Risk Bounds of Accelerated SGD for Overparameterized Linear Regression

Xuheng Li XUHENG.LI@CS.UCLA.EDU
Department of Computer Science, University of California, Los Angeles, CA 90095, USA
Yihe Deng YIHEDENG@CS.UCLA.EDU
Department of Computer Science, University of California, Los Angeles, CA 90095, USA
Jingfeng Wu UUUJF@BERKELEY.EDU
Simons Institute, University of California, Berkeley, CA 94720, USA
Dongruo Zhou DZ13@IU.EDU
Department of Computer Science, Indiana University Bloomington, IN 47408, USA
Quanquan Gu QGU@CS.UCLA.EDU
Department of Computer Science, University of California, Los Angeles, CA 90095, USA

Abstract

Accelerated stochastic gradient descent (ASGD) is a workhorse in deep learning. While existing optimization theory can explain its faster convergence, they fall short in explaining its better generalization. In this paper, we study the generalization of ASGD for overparameterized linear regression. We establish an instance-dependent excess risk bound for ASGD within each eigen-subspace of the data covariance matrix. Our analysis shows that (i) ASGD outperforms SGD in the subspace of small eigenvalues, while in the subspace of large eigenvalues, its bias error decays slower than SGD; and (ii) the variance error of ASGD is always larger than that of SGD. Our result suggests that ASGD can outperform SGD when the difference between the initialization and the true weight vector is mostly confined to the subspace of small eigenvalues.

1. Introduction

Momentum [12] is an important technique in optimization. In the context of convex and smooth optimization, Nesterov's momentum (accelerated gradient descent (AGD)) achieves the minimax optimal convergence rate [13] and provably accelerates the vanilla GD method. Recent work by Liu and Belkin [11] shows that stochastic gradient descent (SGD) can also be accelerated by momentum in the overparameterized setting. However, the effect of momentum on the generalization performance is less studied. It has been empirically shown that ASGD does not always outperform SGD [16], but there has been little theoretical work justifying this observation. Notable exceptions are Jain et al. [8] and [15], which provide excess risk bounds for accelerated SGD (ASGD) (a.k.a., SGD with momentum) for least squares problems in the strongly convex [8] and convex settings [15], respectively. However, both of their results are limited to the classical, finite-dimensional regime, and cannot be applied when the number of parameters exceeds the number of samples. On the other hand, a recent line of work completely characterizes the excess risk of SGD for least squares, even in the overparameterized regime [2–4, 7, 17, 18]. In particular, Wu et al. [17], Zou et al. [18] provide finite-sample and dimension-free excess risk bounds for SGD that are sharp for each least squares instance. Given these results, it becomes imperative to thoroughly investigate whether the inclusion of momentum proves beneficial in terms of generalization, particularly in the context of least squares problems.

Contributions. In this paper, we tackle the question by considering ASGD for (overparameterized) linear regression problems and comparing its performance with SGD.

- Our main result provides an instance-dependent excess risk bound for ASGD that can be applied in the overparameterized regime. Similar to the bounds for SGD in Wu et al. [17], Zou et al. [18], our bound for ASGD is independent of the ambient dimension and comprehensively depends on the spectrum of the data covariance matrix.
- Based on the excess risk bounds, we then compare the excess risk of ASGD and SGD. We find that the variance error of ASGD is always no smaller than that of SGD. Moreover, the bias error of ASGD is smaller than that of SGD along the small eigenvalue directions, but is larger than that of SGD along the large eigenvalue directions, with respect to the spectrum of the data covariance matrix. Thus momentum can help with generalization only if the main signals are aligned with small eigenvalue directions of the data covariance matrix and if the noise is small.

Notation. In this paper, scalars are denoted by non-boldface letters. Vectors and matrices are denoted by lower-case and upper-case boldface letters, respectively. Denote linear operators on matrices by upper-case calligraphic letters. Denote the inner product of vectors by $\langle \mathbf{u}, \mathbf{v} \rangle$. For a vector $\mathbf{v}$, denote its j-th entry as $(\mathbf{v})_j$; For a matrix $\mathbf{M}$, denote its ij-entry as $(\mathbf{M})_{ij}$. For a PSD matrix $\mathbf{M}$, define $\|\mathbf{u}\|_{\mathbf{M}}^2 = \mathbf{u}^\top \mathbf{M}\mathbf{u}$. Denote the 2-norm of vector $\mathbf{v}$ as $\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^\top \mathbf{v}}$. Denote the inner product of matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{2d \times 2d}$ as $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j=1}^{2d} (\mathbf{A})_{ij} (\mathbf{B})_{ij}$. The Kronecker product of matrices is denoted by $\otimes$. The operation of a linear matrix operator on a matrix is denoted by $\circ$.

