[\left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \otimes_k e_i e_i^\top\right] \cdot \mathcal{J}\right],$$

$$= \sum_i \sum_{t=0}^{\infty} \begin{bmatrix}\beta & \alpha\\ \beta & \alpha\end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \begin{bmatrix}\beta & \beta\\ \alpha & \alpha\end{bmatrix} \otimes_k \mathbb{E}\left[\left(H - aa^\top\right) e_i e_i^\top \left(H - aa^\top\right)\right].$$

Hence,

$$\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top\right)^{-1} \circ \Upsilon = \sum_i \left(\sum_{t=0}^{\infty} \begin{bmatrix}\beta & \alpha\\ \beta & \alpha\end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \begin{bmatrix}\beta & \beta\\ \alpha & \alpha\end{bmatrix}\right) \otimes_k \mathbb{E}\left[\left(H - aa^\top\right) e_i e_i^\top \left(H - aa^\top\right)\right]. \tag{60}$$

From the definition of Υ_i,

$$\lambda_i \left(\sum_{t=0}^{\infty} \begin{bmatrix}\beta & \alpha\\ \beta & \alpha\end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \begin{bmatrix}\beta & \alpha\\ \beta & \alpha\end{bmatrix}\right) = \sum_{t=0}^{\infty} \begin{bmatrix}\beta\lambda_i & \alpha\lambda_i\\ \beta\lambda_i & \alpha\lambda_i\end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} \cdot \mathcal{A}_i^t\right) \begin{bmatrix}\beta\lambda_i & \beta\lambda_i\\ \alpha\lambda_i & \alpha\lambda_i\end{bmatrix}.$$

With $b = \beta\lambda_i, a = \alpha\lambda_i$,

$$\mathcal{A}_i = \begin{bmatrix}1-b & 1-b\\ -a & 1-a\end{bmatrix}.$$

Hence to compute this series we can invoke Lemma 38 with $\Gamma = \mathcal{A}_i$,

$$\sum_{t=0}^{\infty} \begin{bmatrix}\beta\lambda_i & \alpha\lambda_i\\ \beta\lambda_i & \alpha\lambda_i\end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} \cdot \mathcal{A}_i^t\right) \begin{bmatrix}\beta\lambda_i & \beta\lambda_i\\ \alpha\lambda_i & \alpha\lambda_i\end{bmatrix} = \left(\frac{2a}{b(4-(a+2b))} + \frac{a+2b}{(4-(a+2b))}\right)\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix},$$

$$\lambda_i \left(\sum_{t=0}^{\infty} \begin{bmatrix}\beta & \alpha\\ \beta & \alpha\end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \begin{bmatrix}\beta & \alpha\\ \beta & \alpha\end{bmatrix}\right) = \left(\frac{2\alpha\lambda_i}{\beta\lambda_i(4-(\alpha\lambda_i+2\beta\lambda_i))} + \frac{\alpha\lambda_i+2\beta\lambda_i}{(4-(\alpha\lambda_i+2\beta\lambda_i))}\right)\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix},$$

$$\left(\sum_{t=0}^{\infty} \begin{bmatrix}\beta & \alpha\\ \beta & \alpha\end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t\right) \begin{bmatrix}\beta & \alpha\\ \beta & \alpha\end{bmatrix}\right) = \left(\frac{2\alpha}{\beta\lambda_i(4-(\alpha\lambda_i+2\beta\lambda_i))} + \frac{\alpha+2\beta}{(4-(\alpha\lambda_i+2\beta\lambda_i))}\right)\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}.$$

We have

$$(\alpha+2\beta)\lambda_i \le (\alpha+2\beta)\lambda_{\max} \le (\alpha+2\beta)R^2 \le 1,$$

$$\text{Hence, } 4 - ((\alpha+2\beta)\lambda_i) \ge 3,$$

$$\frac{1}{4-((\alpha+2\beta)\lambda_i)} \le \frac{1}{3},$$

$$\frac{2\alpha}{\beta\lambda_i(4-(\alpha\lambda_i+2\beta\lambda_i))} + \frac{\alpha+2\beta}{(4-(\alpha\lambda_i+2\beta\lambda_i))} \le \frac{1}{3}\left(\frac{2\alpha}{\beta\lambda_i} + (\alpha+2\beta)\right).$$

As the matrix $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ is PSD,

$$\left(\sum_{t=0}^{\infty} \begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t \right) \begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} \right) \preccurlyeq \frac{1}{3} \left(\frac{2\alpha}{\beta\lambda_i} + (\alpha + 2\beta) \right) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Using the Property 1,

$$\mathbb{E}\left[\left(H - aa^\top \right) e_i e_i^\top \left(H - aa^\top \right) \right] \preccurlyeq \mathbb{E}\left[aa^\top \cdot e_i e_i^\top \cdot aa^\top \right].$$

Using the above two results and the fact that for any PSD matrices P, Q, R, S, $P \preccurlyeq Q$ and $R \preccurlyeq S$ then $P \otimes_k R \preccurlyeq Q \otimes_k S$. Hence from Eq.(60) we can get the bound as follows

$$\begin{aligned}
\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top \right)^{-1} \circ \Upsilon &= \sum_i \left(\sum_{t=0}^{\infty} \begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} \left((\mathcal{A}_i^t)^\top \cdot \Upsilon_i \cdot \mathcal{A}_i^t \right) \begin{bmatrix} \beta & \beta \\ \alpha & \alpha \end{bmatrix} \right) \otimes_k \mathbb{E}\left[\left(H - aa^\top \right) e_i e_i^\top \left(H - aa^\top \right) \right], \\
&\preccurlyeq \sum_i \frac{1}{3} \left(\frac{2\alpha}{\beta\lambda_i} + (\alpha + 2\beta) \right) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes_k \mathbb{E}\left[aa^\top \cdot e_i e_i^\top \cdot aa^\top \right], \\
&= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes_k \mathbb{E}\left[aa^\top \cdot \sum_i \left[\left(\frac{2\alpha}{3\beta\lambda_i} + \frac{(\alpha + 2\beta)}{3} \right) e_i e_i^\top \right] \cdot aa^\top \right], \\
&= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes_k \mathbb{E}\left[aa^\top \cdot \left[\frac{2\alpha}{3\beta} \mathbf{H}^{-1} + \frac{(\alpha + 2\beta)}{3} \mathbf{I} \right] \cdot aa^\top \right], \\
&= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes_k \left[\frac{2\alpha}{3\beta} \mathbb{E}\left[\|a\|_{\mathbf{H}^{-1}}^2 aa^\top \right] + \frac{(\alpha + 2\beta)}{3} \mathbb{E}\left[\|a\|^2 aa^\top \right] \right].
\end{aligned}$$

Using the Assumptions 1, 3 of the feature distribution, we have

$$\frac{2\alpha}{3\beta} \mathbb{E}\left[\|a\|_{\mathbf{H}^{-1}}^2 aa^\top \right] + \frac{(\alpha + 2\beta)}{3} \mathbb{E}\left[\|a\|^2 aa^\top \right] \preccurlyeq \left(\frac{2\alpha\tilde{\kappa}}{3\beta} + \frac{(\alpha + 2\beta)R^2}{3} \right) \mathbf{H}.$$