2. Preliminaries

Linear Regression and ASGD. The goal of linear regression is to minimize the following risk:

$$L(\mathbf{w}) := 1/2 \cdot \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(y - \langle \mathbf{w}, \mathbf{x} \rangle)^2 \right],$$

where $\mathbf{x}$ is an input feature vector belonging to a Hilbert space, $y \in \mathbb{R}$ is the response, $\mathbf{w} \in \mathcal{H}$ is the weight vector to be optimized, and $\mathcal{D}$ is an underlying unknown distribution of the data. Let $\mathbf{H} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}|_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top]$ be the second-order moment matrix of the distribution $\mathcal{D}$. Let the eigen-decomposition of the Hessian be $\mathbf{H} = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of $\mathbf{H}$ sorted in descending order with $\mathbf{v}_i$'s being the corresponding eigenvectors.

In the t-th iteration, a sample $(\mathbf{x}_t, y_t) \sim \mathcal{D}$ is observed. The stochastic gradient is calculated by

$$\widehat{\nabla} L(\mathbf{w}) = -(y_t - \langle \mathbf{w}, \mathbf{x}_t \rangle)\mathbf{x}_t. \tag{2.1}$$

We follow the classical ASGD scheme [13], which maintains three sequences $\mathbf{w}_t$, $\mathbf{v}_t$ and $\mathbf{u}_t$. Let N be the number of samples observed, then for any $1 \le t \le N$, the update rules of $\mathbf{w}_t, \mathbf{v}_t, \mathbf{u}_t$ are:

$$\mathbf{u}_{t-1} = \alpha \mathbf{w}_{t-1} + (1-\alpha)\mathbf{v}_{t-1}, \tag{2.2}$$

$$\mathbf{w}_t = \mathbf{u}_{t-1} - \delta \widehat{\nabla} L(\mathbf{u}_{t-1}), \tag{2.3}$$

$$\mathbf{v}_t = \beta \mathbf{u}_{t-1} + (1-\beta)\mathbf{v}_{t-1} - \gamma \widehat{\nabla} L(\mathbf{u}_{t-1}), \tag{2.4}$$

where $\alpha, \beta, \gamma, \delta > 0$ are hyperparameters. The $\mathbf{v}_t$ sequence is initialized at $\mathbf{w}_0 \in \mathcal{H}$. In this work, following Jain et al. [8] and Zou et al. [18], we consider ASGD with tail averaging. The tail-averaged final output is $\overline{\mathbf{w}}_{s,s+N} := N^{-1} \sum_{t=s}^{s+N-1} \mathbf{w}_t$. With certain assumptions, $L(\mathbf{w})$ admits a unique global optimum denoted by $\mathbf{w}^* := \mathrm{argmin}_{\mathbf{w}} L(\mathbf{w})$. We focus on the overparameterized setting, where $d \gg N$ (or possibly countably infinite).

Assumptions. We introduce assumptions required in our analysis, following those of Wu et al. [17], Zou et al. [18].

Assumption 2.1 (Regularity conditions) *The second moment $\mathbf{H}$ exists, and $\mathrm{tr}(\mathbf{H})$ is finite. $\mathbf{H}$ is strictly positive definite, i.e., $\mathbf{H} \succ \mathbf{0}$. Thus, $L(\mathbf{w})$ admits a unique global optimum $\mathbf{w}^*$. The second-order moment of labels $\mathbb{E}[y^2]$ is also finite. Let $\mathcal{M} := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}]$ denote the fourth moment of $\mathbf{x}$, then $\mathcal{M}$ exists and is finite.*

Assumption 2.2 (Fourth moment condition) *Assume there exists a positive constant $\psi > 0$, such that for any PSD matrix $\mathbf{A}$, we have $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] \preceq \psi\, \mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}$.*

Specifically, when $\mathcal{D}$ is a Gaussian distribution, we have $\psi = 3$.

Assumption 2.3 (Noise condition) *Assume that*

$$\mathbf{\Sigma} := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\widehat{\nabla} L(\mathbf{w}^*) \otimes \widehat{\nabla} L(\mathbf{w}^*)] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \langle \mathbf{w}^*, \mathbf{x} \rangle)^2 \mathbf{x}\mathbf{x}^\top],$$

and $\sigma^2 := \|\mathbf{H}^{-\frac{1}{2}} \mathbf{\Sigma} \mathbf{H}^{-\frac{1}{2}}\|_2$ exist and are finite.

3. Main Results

We now provide an excess risk upper bound for ASGD.