Using

$$(\alpha + 2\beta)R^2 \leq 1, \alpha \leq \frac{\beta}{2\tilde{\kappa}},$$

$$\frac{2\alpha}{3\beta} \mathbb{E}\left[\|a\|_{\mathbf{H}^{-1}}^2 aa^\top \right] + \frac{(\alpha + 2\beta)}{3} \mathbb{E}\left[\|a\|^2 aa^\top \right] \preccurlyeq \left(\frac{1}{3} + \frac{1}{3} \right) \mathbf{H} = \frac{2}{3} \mathbf{H}.$$

Using this bound, kronecker product of two PSD matrices is positive completes the proof.

$$\mathcal{M}^\top \circ \left(\mathcal{I} - \widetilde{\mathcal{T}}^\top \right)^{-1} \circ \Upsilon \preccurlyeq \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes_k \frac{2}{3} \mathbf{H} = \frac{2}{3} \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{bmatrix}.$$

■

Appendix E. Technical Lemmas

Property 2 (Eigen Decomposition of Γ) *For the matrix*

$$\Gamma = \begin{bmatrix} 1-b & 1-b \\ -a & 1-a \end{bmatrix} \tag{61}$$

The eigen values of Γ *are given by*

$$r_+ = 1 - \frac{(a+b)}{2} + \sqrt{\left(\frac{a+b}{2}\right)^2 - a} \qquad r_- = 1 - \frac{(a+b)}{2} - \sqrt{\left(\frac{a+b}{2}\right)^2 - a} \tag{62}$$

The eigen decomposition $\Gamma = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$ *where*

$$\mathbf{U} = \frac{1}{\mathbf{\Delta}} \begin{bmatrix} \frac{r_-}{(1-r_-)} & 1 \\ 1 & \frac{1-r_+}{r_+} \end{bmatrix} \qquad \mathbf{U}^{-1} = \begin{bmatrix} -a & (1-r_-)r_+ \\ (1-r_-)r_+ & -r_+r_- \end{bmatrix} \tag{63}$$

$$\mathbf{\Lambda} = \begin{bmatrix} r_+ & 0 \\ 0 & r_- \end{bmatrix} \tag{64}$$

$$\mathbf{\Delta} = r_+ - r_- \tag{65}$$

The following observations hold

- $\mathbf{U}$ *and* $\mathbf{U}^{-1}$ *are symmetric and* r_+, r_- *can be complex as* Γ *is not symmetric.*
- *For* $0 < a, b < 1$ *,* $|r_+|, |r_-| < 1$

Lemma 35 *For* r_+, r_- *given in the Property 2, the following bound holds*

$$\left[\frac{(1-r_+)\, r_+^t - (1-r_-)\, r_-^t}{\mathbf{\Delta}}\right]^2 \le 1 + \frac{a}{(1-b)^2}. \tag{66}$$

Proof To prove this conside the one-dimensional nesterov sequences starting form $x_0 = 1, z_0 = 0$

$$y_{t+1} = x_t - bx_t, \tag{67a}$$

$$z_{t+1} = \boldsymbol{z}_t - a(t+1)x_t, \tag{67b}$$

$$(t+2)x_{t+1} = (t+1)y_{t+1} + z_{t+1} \tag{67c}$$

We can easily check that, for $t \ge 0$,

$$\begin{bmatrix} (t+1)y_{t+1} \\ z_{t+1} \end{bmatrix} = \Gamma^{t+1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{68}$$

Using the eigen decomposition of $\Gamma = U\Lambda U^{-1}$ we can check that

$$U^{-1}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -a & (1-r_-)r_+ \\ (1-r_-)r_+ & -r_+r_- \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -a \\ (1-r_-)r_+ \end{bmatrix},$$

$$U\Lambda^{t+1}U^{-1}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\mathbf{\Delta}}\begin{bmatrix} \frac{r_-}{(1-r_-)} & 1 \\ 1 & \frac{1-r_+}{r_+} \end{bmatrix}\begin{bmatrix} -ar_+^{t+1} \\ (1-r_-)r_+r_-^{t+1} \end{bmatrix} = \frac{1}{\mathbf{\Delta}}\begin{bmatrix} -r_+r_- \left[(1-r_+)r_+^t - (1-r_-)r_-^t\right] \\ -a\left[r_+^{t+1} - r_-^{t+1}\right] \end{bmatrix},$$

$$\Gamma^{t+1}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\mathbf{\Delta}}\begin{bmatrix} -r_+r_- \left[(1-r_+)r_+^t - (1-r_-)r_-^t\right] \\ -a\left[r_+^{t+1} - r_-^{t+1}\right]. \end{bmatrix}$$

Hence,

$$\begin{bmatrix}(t+1)y_{t+1}\\ z_{t+1}\end{bmatrix} = \frac{1}{\Delta}\begin{bmatrix}-r_+r_-\left[(1-r_+)r_+^t-(1-r_-)r_-^t\right]\\ -a\left[r_+^{t+1}-r_-^{t+1}\right].\end{bmatrix} \tag{69}$$

Now use the potential function defined by $V_t = t^2y_t^2 + \frac{1}{a}z_t^2$, for $t \geq 1$. If $0 < a \leq b < 1$ then for $t \geq 1$ we can show that $V_{t+1} \leq V_t$. Hence, for any $t \geq 1$, $V_t \leq V_1$. Note that $V_1 < V_0$ does not hold due to different initialization. From this,

$$(t+1)^2y_{t+1}^2 \leq (t+1)^2y_{t+1}^2 + \frac{1}{a}z_{t+1}^2 \leq V_1, \tag{70}$$

$$V_1 = (1-b)^2 + \frac{1}{a}a^2 = (1-b)^2 + a. \tag{71}$$

Using the expression of $(t+1)^2y_{t+1}^2$, we get

$$\begin{aligned}\left[\frac{(1-r_+)\,r_+^t-(1-r_-)\,r_-^t}{\Delta}\right]^2 &\leq \frac{1}{(r_+r_-)^2}\left[(1-b)^2+a\right],\\ &= 1+\frac{a}{(1-b)^2}.\end{aligned}$$

This proves the lemma. ∎

Lemma 36 *For* $0 < a \leq b < 1$, *for* $\Gamma, \aleph$ *in Lemma 37,*

$$\left\langle \begin{bmatrix}0 & 0\\ 0 & \frac{1}{a^2}\end{bmatrix}, \Gamma^T\aleph\left(\Gamma^T\right)^\top \right\rangle \leq 1+\frac{a}{(1-b)^2}. \tag{72}$$

Proof In the following Lemma 37, we compute the closed form for $\Gamma^T\aleph\left(\Gamma^T\right)^\top = \begin{bmatrix}\nu_{11}(t) & \nu_{12}(t)\\ \nu_{21}(t) & \nu_{22}(t)\end{bmatrix}$. Using this

$$\left\langle \begin{bmatrix}0 & 0\\ 0 & \frac{1}{a^2}\end{bmatrix}, \Gamma^T\aleph\left(\Gamma^T\right)^\top \right\rangle = \frac{\nu_{22}(t)}{a^2}.$$

From Eq.(76),

$$\left\langle \begin{bmatrix}0 & 0\\ 0 & \frac{1}{a^2}\end{bmatrix}, \Gamma^T\aleph\left(\Gamma^T\right)^\top \right\rangle = \left[\frac{(1-r_+)\,r_+^t-(1-r_-)\,r_-^t}{\Delta}\right]^2. \tag{73}$$

From Lemma 35, the lemma holds. ∎

Lemma 37 *For* $0 < a, b < 1$, *with* Γ *and* $\aleph$ *of form*

$$\Gamma = \begin{bmatrix}1-b & 1-b\\ -a & 1-a\end{bmatrix}, \qquad \aleph = \begin{bmatrix}b^2 & ba\\ ba & a^2\end{bmatrix}.$$

The series

$$\sum_{t=0}^{\infty}\Gamma^t\aleph\left(\Gamma^t\right)^\top = \frac{1}{b(4-(a+2b))}\begin{bmatrix}2a+b(2b-3a) & a(2b-a)\\ a(2b-a) & 2a^2\end{bmatrix}. \tag{74}$$

Proof To calculate the exponents of Γ we use the eigen decomposition in Property 2,

$$\Gamma = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1},$$
$$\Gamma^t = \mathbf{U}\mathbf{\Lambda}^t\mathbf{U}^{-1},$$
$$\Gamma^t \cdot \aleph \cdot \left(\Gamma^t\right)^\top = \mathbf{U}\mathbf{\Lambda}^t\mathbf{U}^{-1} \cdot \aleph \cdot \left(\mathbf{U}\mathbf{\Lambda}^t\mathbf{U}^{-1}\right)^\top,$$

From Property 2 that $\mathbf{U}, \mathbf{U}^{-1}$ are symmetric.

$$\Gamma^t \cdot \aleph \cdot \left(\Gamma^t\right)^\top = \mathbf{U}\mathbf{\Lambda}^t\left[\mathbf{U}^{-1}\aleph\mathbf{U}^{-1}\right]\mathbf{\Lambda}^t\mathbf{U}.$$

Computing $U^{-1}\aleph\mathbf{U}^{-1}$: From Property 2,

$$\mathbf{U} = \frac{1}{\mathbf{\Delta}}\begin{bmatrix} \frac{r_-}{(1-r_-)} & 1 \\ 1 & \frac{1-r_+}{r_+} \end{bmatrix}, \qquad \mathbf{U}^{-1} = \begin{bmatrix} -a & (1-r_-)r_+ \\ (1-r_-)r_+ & -r_+r_- \end{bmatrix},$$

$$\mathbf{U}^{-1}\aleph\mathbf{U}^{-1} = \mathbf{U}^{-1}\begin{bmatrix} b^2 & ab \\ ab & a^2 \end{bmatrix}\mathbf{U}^{-1} = \mathbf{U}^{-1}\left(\begin{bmatrix} b \\ a \end{bmatrix}\right) \otimes \left(\begin{bmatrix} b \\ a \end{bmatrix}\right)\mathbf{U}^{-1},$$
$$= \left(\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix}\right) \otimes \left(\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix}\right),$$
$$\mathbf{U}\mathbf{\Lambda}^t\left[\mathbf{U}^{-1}\aleph\mathbf{U}^{-1}\right]\mathbf{\Lambda}^t\mathbf{U} = \left(\mathbf{U}\mathbf{\Lambda}^t\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix}\right) \otimes \left(\mathbf{U}\mathbf{\Lambda}^t\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix}\right),$$
$$\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} -a & (1-r_-)r_+ \\ (1-r_-)r_+ & -r_+r_- \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} -ab + (1-r_-)r_+a \\ (1-r_-)r_+b - r_+r_-a \end{bmatrix},$$

using $a = (1-r_+)(1-r_-), b = (1-r_+r_-)$,

$$= \begin{bmatrix} -a(1-r_+r_-) + (1-r_-)r_+a \\ (1-r_-)r_+(1-r_+r_-) - r_+r_-(1-r_+)(1-r_-) \end{bmatrix},$$
$$= \begin{bmatrix} a\left(-(1-r_+r_-) + (1-r_-)r_+\right) \\ r_+(1-r_-)\left(1-r_-r_+ - r_-(1-r_+)\right) \end{bmatrix} = \begin{bmatrix} a(r_+-1) \\ r_+(1-r_-)^2 \end{bmatrix}.$$

$$\mathbf{\Lambda}^t\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} r_+^t & 0 \\ 0 & r_-^t \end{bmatrix}\begin{bmatrix} a(r_+-1) \\ r_+(1-r_-)^2 \end{bmatrix} = \begin{bmatrix} ar_+^t(r_+-1) \\ r_+r_-^t(1-r_-)^2 \end{bmatrix}.$$

$$U\mathbf{\Lambda}^t\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix} = \frac{1}{\mathbf{\Delta}}\begin{bmatrix} \frac{r_-}{(1-r_-)} & 1 \\ 1 & \frac{1-r_+}{r_+} \end{bmatrix}\begin{bmatrix} ar_+^t(r_+-1) \\ r_+r_-^t(1-r_-)^2 \end{bmatrix},$$
$$= \frac{1}{\mathbf{\Delta}}\begin{bmatrix} \frac{r_-}{(1-r_-)}ar_+^t(r_+-1) + r_+r_-^t(1-r_-)^2 \\ ar_+^t(r_+-1) + (1-r_+)r_-^t(1-r_-)^2 \end{bmatrix}.$$

Using $a = (1-r_+)(1-r_-), b = (1-r_+r_-)$,

$$= \frac{1}{\mathbf{\Delta}}\begin{bmatrix} -r_-r_+^t(r_+-1)^2 + r_+r_-^t(1-r_-)^2 \\ ar_+^t(r_+-1) + ar_-^t(1-r_-) \end{bmatrix},$$
$$= \frac{1}{\mathbf{\Delta}}\begin{bmatrix} -\left[r_-(r_+-1)^2r_+^t - r_+(1-r_-)^2r_-^t\right] \\ -a\left[(1-r_+)r_+^t - (1-r_-)r_-^t\right] \end{bmatrix}.$$

$$
\begin{aligned}
\Gamma^t \cdot \aleph \cdot \left(\Gamma^t\right)^\top &= \left(U\boldsymbol{\Lambda}^t\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix}\right) \otimes \left(U\boldsymbol{\Lambda}^t\mathbf{U}^{-1}\begin{bmatrix} b \\ a \end{bmatrix}\right), \\
&= \frac{1}{\boldsymbol{\Delta}}\begin{bmatrix} -\left[r_-\left(1-r_+\right)^2 r_+^t - r_+\left(1-r_-\right)^2 r_-^t\right] \\ -a\left[\left(1-r_+\right) r_+^t - \left(1-r_-\right) r_-^t\right] \end{bmatrix} \otimes \frac{1}{\boldsymbol{\Delta}}\begin{bmatrix} -\left[r_-\left(r_+-1\right)^2 r_+^t - r_+\left(1-r_-\right)^2 r_-^t\right] \\ -a\left[\left(1-r_+\right) r_+^t - \left(1-r_-\right) r_-^t\right] \end{bmatrix}, \\
&= \begin{bmatrix} \nu_{11}(t) & \nu_{12}(t) \\ \nu_{12}(t) & \nu_{22}(t) \end{bmatrix}.
\end{aligned}
$$

where

$$\boldsymbol{\Delta}^2\,\nu_{11}(t) \stackrel{\text{def}}{=} \left[r_-\left(1-r_+\right)^2 r_+^t - r_+\left(1-r_-\right)^2 r_-^t\right]^2, \tag{75}$$

$$\boldsymbol{\Delta}^2\,\nu_{22}(t) \stackrel{\text{def}}{=} a^2\left[\left(1-r_+\right) r_+^t - \left(1-r_-\right) r_-^t\right]^2, \tag{76}$$

$$\boldsymbol{\Delta}^2\,\nu_{12}(t) \stackrel{\text{def}}{=} a\left(r_-\left(1-r_+\right)^2 r_+^t - r_+\left(1-r_-\right)^2 r_-^t\right)\left(\left(1-r_+\right) r_+^t - \left(1-r_-\right) r_-^t\right). \tag{77}$$