3.1. Risk Bound of ASGD in the High-Dimensional Setting

Parameter choice. We select hyperparameters of ASGD as follows: We first pick a non-negative integer $\widetilde{\kappa}$. We then select parameters $\delta, \gamma, \beta, \alpha$ as follows, based on $\widetilde{\kappa}$:

$$\delta \leq \frac{1}{2\psi\, \mathrm{tr}(\mathbf{H})}, \quad \gamma \in \left[\delta,\ \frac{1}{2\psi \sum_{i>\widetilde{\kappa}} \lambda_i}\right], \quad \beta = \frac{\delta}{\psi \widetilde{\kappa} \gamma}, \quad \alpha = \frac{1}{1+\beta}. \tag{3.1}$$

we first introduce three quantities which are cutoffs of the spectrum of $\mathbf{H}$. The eigenvalues of $\mathbf{A}_i$ can be either complex or real, which depends on the range of λ_i. Define

$$\begin{aligned} k^\ddagger &:= \max\{i : \lambda_i \geq (\sqrt{q - c\delta} + \sqrt{c(q-\delta)})^2 / q^2\}, \\ k^\dagger &:= \max\{i : \lambda_i > (\sqrt{q - c\delta} - \sqrt{c(q-\delta)})^2 / q^2\}. \end{aligned} \tag{3.2}$$

We also define $\widehat{k}$ as $\widehat{k} := \max\{i : \lambda_i \geq (1-c)/\delta\}$. We can show that with our choice of parameters, we have $k^\ddagger \leq \widehat{k} \leq k^\dagger$ (see Appendix G.1).

For integers $0 \leq k_1 \leq k_2$, denote $\mathbf{H}_{k_1:k_2} := \sum_{i=k_1+1}^{k_2} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ and $\mathbf{H}_{k_1:\infty} := \sum_{i=k_1+1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$. Now we present the main result:

Theorem 3.1 *Under Assumptions 2.1, 2.2 and 2.3, with the parameter choice in (3.1), if $N(1-c) \geq 2$, the excess risk of tail-averaged iterate from ASGD satisfies:*

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s,s+N})] - L(\mathbf{w}^*) \leq 2 \cdot \textit{EffectiveVar} + 2 \cdot \textit{EffectiveBias}. \tag{3.3}$$

where the effective variance is bounded by

$$\textit{EffectiveVar} \leq \sigma^2 r\left[\frac{27k^*}{2N} + 18(s+N)\gamma^2 \sum_{i>k^*}\lambda_i^2\right] + \frac{\psi r}{N}\left[\frac{9k^*}{N} + 36N\gamma^2\sum_{i>k^*}\lambda_i^2\right] \cdot \left[\frac{14}{\delta}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:\widehat{k}}}\right.$$
$$\left. + \frac{10}{1-c}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}} + \frac{2}{\gamma+\delta}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{k^\dagger:k^*}} + 4(s+N)\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right],$$

and the effective bias is bounded by

$$\textit{EffectiveBias} \leq \frac{8(c\delta/q)^{2s}}{N^2\delta^2}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^\ddagger}} + \frac{4s^2}{N^2}c^s\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}_{k^\ddagger:k^\dagger}}$$
$$+ \frac{16c^s}{N^2\delta^2}\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{k^\ddagger:\widehat{k}}} + \frac{100c^s}{N^2(1-c)^2}\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}}$$
$$+ \frac{18}{N^2(\gamma+\delta)^2}\left\|\left(\mathbf{I} - \frac{\gamma+\delta}{2}\mathbf{H}\right)^s(\mathbf{w}_0 - \mathbf{w}^*)\right\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}} + 18\left\|\left(\mathbf{I} - \frac{\gamma+\delta}{2}\mathbf{H}\right)^s(\mathbf{w}_0 - \mathbf{w}^*)\right\|^2_{\mathbf{H}_{k^*:\infty}},$$

with $k^* = \max\{k : \lambda_k \geq 1/((\gamma+\delta)N)\}$, and

$$r := (1-\psi l)^{-1}, \quad l := \delta\,\mathrm{tr}(\mathbf{H})/2 + 1/(2\psi) + \gamma\sum_{i>\widetilde{\kappa}}\lambda_i/4.$$

Theorem 3.1 establishes the excess risk bound of ASGD under the overparameterized setting. To our knowledge, this is the first instance-dependent bound of ASGD within each eigen-subspace of $\mathbf{H}$. Our excess bound includes both the variance term, which depends on the randomness coming from the data distribution $\mathcal{D}$, and the bias term, which includes "accelerated convergence" terms brought by the ASGD.

Remark 3.2 *The cutoff index k^* is referred to as the effective dimension, which can be much smaller than the model dimensionality d, especially when the eigenvalues decay fast. We want to emphasize that similar effective dimension has also appeared in the previous work which analyzes the convergence of SGD under the overparameterized model setting [17, 18]. Nevertheless, the effective dimension of SGD is $k^*_{\mathrm{SGD}} := \max\{k : \lambda_k \geq 1/(\delta N)\}$, which is smaller than that in ASGD. In Section 4, we will provide a comparison of the risk bounds between SGD and ASGD.*