Using these,

$$\sum_{t=0}^{\infty} \Gamma^t \cdot \aleph \cdot \left(\Gamma^t\right)^\top = \begin{bmatrix} \sum_{t=0}^{\infty} \nu_{11}(t) & \sum_{t=0}^{\infty} \nu_{12}(t) \\ \sum_{t=0}^{\infty} \nu_{12}(t) & \sum_{t=0}^{\infty} \nu_{22}(t) \end{bmatrix}.$$

Evaluating $\sum_{t=0}^{\infty} \nu_{11}(t)$:

$$
\begin{aligned}
\boldsymbol{\Delta}^2 \sum_{t=0}^{\infty} \nu_{11}(t) &= \sum_{t=0}^{\infty} \left[r_-\left(1-r_+\right)^2 r_+^t - r_+\left(1-r_-\right)^2 r_-^t\right]^2, \\
&= \left[r_-\left(1-r_+\right)^2 - r_+\left(1-r_-\right)^2\right]^2 \\
&\quad + \sum_{t=1}^{\infty}\left[r_-^2\left(1-r_+\right)^4 r_+^{2t} + r_+^2\left(1-r_-\right)^4 r_-^{2t} - 2r_-\left(1-r_+\right)^2 r_+\left(1-r_-\right)^2 r_+^t r_-^t\right].
\end{aligned}
$$

From Property 2, when $0 < a, b < 1$ then $|r_+|, |r_-| < 1$. Hence, the following holds,

$$\sum_{t=1}^{\infty} r_+^{2t} = \frac{r_+^2}{1-r_+^2}, \quad \sum_{t=1}^{\infty} r_-^{2t} = \frac{r_-^2}{1-r_-^2}, \quad \sum_{t=1}^{\infty} r_+^t r_-^t = \frac{r_+ r_-}{1-r_+ r_-},$$

$$
\begin{aligned}
r_-\left(1-r_+\right)^2 - r_+\left(1-r_-\right)^2 &= \left(1-\left(1-r_-\right)\right)\left(1-r_+\right)^2 - \left(1-\left(1-r_+\right)\right)\left(1-r_-\right)^2 \\
&= \left(1-r_+\right)^2 - \left(1-r_-\right)^2 - \left(1-r_-\right)\left(1-r_+\right)^2 + \left(1-r_+\right)\left(1-r_-\right)^2, \\
&= \left(r_- - r_+\right)\left(2-r_- - r_+\right) - \left(1-r_-\right)\left(1-r_+\right)\left(r_- - r_+\right), \\
&= \left(r_- - r_+\right)\left[\left(2-r_- - r_+\right) - \left(1-r_-\right)\left(1-r_+\right)\right], \\
&= \left(r_- - r_+\right)\left[1 - r_+ r_-\right] = -\boldsymbol{\Delta} b.
\end{aligned}
$$

From here, the first term of the sumation is as follows,

$$r_-\left(1-r_+\right)^2 - r_+\left(1-r_-\right)^2 = -\boldsymbol{\Delta} b. \tag{78}$$

To calculate the sum of the remaining terms,

$$\begin{aligned}\Delta^2\left(\sum_{t=0}^{\infty}\nu_{11}(t)-b^2\right) &= \sum_{t=1}^{\infty}\left[r_-^2\left(1-r_+\right)^4 r_+^{2t}+r_+^2\left(1-r_-\right)^4 r_-^{2t}-2r_-\left(1-r_+\right)^2 r_+\left(1-r_-\right)^2 r_+^t r_-^t\right],\\ &= r_+^2 r_-^2\sum_{t=0}^{\infty}\left[\left(1-r_+\right)^4 r_+^{2t}+\left(1-r_-\right)^4 r_-^{2t}-2r_-\left(1-r_+\right)^2 r_+\left(1-r_-\right)^2 r_+^t r_-^t\right],\\ &= r_+^2 r_-^2\sum_{t=0}^{\infty}\left[r_-^2\left(1-r_+\right)^2 r_+^t-\left(1-r_-\right)^2 r_-^t\right]^2.\end{aligned}$$

Invoking Lemma 39,

$$\begin{aligned}\Delta^2\left(\sum_{t=0}^{\infty}\nu_{11}(t)-b^2\right) &= \Delta^2 r_+^2 r_-^2\left[\frac{a(4-(a+2b))+(a+2b)^2}{2b(4-(a+2b))}\right],\\ \left(\sum_{t=0}^{\infty}\nu_{11}(t)-b^2\right) &= (1-b)^2\left[\frac{a(4-(a+2b))+(a+2b)^2}{2b(4-(a+2b))}\right].\end{aligned}$$

Using simple algebraic manipulations summation of $\nu_{11}(t)$'s can be compactly written as follows,

$$\begin{aligned}\sum_{t=0}^{\infty}\nu_{11}(t) &= b^2+(1-b)^2\left[\frac{a(4-(a+2b))+(a+2b)^2}{2b(4-(a+2b))}\right],\\ &= b^2+(1-b)^2\left[\frac{4a+(a+2b)\left((a+2b)-a\right)}{2b(4-(a+2b))}\right],\\ &= b^2+(1-b)^2\left[\frac{4a+2b(a+2b)}{2b(4-(a+2b))}\right],\\ &= b^2+\frac{(a+2b)(1-b)^2}{(4-(a+2b))}+\left[\frac{4a(1-b)^2}{2b(4-(a+2b))}\right],\\ &= \frac{b^2(4-(a+2b))+(a+2b)(1-b)^2}{(4-(a+2b))}+\left[\frac{4a(1-b)^2}{2b(4-(a+2b))}\right].\\ b^2(4-(a+2b))+(a+2b)(1-b)^2 &= b^2(4-(a+2b))+(a+2b)\left(1-2b+b^2\right)\\ &= 4b^2+(a+2b)\left(1-2b\right)=a+2b-2ab,\\ \sum_{t=0}^{\infty}\nu_{11}(t) &= \frac{a+2b-2ab}{(4-(a+2b))}+\left[\frac{2a(1-b)^2}{b(4-(a+2b))}\right],\\ &= \frac{b(a+2b-2ab)+2a(1-b)^2}{b(4-(a+2b))}\\ &= \frac{ab+2b^2-2ab^2+2a-4ab+2ab^2}{b(4-(a+2b))}=\frac{2a+b(2b-3a)}{b(4-(a+2b))}.\end{aligned}$$

Hence,

$$\sum_{t=0}^{\infty}\nu_{11}(t)=\frac{2a}{b(4-(a+2b))}+\frac{2b-3a}{4-(a+2b)}. \tag{79}$$

Evaluating $\sum_{t=0}^{\infty} \nu_{22}(t)$: From Eq.(77),

$$\Delta^2 \, \nu_{22}(t) = a^2 \left[(1-r_+)\, r_+^t - (1-r_-)\, r_-^t\right]^2,$$
$$\Delta^2 \sum_{t=0}^{\infty} \nu_{22}(t) = a^2 \sum_{t=0}^{\infty} (1-r_+)^2\, r_+^{2t} + (1-r_-)^2\, r_-^{2t} - 2\,(1-r_+)\,(1-r_-)\, r_+^t r_-^t.$$

From Property 2, when $0 < a, b < 1$ then $|r_+|, |r_-| < 1$. Hence, the following holds,

$$\sum_{t=0}^{\infty} r_+^{2t} = \frac{1}{1-r_+^2}, \quad \sum_{t=0}^{\infty} r_-^{2t} = \frac{1}{1-r_-^2}, \quad \sum_{t=0}^{\infty} r_+^t r_-^t = \frac{1}{1-r_+r_-},$$
$$\Delta^2 \sum_{t=0}^{\infty} \nu_{22}(t) = a^2 \left[(1-r_+)^2 \frac{1}{1-r_+^2} + (1-r_-)^2 \frac{1}{1-r_-^2} - 2\frac{(1-r_+)\,(1-r_-)}{1-r_+r_-}\right].$$
$$= a^2 \left[\frac{1-r_+}{1+r_+} + \frac{1-r_-}{1+r_-} - 2\frac{(1-r_+)\,(1-r_-)}{1-r_+r_-}\right].$$

$$\Delta^2 \sum_{t=0}^{\infty} \nu_{22}(t) = a^2 \left[\frac{1-r_+}{1+r_+} + \frac{1-r_-}{1+r_-} - 2\frac{(1-r_+)\,(1-r_-)}{1-r_+r_-}\right]. \tag{80}$$

Considering the computation in the right part,

$$\frac{1-r_+}{1+r_+} + \frac{1-r_-}{1+r_-} = \frac{(1-r_+)\,(1+r_-) + (1-r_-)\,(1+r_+)}{(1+r_+)\,(1+r_-)} = \frac{2\,(1-r_+r_-)}{(1+r_+)\,(1+r_-)},$$
$$\frac{1-r_+}{1+r_+} + \frac{1-r_-}{1+r_-} - 2\frac{(1-r_+)\,(1-r_-)}{1-r_+r_-} = \frac{2\,(1-r_+r_-)}{(1+r_+)\,(1+r_-)} - 2\frac{(1-r_+)\,(1-r_-)}{1-r_+r_-}$$
$$= 2\frac{(1-r_+r_-)^2 - \left(1-r_+^2\right)\left(1-r_-^2\right)}{(1+r_+)\,(1+r_-)\,(1-r_+r_-)}.$$

Computing the numerator, we get the following,

$$(1-r_+r_-)^2 - \left(1-r_+^2\right)\left(1-r_-^2\right) = 1 - 2r_+r_- + r_+^2 r_-^2 - \left(1 - r_+^2 - r_-^2 + r_+^2 r_-^2\right),$$
$$= r_+^2 + r_-^2 - 2r_+r_- = \Delta^2.$$

The denominator from Eq.(94),

$$(1+r_-)\,(1+r_+)\,(1-r_+r_-) = b(4-(a+2b)).$$

Substituting these back in Eq.(80), we get

$$\Delta^2 \sum_{t=0}^{\infty} \nu_{22}(t) = \frac{2a^2\Delta^2}{b(4-(a+2b))}.$$

Hence,

$$\sum_{t=0}^{\infty} \nu_{22}(t) = \frac{2a^2}{b(4-(a+2b))}. \tag{81}$$

Evaluating $\sum_{t=0}^{\infty} \nu_{12}(t)$: From Eq.(77),

$$\begin{aligned}
\boldsymbol{\Delta}^2\, \nu_{12}(t) &= a\left(r_-\left(1-r_+\right)^2 r_+^t - r_+\left(1-r_-\right)^2 r_-^t\right)\left(\left(1-r_+\right) r_+^t - \left(1-r_-\right) r_-^t\right), \\
&= a\left[r_-\left(1-r_+\right)^3 r_+^{2t} + r_+\left(1-r_-\right)^3 r_-^{2t} - \left\{r_-\left(1-r_+\right) + r_+\left(1-r_-\right)\right\}\left(1-r_+\right)\left(1-r_-\right) r_+^t r_-^t\right],
\end{aligned}$$

$$\boldsymbol{\Delta}^2 \sum_{t=0}^{\infty} \nu_{12}(t) = a\sum_{t=0}^{\infty}\left[r_-\left(1-r_+\right)^3 r_+^{2t} + r_+\left(1-r_-\right)^3 r_-^{2t} - \left\{r_-\left(1-r_+\right) + r_+\left(1-r_-\right)\right\}\left(1-r_+\right)\left(1-r_-\right) r_+^t r_-^t\right].$$

For $0 < a, b < 1$, we have $|r_+|, |r_-| < 1$ from Property 2. Hence, the following holds,

$$\sum_{t=0}^{\infty} r_+^{2t} = \frac{1}{1-r_+^2}, \quad \sum_{t=0}^{\infty} r_-^{2t} = \frac{1}{1-r_-^2}, \quad \sum_{t=0}^{\infty} r_+^t r_-^t = \frac{1}{1-r_+ r_-}.$$

$$\begin{aligned}
&= a\left[\frac{r_-\left(1-r_+\right)^3}{1-r_+^2} + \frac{r_+\left(1-r_-\right)^3}{1-r_-^2} - \left\{r_-\left(1-r_+\right) + r_+\left(1-r_-\right)\right\}\frac{\left(1-r_+\right)\left(1-r_-\right)}{1-r_+ r_-}\right]. \\
&= a\left[\frac{r_-\left(1-r_+\right)^2}{1+r_+} + \frac{r_+\left(1-r_-\right)^2}{1+r_-} - \left\{r_-\left(1-r_+\right) + r_+\left(1-r_-\right)\right\}\frac{\left(1-r_+\right)\left(1-r_-\right)}{1-r_+ r_-}\right].
\end{aligned}$$

Using

$$2\left(1-r_+ r_-\right) = \left(1-r_-\right)\left(1+r_+\right) + \left(1-r_+\right)\left(1+r_-\right),$$

$$2r_-\left(1+r_-\right)\left(1-r_+\right)^2\left(1-r_+ r_-\right) = r_-\left(1-r_+\right)\left(1-r_-^2\right)\left(1-r_+^2\right) + r_-\left(1+r_-\right)^2\left(1-r_+\right)^3. \tag{82}$$

Similarly by symmetry

$$2r_+\left(1+r_+\right)\left(1-r_-\right)^2\left(1-r_+ r_-\right) = r_+\left(1-r_-\right)\left(1-r_-^2\right)\left(1-r_+^2\right) + r_+\left(1+r_+\right)^2\left(1-r_-\right)^3. \tag{83}$$

$$\begin{aligned}
&\left[r_-\left(1-r_+\right) + r_+\left(1-r_-\right)\right]\left(1-r_+^2\right)\left(1-r_-^2\right), \\
&\qquad = r_-\left(1-r_+\right)\left(1-r_+^2\right)\left(1-r_-^2\right) + r_+\left(1-r_-\right)\left(1-r_+^2\right)\left(1-r_-^2\right).
\end{aligned} \tag{84}$$