4. Comparison between ASGD and SGD

In this section, we first introduce the SGD update, which is given by $\mathbf{w}_t^{\mathrm{SGD}} = \mathbf{w}_{t-1}^{\mathrm{SGD}} - \delta\widehat{\nabla}L(\mathbf{w}_{t-1}^{\mathrm{SGD}})$, where δ satisfies the requirement in (3.1). Analogous to ASGD, tail-averaged SGD is defined as $\overline{\mathbf{w}}^{\mathrm{SGD}}_{s:s+N} := N^{-1}\sum_{t=s}^{s+N-1}\mathbf{w}_t^{\mathrm{SGD}}$.. The following theorem shows the existence of linear regression instances where ASGD outperforms SGD (the proof is given in Appendix F.2):

Theorem 4.1 (Informal) *There exists a class of linear regression instances and corresponding choice of parameter such that the excess risk bound of tail-averaged ASGD satisfies*

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) = \mathcal{O}(\sigma^2(N^{-1/2} + N^{-2} \cdot 0.9873^s)),$$

and the excess risk bound of tail-averaged SGD satisfies

$$\mathbb{E}[L(\overline{\mathbf{w}}^{SGD}_{s:s+N})] - L(\mathbf{w}^*) = \Omega(\sigma^2(N^{-1/2} + N^{-2} \cdot 0.996^s)).$$

Theorem 4.1 is inspired by the following comparison of the effective variance and bias of SGD and ASGD with the assumption that $s = \mathcal{O}(N)$. Under Assumptions 2.1, 2.2 and 2.3, Zou et al. [18] prove that, effective variance and effective bias of SGD satisfy:

$$\begin{aligned}
\mathrm{EffectiveVar} &\leq \sigma^2 r_{\mathrm{SGD}} \cdot \left[\frac{k^*_{\mathrm{SGD}}}{N} + (s+N)\delta^2 \sum_{i>k^*_{\mathrm{SGD}}} \lambda_i^2\right] \\
&+ \frac{4\psi r_{\mathrm{SGD}}}{N} \cdot \left[\frac{1}{\delta}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*_{\mathrm{SGD}}}} + (s+N)\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*_{\mathrm{SGD}}:\infty}}\right] \cdot \left[\frac{k^*_{\mathrm{SGD}}}{N} + N\delta^2 \sum_{i>k^*_{\mathrm{SGD}}} \lambda_i^2\right], \\
\mathrm{EffectiveBias} &\leq \frac{1}{\delta^2 N^2}\|(\mathbf{I} - \delta\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{0:k^*_{\mathrm{SGD}}}} + \|(\mathbf{I} - \delta\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}_{k^*_{\mathrm{SGD}}:\infty}},
\end{aligned}$$

where $r_{\mathrm{SGD}} = (1 - \psi\delta\,\mathrm{tr}(\mathbf{H}))^{-1}$ and $k^*_{\mathrm{SGD}} = \max\{i : \lambda_i \geq 1/(\delta N)\}$.

Comparison of effective variance. Assuming that the initial variance $\mathbf{w}_0 - \mathbf{w}^*$ is bounded, the effective variances of both ASGD and SGD are dominated by the terms with the σ^2 coefficient, i.e., the real variance. Thus, ignoring σ^2, r and r_{SGD} and constants, effective variance of ASGD in the subspace of λ_i is $\mathcal{O}(\min\{1/N, N\gamma^2\lambda_i^2\})$, compared to $\mathcal{O}(\min\{1/N, N\delta^2\lambda_i^2\})$ for SGD. With $\gamma \geq \delta$ according to the choice of parameters in (3.1), we conclude that the excess variance of ASGD in every subspace is larger than that of SGD.

Comparison of effective bias. Effective bias of both SGD and ASGD decay exponentially in s within each subspace. The decay rate of SGD is $(1 - \delta\lambda_i)^s$ in the subspace of λ_i. For ASGD,

1. When $i \leq k^\ddagger$, the decay rate in the subspace of λ_i is $(c\delta/q)^s$. By definition of $k^\ddagger$, we have $1 - \delta\lambda_i \leq c\delta/q$ (see Appendix G.1 for detailed proof).

2. When $k^\ddagger < i \leq k^\dagger$, the decay rate in the subspace of λ_i is $[c(1 - \delta\lambda_i)]^{s/2}$. According to the definition of $\widehat{k}$, when $k^\ddagger < i \leq \widehat{k}$, we have $1 - \delta\lambda_i \leq \sqrt{c(1 - \delta\lambda_i)}$; When $\widehat{k} < i \leq k^\dagger$, we have $1 - \delta\lambda_i \geq \sqrt{c(1 - \delta\lambda_i)}$.

3. When $i > k^\dagger$, the decay rate in the subspace of λ_i is $(1 - (\gamma + \delta)\lambda_i/2)^s$. By the choice of parameters (3.1), we have $\gamma \geq \delta$, so $1 - (\gamma + \delta)\lambda_i/2 \leq 1 - \delta\lambda_i$.

Combining the three cases above, we conclude that the effective bias of ASGD decays faster than that of SGD in eigen-subspaces of λ_i where $i > \widehat{k}$, while it decays slower than SGD in subspaces of λ_i where $i \leq \widehat{k}$. This phenomenon is illustrated in Figure 1. Therefore, ASGD can perform better than SGD if $\mathbf{w}_0 - \mathbf{w}^*$ is mostly refined to the eigen-subspaces of λ_i where $i > \widehat{k}$.

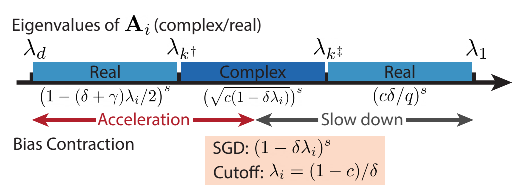


Figure 1: Illustration of the eigenspectrum.

5. Conclusion

In this work, we provide instance-dependent risk bounds for ASGD that are comprehensively dependent on the spectrum of the data covariance matrix. We show that ASGD always exhibit higher variance and higher bias along the large eigenvalues subspace than SGD. However, along the small eigenvalues subspace, its bias error is smaller than that of SGD. These together suggest that ASGD outperforms SGD only if the signals mostly align with the small eigenvalues subspaces of the data covariance and that the noise is small.

References

[1] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[2] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *arXiv preprint arXiv:2006.08212*, 2020.

[3] Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pages 205–213, 2015.

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

[5] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

[6] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017.

[7] Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1):8258–8299, 2017.

[8] Prateek Jain, M. Sham Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. *COLT*, pages 545–604, 2018.

[9] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[10] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

[11] Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. *arXiv preprint arXiv:1810.13395*, 2018.

[12] Yu E Nesterov. A method for solving the convex programming problem with convergence rate $o(\frac{1}{k^2})$. In *Dokl. Akad. Nauk SSSR,*, volume 269, pages 543–547, 1983.

[13] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

[14] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[15] Aditya Varre and Nicolas Flammarion. Accelerated sgd for non-strongly-convex least squares, 2022.

[16] Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate sgd. *arXiv preprint arXiv:2307.15196*, 2023.

[17] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. *The 39th International Conference on Machine Learning*, 2022.

[18] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *The 34th Annual Conference on Learning Theory*, 2021.

The organization of the appendix is as follows.

- In Appendix A, we discuss the related work.
- In Appendix B, we present empirical results on synthetic experiments to confirm our theoretical findings.
- In Appendix C, we present the detailed preliminaries of our results.
- In Appendix D, we discuss the choice of hyperparameters mentioned in (3.1).
- In Appendix E, we extend our main results to the Stochastic Heavy Ball method.
- In Appendix F.1, we prove Theorem 3.1, which depends on two key lemmas: Lemma F.3 to bound the variance term, and Lemma F.4 to bound the bias term.
- In Appendix F.2, we prove Theorem 4.1.
- In Appendix G, we provide the upper bounds for $\mathbf{A}$, which is the population version of $\widehat{\mathbf{A}}_t$, the update matrix of the noise $\boldsymbol{\eta}_t$. The upper bound of $\mathbf{A}$ is crucial to our proof.
- In Appendix H, we provide the upper bounds for a group of linear operators, which are crucial to our proof.
- In Appendix I, we provide the detailed proof of Lemma F.3.
- In Appendix J, we provide the detailed proof of Lemma F.4.
- In Appendix K, we prove Corollary K.1.
- In Appendix L, we provide the detailed proofs for the setting of standard basis.
- In Appendix M, we provide the proof of all remaining lemmas.

We note that, from a technical perspective, we extend the analysis of the stationary covariance matrix in Jain et al. [8] to the overparameterized setting, where we remove all dimension-dependent factors with a fine-grained analysis of the ASGD iterates. Our techniques might be of independent interest for analyzing ASGD in other settings.

Appendix A. Related Work

The generalization performances of SGD and ASGD applied to *underparameterized* linear regression have been studied in a line of works, based on the technique of bias-variance decomposition. It is shown that for SGD with iterate averaging from the beginning, bias error has a convergence rate of $\mathcal{O}(1/N^2)$ and variance has a convergence rate of $\mathcal{O}(d/N)$, where N is the number of calls of the stochastic oracle and d is the model dimension [3, 5, 6]. If the eigenvalue of the data covariance matrix is bounded away from zero, then the convergence rate of the bias error can be further improved with additional exponential shrinkage by taking tail averaging of the iterates [7].

For ASGD applied to linear regression, there are two cases: one with the assumption that the eigenvalue spectrum of the data covariance matrix is bounded away from zero (strongly convex) and the other without such assumption (general convex). For strongly convex linear regression, [8] show an accelerated convergence rate for the bias error of ASGD with constant stepsize and tail averaging, compared to that of tail-averaged SGD in [7]. We extend the use of linear operators and the techniques for bounding the operator spectrum in [8].