Combining them,

$$\begin{aligned}
\text{Eq.}(82) + \text{Eq.}(83) - 2 * \text{Eq.}(84) &= r_-\left(1+r_-\right)^2\left(1-r_+\right)^3 + r_+\left(1+r_+\right)^2\left(1-r_-\right)^3 \\
&\quad - r_-\left(1-r_+\right)\left(1-r_+^2\right)\left(1-r_-^2\right) - r_+\left(1-r_-\right)\left(1-r_+^2\right)\left(1-r_-^2\right), \\
&= r_-\left(1+r_-\right)^2\left(1-r_+\right)^3 - r_-\left(1-r_+\right)\left(1-r_+^2\right)\left(1-r_-^2\right), \\
&\quad + r_+\left(1+r_+\right)^2\left(1-r_-\right)^3 - r_+\left(1-r_-\right)\left(1-r_+^2\right)\left(1-r_-^2\right), \\
&= r_-\left(1+r_-\right)\left(1-r_+\right)^2\left[\left(1+r_-\right)\left(1-r_+\right) - \left(1+r_+\right)\left(1-r_-\right)\right], \\
&\quad + r_+\left(1+r_+\right)\left(1-r_-\right)^2\left[\left(1+r_+\right)\left(1-r_-\right) - \left(1-r_+\right)\left(1+r_-\right)\right], \\
&= 2r_-\left(1+r_-\right)\left(1-r_+\right)^2\left[r_- - r_+\right] + 2r_+\left(1+r_+\right)\left(1-r_-\right)^2\left[r_+ - r_-\right], \\
&= -2\boldsymbol{\Delta}\left[r_-\left(1+r_-\right)\left(1-r_+\right)^2 - r_+\left(1+r_+\right)\left(1-r_-\right)^2\right].
\end{aligned}$$

Evaluating

$$r_-\left(1+r_-\right)\left(1-r_+\right)^2 - r_+\left(1+r_+\right)\left(1-r_-\right)^2 = r_-\left(1-r_+\right)^2 - r_+\left(1-r_-\right)^2 + r_-^2\left(1-r_+\right)^2 - r_+^2\left(1-r_-\right)^2 .$$

From Eq.(78), we have the following,

$$\begin{aligned}
r_-\left(1-r_+\right)^2 - r_+\left(1-r_-\right)^2 &= -\mathbf{\Delta} b,\\
r_-^2\left(1-r_+\right)^2 - r_+^2\left(1-r_-\right)^2 &= \left[r_-\left(1-r_+\right) - r_+\left(1-r_-\right)\right]\left[r_-\left(1-r_+\right) + r_+\left(1-r_-\right)\right],\\
&= -\mathbf{\Delta}\left[r_+ + r_- - 2r_+r_-\right] = -\mathbf{\Delta}\left[2-(a+b)-2(1-b)\right],\\
&= -\mathbf{\Delta}\left[b-a\right].
\end{aligned}$$

$$r_-\left(1+r_-\right)\left(1-r_+\right)^2 - r_+\left(1+r_+\right)\left(1-r_-\right)^2 = -\mathbf{\Delta} b - \mathbf{\Delta}\left[b-a\right] = -\mathbf{\Delta}\left(2b-a\right).$$

So the numerator of $\mathbf{\Delta}^2 \sum_{t=0}^{\infty} \nu_{12}(t)$

$$\begin{aligned}
\frac{a}{2}\left(\text{Eq.}(82) + \text{Eq.}(83) - 2 * \text{Eq.}(84)\right) &= \frac{a}{2}\left(-2\mathbf{\Delta}(-\mathbf{\Delta}\left(2b-a\right))\right),\\
&= \mathbf{\Delta}^2 a\left(2b-a\right).
\end{aligned}$$

the denominator is

$$\left(1+r_-\right)\left(1+r_+\right)\left(1-r_+r_-\right) = b(4-(a+2b)).$$

Finally, we have

$$\sum_{t=0}^{\infty} \nu_{12}(t) = \frac{a(2b-a)}{b(4-(a+2b))}. \tag{85}$$

From Eq.(85), Eq.(81), Eq.(79)

$$\begin{aligned}
\sum_{t=0}^{\infty} \Gamma^t \aleph \left(\Gamma^t\right)^\top &= \begin{bmatrix} \frac{2a}{b(4-(a+2b))} + \frac{2b-3a}{4-(a+2b)} & \frac{a(2b-a)}{b(4-(a+2b))} \\ \frac{a(2b-a)}{b(4-(a+2b))} & \frac{2a^2}{b(4-(a+2b))} \end{bmatrix}\\
&= \frac{1}{b(4-(a+2b))}\begin{bmatrix} 2a+b(2b-3a) & a(2b-a) \\ a(2b-a) & 2a^2 \end{bmatrix}.
\end{aligned}$$

This proves the lemma. ∎

Lemma 38 *For $0 < a, b < 1$, with Γ and $\aleph$ of form*

$$\Gamma = \begin{bmatrix} 1-b & 1-b \\ -a & 1-a \end{bmatrix}, \qquad \aleph = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

The series

$$\begin{bmatrix} b & a \\ b & a \end{bmatrix} \sum_{t=0}^{\infty} \left(\Gamma^t\right)^\top \aleph \Gamma^t \begin{bmatrix} b & b \\ a & a \end{bmatrix} = \left(\frac{2a}{b(4-(a+2b))} + \frac{a+2b}{(4-(a+2b))}\right)\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{86}$$

Proof To calculate the exponents of Γ we use the eigendecomposition from Property 2,

$$\Gamma = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1},$$
$$\Gamma^t = \mathbf{U}\mathbf{\Lambda}^t\mathbf{U}^{-1},$$
$$\left(\Gamma^t\right)^\top \cdot \aleph \cdot \Gamma^t = \mathbf{U}^{-1}\mathbf{\Lambda}^t\mathbf{U} \cdot \aleph \cdot \left(\mathbf{U}^{-1}\mathbf{\Lambda}^t\mathbf{U}\right)^\top.$$

Using the fact in Property 2, that $\mathbf{U}, \mathbf{U}^{-1}$ are symmetric.

$$\begin{bmatrix} b & a \\ b & a \end{bmatrix} \left(\Gamma^t\right)^\top \cdot \aleph \cdot \Gamma^t \begin{bmatrix} b & a \\ b & a \end{bmatrix}^\top = \begin{bmatrix} b & a \\ b & a \end{bmatrix} \mathbf{U}^{-1}\mathbf{\Lambda}^t \left[\mathbf{U}\aleph\mathbf{U}\right] \mathbf{\Lambda}^t\mathbf{U}^{-1} \begin{bmatrix} b & a \\ b & a \end{bmatrix}^\top.$$

$$\mathbf{U}\aleph\mathbf{U} = \mathbf{U}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\mathbf{U} = \mathbf{U}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \otimes \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)\mathbf{U},$$
$$= \left(\mathbf{U}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \otimes \left(\mathbf{U}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right),$$
$$\mathbf{U}^{-1}\mathbf{\Lambda}^t\left[\mathbf{U}\aleph\mathbf{U}\right]\mathbf{\Lambda}^t\mathbf{U}^{-1} = \left(\mathbf{U}^{-1}\mathbf{\Lambda}^t\mathbf{U}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \otimes \left(\mathbf{U}^{-1}\mathbf{\Lambda}^t\mathbf{U}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right),$$
$$\begin{bmatrix} b & a \\ b & a \end{bmatrix} \left(\Gamma^t\right)^\top \cdot \aleph \cdot \Gamma^t \begin{bmatrix} b & a \\ b & a \end{bmatrix}^\top = \left(\begin{bmatrix} b & a \\ b & a \end{bmatrix}\mathbf{U}^{-1}\mathbf{\Lambda}^t\mathbf{U}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \otimes \left(\begin{bmatrix} b & a \\ b & a \end{bmatrix}\mathbf{U}^{-1}\mathbf{\Lambda}^t\mathbf{U}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right).$$

From eigendecomposition given in Property 2,

$$\mathbf{\Delta U} = \begin{bmatrix} \frac{r_-}{(1-r_-)} & 1 \\ 1 & \frac{1-r_+}{r_+} \end{bmatrix},$$
$$\mathbf{\Delta U}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{r_-}{(1-r_-)} & 1 \\ 1 & \frac{1-r_+}{r_+} \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \dfrac{1}{(1-r_-)} \\ \dfrac{1}{r_+} \end{bmatrix},$$
$$\mathbf{\Delta \Lambda}^t\mathbf{U}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} r_+^t & 0 \\ 0 & r_-^t \end{bmatrix}\begin{bmatrix} \dfrac{1}{(1-r_-)} \\ \dfrac{1}{r_+} \end{bmatrix} = \begin{bmatrix} \dfrac{r_+^t}{(1-r_-)} \\ \dfrac{r_-^t}{r_+} \end{bmatrix}.$$

Again from Property 2 using $\mathbf{U}^{-1}$

$$\mathbf{U}^{-1} = \begin{bmatrix} -a & (1-r_-)r_+ \\ (1-r_-)r_+ & -r_+r_- \end{bmatrix},$$
$$\mathbf{\Delta U}^{-1}\mathbf{\Lambda}^t\mathbf{U}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -a & (1-r_-)r_+ \\ (1-r_-)r_+ & -r_+r_- \end{bmatrix}\begin{bmatrix} \frac{r_+^t}{(1-r_-)} \\ \frac{r_-^t}{r_+} \end{bmatrix},$$
$$= \begin{bmatrix} \frac{-ar_+^t}{(1-r_-)} + (1-r_-)r_-^t \\ r_+^{t+1} - r_-^{t+1} \end{bmatrix}.$$

Using $a = (1 - r_+)\,(1 - r_-)$,

$$\boldsymbol{\Delta}\mathbf{U}^{-1}\boldsymbol{\Lambda}^t\mathbf{U}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}-(1-r_+)r_+^t + (1-r_-)r_-^t\\ r_+^{t+1} - r_-^{t+1}\end{bmatrix} = \begin{bmatrix}-\left[(1-r_+)r_+^t - (1-r_-)r_-^t\right]\\ r_+^{t+1} - r_-^{t+1}\end{bmatrix}.$$

$$\begin{aligned}\boldsymbol{\Delta}\begin{bmatrix}b & a\\ b & a\end{bmatrix}\mathbf{U}^{-1}\boldsymbol{\Lambda}^t\mathbf{U}\begin{bmatrix}1\\1\end{bmatrix} &= \begin{bmatrix}b & a\\ b & a\end{bmatrix}\begin{bmatrix}-\left[(1-r_+)r_+^t - (1-r_-)r_-^t\right]\\ r_+^{t+1} - r_-^{t+1}\end{bmatrix},\\ &= \begin{bmatrix}-b\left((1-r_+)r_+^t - (1-r_-)r_-^t\right) + a\left(r_+^{t+1} - r_-^{t+1}\right)\\ -b\left((1-r_+)r_+^t - (1-r_-)r_-^t\right) + a\left(r_+^{t+1} - r_-^{t+1}\right)\end{bmatrix},\\ &= \begin{bmatrix}-\left[\left(b(1-r_+) - a(r_+)\right)r_+^t - \left(b(1-r_-) - a(r_-)\right)r_-^t\right]\\ -\left[\left(b(1-r_+) - a(r_+)\right)r_+^t - \left(b(1-r_-) - a(r_-)\right)r_-^t\right].\end{bmatrix}\end{aligned}$$

Using $a = (1 - r_+)\,(1 - r_-)\,, b = 1 - r_- r_+$,

$$\begin{aligned}b(1-r_+) - a(r_+) &= (1 - r_+ r_-)(1 - r_+) - (1 - r_+)\,(1 - r_-)\,r_+,\\ &= (1 - r_+)\left((1 - r_+ r_-) - r_+\,(1 - r_-)\right),\\ &= (1 - r_+)\,(1 - r_+ r_- - r_+ + r_+ r_-) = (1 - r_+)^2\,.\end{aligned}$$

By symmetry,

$$b(1-r_-) - a(r_-) = (1 - r_-)^2\,.$$

Substituting this back we get

$$\boldsymbol{\Delta}\begin{bmatrix}b & a\\ b & a\end{bmatrix}\mathbf{U}^{-1}\boldsymbol{\Lambda}^t\mathbf{U}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}-\left((1-r_+)^2\,r_+^t - (1-r_-)^2\,r_-^t\right)\\ -\left((1-r_+)^2\,r_+^t - (1-r_-)^2\,r_-^t\right)\end{bmatrix} \quad = -\left((1-r_+)^2\,r_+^t - (1-r_-)^2\,r_-^t\right)\begin{bmatrix}1\\1\end{bmatrix}.$$

$$\begin{aligned}\boldsymbol{\Delta}\begin{bmatrix}b & a\\ b & a\end{bmatrix}\mathbf{U}^{-1}\boldsymbol{\Lambda}^t\mathbf{U}\begin{bmatrix}1\\1\end{bmatrix}\otimes\boldsymbol{\Delta}\begin{bmatrix}b & a\\ b & a\end{bmatrix}\mathbf{U}^{-1}\boldsymbol{\Lambda}^t\mathbf{U}\begin{bmatrix}1\\1\end{bmatrix} &= \left((1-r_+)^2\,r_+^t - (1-r_-)^2\,r_-^t\right)^2\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix},\\ \boldsymbol{\Delta}^2\begin{bmatrix}b & a\\ b & a\end{bmatrix}\left(\Gamma^t\right)^\top\cdot\aleph\cdot\Gamma^t\begin{bmatrix}b & a\\ b & a\end{bmatrix}^\top &= \left((1-r_+)^2\,r_+^t - (1-r_-)^2\,r_-^t\right)^2\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}.\\ \begin{bmatrix}b & a\\ b & a\end{bmatrix}\sum_{t=0}^{\infty}\left(\Gamma^t\right)^\top\cdot\aleph\cdot\Gamma^t\begin{bmatrix}b & a\\ b & a\end{bmatrix}^\top &= \left[\sum_{t=0}^{\infty}\left(\frac{(1-r_+)^2\,r_+^t - (1-r_-)^2\,r_-^t}{\boldsymbol{\Delta}}\right)^2\right]\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}.\end{aligned}$$

Using Lemma 39,

$$\begin{bmatrix}b & a\\ b & a\end{bmatrix}\sum_{t=0}^{\infty}\left(\Gamma^t\right)^\top\cdot\aleph\cdot\Gamma^t\begin{bmatrix}b & a\\ b & a\end{bmatrix}^\top = \left(\frac{2a}{b(4 - (a + 2b))} + \frac{a + 2b}{(4 - (a + 2b))}\right)\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}.$$

■

Lemma 39 *With r_+, r_- defined by Property 2, and $\nu(t)$ defined by*

$$\nu(t) \overset{\text{def}}{=} \left[\frac{(1-r_+)^2 r_+^t - (1-r_-)^2 r_-^t}{\boldsymbol{\Delta}}\right]^2 .$$

the series

$$\sum_{t=0}^{\infty} \nu(t) = \frac{2a}{b(4-(a+2b))} + \frac{a+2b}{(4-(a+2b))}.$$