Recently, the generalization of ASGD applied to general convex linear regression is studied by [15]. Their result shows the acceleration of ASGD with time-varying parameters and weighted iterate averaging, especially for large N. The case of general convex linear regression is closer to the overparameterized setting where fast-decaying eigenspectrum is of special interest. However, their result is not applicable to the overparameterized linear regression because of the dimensionality dependence. Additionally, their result does not reveal the exponential bias decay of ASGD with constant stepsize.

The generalization performance of overparameterized linear regression has been studied by a line of works [1, 14]. For SGD applied to overparameterized linear regression, [18] replace the model dimensionality d with the effective dimension defined in terms of the eigenspectrum. This work manages to deal with any data covariance matrix, while prior works require certain assumptions [4]. Wu et al. [17] show a similar result for the last iterate of SGD with exponentially decaying stepsize.

Appendix B. Experiments

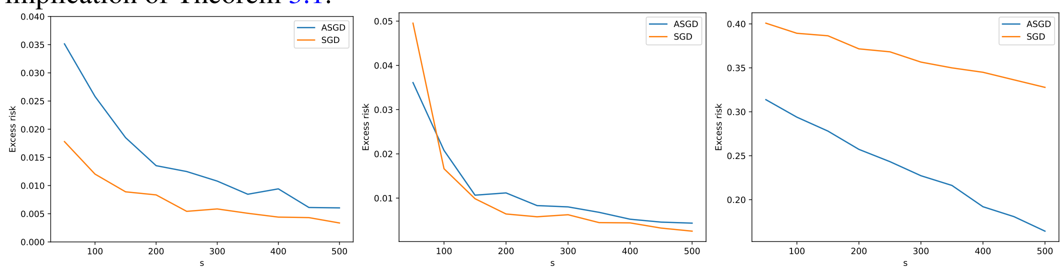
In this section, we empirically verify that ASGD can outperform SGD when $\mathbf{w}_0 - \mathbf{w}^*$ is mainly confined to the eigen-subspace of small eigenvalues.

Data model. Our experiments are based on the setting of overparameterized linear regression, where the model dimenstion is $d = 2000$. The data covariance matrix $\mathbf{H}$ is diagonal with eigenvalues $\lambda_i = i^{-2}$. Samples $\mathbf{x}_t$ follow Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{H})$, so Assumption 2.2 holds with $\psi = 3$. The ground truth is $\mathbf{w}^* = \mathbf{0}$, and the labels y_t follow the distribution $\mathcal{N}(0, \sigma^2)$ where $\sigma^2 = 0.01$.

Hyperparameters of ASGD and SGD. We select parameters of ASGD so that it satisfies the requirements in (3.1). We first let $\widetilde{\kappa} = 5$. According to (3.1), δ satisfies $\delta \le 1/\pi^2$, so we pick $\delta = 0.1$, which is also the stepsize of SGD. We then let $\alpha = 0.9875$, so that $(1-c)/\delta = 2(1-\alpha)/\delta = 0.25 = \lambda_2$, which implies that $\widehat{k} = 2$. Finally, we select $\beta = (1-\alpha)/\alpha$ and $\gamma = \delta/(\psi\widetilde{\kappa}\beta)$. We can verify that the parameters satisfy all requirements in (3.1).

We fix the length of tail averaging as $N = 500$, and conduct experiments on different s where $s = 50, 100, 150, \ldots, 500$. In each experiment, we measure $\overline{\mathbf{w}}_{s:s+N}^\top \mathbf{H} \overline{\mathbf{w}}_{s:s+N}$. For each s, we run the experiment 10 times and take the average of the test results.

We examine three different initializations: (a) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_1$, representing the case where $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of large eigenvalues, (b) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_2$, representing the case where $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of $\lambda_{\widehat{k}}$, and (c) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_{20}$, representing the case where $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of small eigenvalues. Experiment results are shown in Figure 2. We observe that ASGD indeed outperforms SGD in the scenario where $\mathbf{w}_0 - \mathbf{w}^*$ is mostly refined to the subspace of small eigenvalues, and performs worse than SGD when $\mathbf{w}_0 - \mathbf{w}^*$ is refined to the subspace of large eigenvalues. Additionally, the excess risks of SGD and ASGD are similar when $\mathbf{w}_0 - \mathbf{w}^*$ aligns with the subspace corresponding to $\lambda_{\widehat{k}}$, which is also aligns with the implication of Theorem 3.1.



(*a*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_1$. (*b*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_2$. (*c*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_{20}$.