Proof

$$\boldsymbol{\Delta}^2 \left(\sum_{t=0}^{\infty} \nu(t)\right) = \sum_{t=0}^{\infty} \left[(1-r_+)^4 r_+^{2t} + (1-r_-)^4 r_-^{2t} - 2(1-r_+)^2 (1-r_-)^2 r_+^t r_-^t\right].$$

From Property 2 when $0 < a, b < 1$ then $|r_+|, |r_-| < 1$. Hence, the following holds,

$$\sum_{t=0}^{\infty} r_+^{2t} = \frac{1}{1-r_+^2}, \quad \sum_{t=0}^{\infty} r_-^{2t} = \frac{1}{1-r_-^2}, \quad \sum_{t=0}^{\infty} r_+^t r_-^t = \frac{1}{1-r_+ r_-}.$$

$$\begin{aligned}\boldsymbol{\Delta}^2 \left(\sum_{t=0}^{\infty} \nu(t)\right) &= (1-r_+)^4 \frac{1}{1-r_+^2} + (1-r_-)^4 \frac{1}{1-r_-^2} - 2(1-r_+)^2 (1-r_-)^2 \frac{1}{1-r_+ r_-}, \\ &= \left[\frac{(1-r_+)^3}{1+r_+} + \frac{(1-r_-)^3}{1+r_-} - 2\frac{(1-r_+)^2 (1-r_-)^2}{1-r_+ r_-}\right].\end{aligned}$$

$$\boldsymbol{\Delta}^2 \left(\sum_{t=0}^{\infty} \nu(t)\right) = \left[\frac{(1-r_+)^3}{1+r_+} + \frac{(1-r_-)^3}{1+r_-} - 2\frac{(1-r_+)^2 (1-r_-)^2}{1-r_+ r_-}\right]. \tag{87}$$

$$\begin{aligned}\boldsymbol{\Delta}^2 \left(\sum_{t=0}^{\infty} \nu(t)\right) = &\frac{\left[(1-r_+)^3 (1+r_-) + (1-r_-)^3 (1+r_+)\right] [1-r_+ r_-]}{(1+r_+)(1+r_-)[1-r_+ r_-]} \\ &- \frac{2(1-r_+)^2 (1-r_-)^2 (1+r_-)(1+r_+)}{(1+r_+)(1+r_-)[1-r_+ r_-]}.\end{aligned} \tag{88}$$

Note,

$$(1+r_-)(1+r_+) = 1 + r_- + r_+ + r_- r_+.$$

Using $r_- + r_+ = 2 - (a+b)$, $r_- r_+ = 1 - b$,

$$(1+r_-)(1+r_+) = 4 - (a+2b). \tag{89}$$

Using

$$2[1-r_+ r_-] = (1-r_+)(1+r_-) + (1+r_+)(1-r_-),$$
$$2[1-r_+ r_-](1-r_+)^3 (1+r_-) = [(1-r_+)(1+r_-) + (1+r_+)(1-r_-)](1-r_+)^3 (1+r_-).$$

$$2\left[1-r_{+}r_{-}\right](1-r_{+})^{3}(1+r_{-}) = (1-r_{+})^{4}(1+r_{-})^{2} + \left(1-r_{-}^{2}\right)\left(1-r_{+}^{2}\right)(1-r_{+})^{2}, \tag{90}$$

Symetrically,

$$2\left[1-r_{+}r_{-}\right](1-r_{-})^{3}(1+r_{+}) = (1-r_{-})^{4}(1+r_{+})^{2} + \left(1-r_{+}^{2}\right)\left(1-r_{-}^{2}\right)(1-r_{-})^{2}. \tag{91}$$

$$\begin{aligned} 4(1-r_{+})^{2}(1-r_{-})^{2}(1+r_{-})(1+r_{+}) = {} & 2(1+r_{-})(1-r_{+})^{2}(1+r_{+})(1-r_{-})^{2} \\ & + 2\left(1-r_{-}^{2}\right)\left(1-r_{+}^{2}\right)(1-r_{-})(1-r_{+}). \end{aligned} \tag{92}$$

Combining the above calculations,

$$\text{Eq.}(90) + \text{Eq.}(91) - \text{Eq.}(92) = \left[(1-r_{+})^{2}(1+r_{-}) - (1-r_{-})^{2}(1+r_{+})\right]^{2} + \left(1-r_{-}^{2}\right)\left(1-r_{+}^{2}\right)\left[(1-r_{+}) - (1-r_{-})\right]^{2} \tag{93}$$

Computing the two terms,

$$\begin{aligned} (1-r_{+})^{2}(1+r_{-}) - (1-r_{-})^{2}(1+r_{+}) &= (1-r_{+})^{2}(2-(1-r_{-})) - (1-r_{-})^{2}(2-(1-r_{+})), \\ &= 2\left[(1-r_{+})^{2} - (1-r_{-})^{2}\right] - \left[(1-r_{+})^{2}(1-r_{-}) - (1-r_{-})^{2}(1-r_{+})\right] \\ &= 2(r_{-}-r_{+})(2-r_{-}-r_{+}) - (1-r_{-})(1-r_{+})(r_{-}-r_{+}), \\ &= (r_{-}-r_{+})(4-2r_{-}-2r_{+}-1+r_{-}+r_{+}-r_{-}r_{+}) \\ &= -\boldsymbol{\Delta}(4-(1+r_{-})(1+r_{+})) = -\boldsymbol{\Delta}(a+2b), \quad \text{from Eq.(89)}, \\ \left(1-r_{-}^{2}\right)\left(1-r_{+}^{2}\right)\left[(1-r_{+})-(1-r_{-})\right]^{2} &= \boldsymbol{\Delta}^{2}(1-r_{-})(1-r_{+})(1+r_{-})(1+r_{+}), \\ &= \boldsymbol{\Delta}^{2}a(4-(a+2b)). \end{aligned}$$

The numerator of Eq.(88) as per Eq.(93) is $\frac{1}{2}\left(\boldsymbol{\Delta}^{2}a(4-(a+2b)) + \boldsymbol{\Delta}^{2}(a+2b)^{2}\right)$. From Eq.(89), the denominator is

$$(1+r_{-})(1+r_{+})(1-r_{+}r_{-}) = b(4-(a+2b)). \tag{94}$$

Now from Eq.(88), we have

$$\boldsymbol{\Delta}^{2}\left(\sum_{t=0}^{\infty}\nu(t)\right) = \boldsymbol{\Delta}^{2}\left[\frac{a(4-(a+2b)) + (a+2b)^{2}}{2b(4-(a+2b))}\right].$$

Hence,

$$\begin{aligned} \left(\sum_{t=0}^{\infty}\nu(t)\right) &= \left[\frac{a(4-(a+2b)) + (a+2b)^{2}}{2b(4-(a+2b))}\right], \\ &= \frac{4a + (a+2b)(a+2b-a)}{2b(4-(a+2b))} = \frac{4a}{2b(4-(a+2b))} + \frac{2b(a+2b)}{2b(4-(a+2b))}, \\ &= \frac{2a}{b(4-(a+2b))} + \frac{a+2b}{(4-(a+2b))}. \end{aligned}$$

This completes the proof. ■