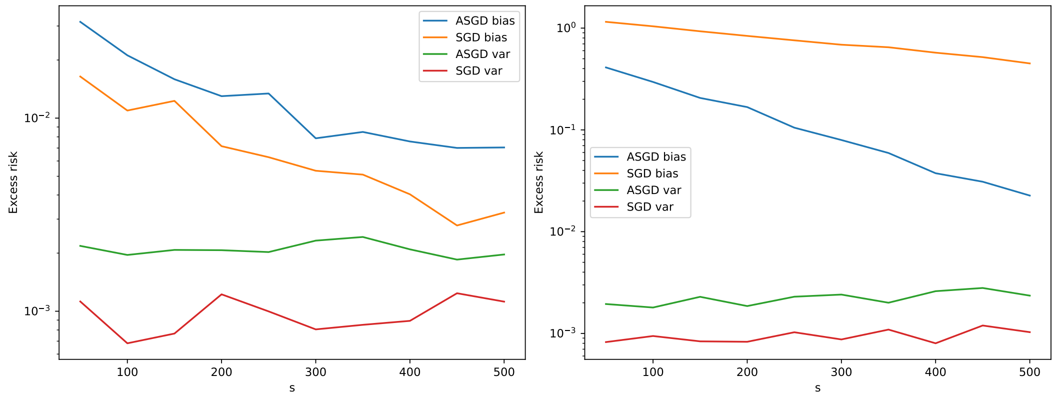
Figure 2: Comparison of excess risk of ASGD and SGD. The noise scale is $\sigma^2 = 0.01$. We run each experiment 10 times and take the average of the excess risk in the 10 trials.

We now provide additional experiments that display the bias and variance errors separately with different data covariance matrices $\mathbf{H}$. Similar to the experiments provided earlier in this section, the model dimension is set to be $d = 2000$, and the input $\mathbf{x}_t$ follows Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{H})$. We consider $\mathbf{H}$ with three types of spectrum: (i) $\lambda_k = k^{-2}$, (ii) $\lambda_k = k \log(k+1)$, and (iii) $\lambda_k = e^{-k/2}$. The ground truth weight vector is $\mathbf{w}^* = 0$, and the label y_t follows the distribution $y_t \sim \mathcal{N}(0, \sigma^2)$ where $\sigma = 0.2$.

Hyperparameters. We select the same hyperparameters of ASGD and SGD as the choice earlier in this section, i.e., $\psi = 3$, $\widetilde{\kappa} = 5$, $\delta = 0.1$, $\alpha = 0.9875$, $\beta = (1-\alpha)/\alpha$ and $\gamma = \delta/(\psi\widetilde{\kappa}\beta)$. We fix $N = 500$ and conduct experiments on different s where $s = 50, 100, \dots, 500$.

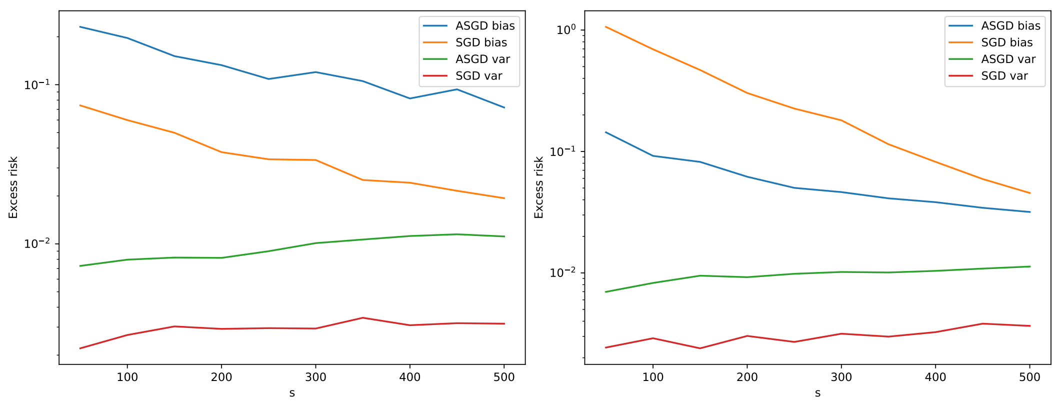
In each experiment, we measure both the bias error $(\overline{\mathbf{w}}^{\text{bias}}_{s:s+N})^\top \mathbf{H} \overline{\mathbf{w}}^{\text{bias}}_{s:s+N}$ and the variance error $(\overline{\mathbf{w}}^{\text{var}}_{s:s+N})^\top \mathbf{H} \overline{\mathbf{w}}^{\text{var}}_{s:s+N}$. For each s, we run the experiment 10 times, and take the average of the test results. We examine two initializations: (a) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_1$, which is the case where $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of large eigenvalues, and (b) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_{10}$, which is the case where $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of small eigenvalues.

The experimental results are shown in Figures 3, 4 and 5. In all experiments, the variance error of ASGD is larger than that of SGD. However, the bias error of ASGD decays faster than that of SGD when $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of small eigenvalues.



(*a*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_1$. (*b*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_{10}$.

Figure 3: Comparison of bias error and variance error of ASGD and SGD. The spectrum of $\mathbf{H}$ is $\lambda_k = k^{-2}$.



(*a*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_1$. (*b*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_{10}$.

Figure 4: Comparison of bias error and variance error of ASGD and SGD. The spectrum of $\mathbf{H}$ is $\lambda_k = k \log(k+1)$.

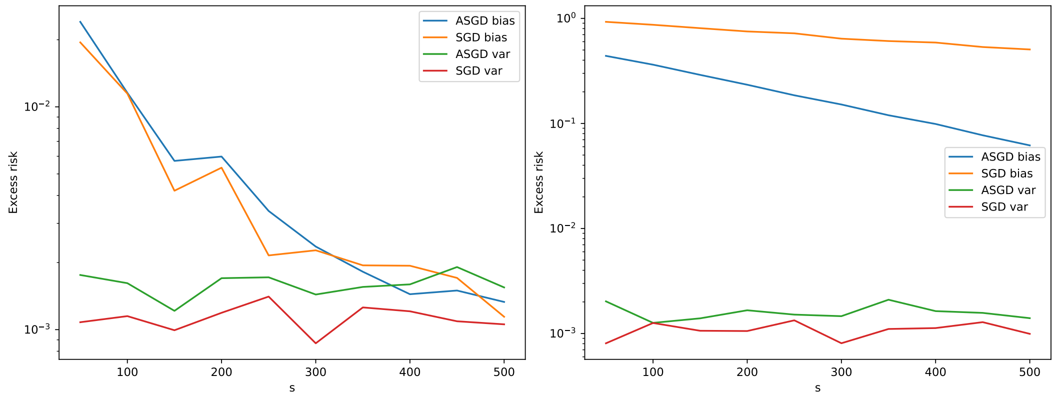
Appendix C. Preliminary

We first introduce some notations. Define the centered ASGD iterate as $\boldsymbol{\eta}_t := \begin{bmatrix} \mathbf{w}_t - \mathbf{w}^* \\ \mathbf{u}_t - \mathbf{w}^* \end{bmatrix}$. Define the tail averaged centered ASGD iterate as $\overline{\boldsymbol{\eta}}_{s,s+N} := N^{-1} \sum_{t=s}^{s+N-1} \boldsymbol{\eta}_t$. The excess risk is then

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s,s+N})] - L(\mathbf{w}^*) = \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}_{s,s+N} \otimes \overline{\boldsymbol{\eta}}_{s,s+N}] \right\rangle.$$

Denote the noise in each sample as $\epsilon_t := y_t - \langle \mathbf{w}^*, \mathbf{x}_t \rangle$. By (2.1), the stochastic gradient at $\mathbf{u}_{t-1}$ can be expressed as

$$\widehat{\nabla} L(\mathbf{u}_{t-1}) = -(\epsilon_t + \langle \mathbf{w}^*, \mathbf{x}_t \rangle - \langle \mathbf{u}_{t-1}, \mathbf{x}_t \rangle)\mathbf{x}_t = \mathbf{x}_t \mathbf{x}_t^\top (\mathbf{u}_{t-1} - \mathbf{w}^*) - \epsilon_t \mathbf{x}_t. \tag{C.1}$$



(*a*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_1$. (*b*) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_{10}$.

Figure 5: Comparison of bias error and variance error of ASGD and SGD. The specturm of $\mathbf{H}$ is $\lambda_k = e^{-k/2}$.

By substituting (C.1) into (2.3) and (2.4) and eliminating $\mathbf{v}_t$ using (2.2), we have

$$\boldsymbol{\eta}_t = \widehat{\mathbf{A}}_t \boldsymbol{\eta}_{t-1} + \boldsymbol{\zeta}_t, \quad \text{where} \quad \widehat{\mathbf{A}}_t := \begin{bmatrix} \mathbf{0} & \mathbf{I} - \delta \mathbf{x}_t \mathbf{x}_t^\top \\ -c\mathbf{I} & (1+c)\mathbf{I} - q\mathbf{x}_t\mathbf{x}_t^\top \end{bmatrix}, \quad \boldsymbol{\zeta}_t := \begin{bmatrix} \delta \cdot \epsilon_t \mathbf{x}_t \\ q \cdot \epsilon_t \mathbf{x}_t \end{bmatrix},$$

and $c := \alpha(1-\beta), q := \alpha\delta + (1-\alpha)\gamma$. Denote the expectation of $\widehat{\mathbf{A}}_t$ as

$$\mathbf{A} := \mathbb{E}[\widehat{\mathbf{A}}_t] = \begin{bmatrix} \mathbf{0} & \mathbf{I} - \delta \mathbf{H} \\ -c\mathbf{I} & (1+c)\mathbf{I} - q\mathbf{H} \end{bmatrix},$$

where $\mathbf{H} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}|_\mathbf{x}}[\mathbf{x}\mathbf{x}^\top]$ is the second-order moment matrix of the distribution $\mathcal{D}$, which is also the Hessian of $L(\mathbf{w})$. Similar to Jain et al. [8], we assume that $\mathbf{H}$ is diagonal, then $\mathbf{A}$ is block diagonal with each block being $\mathbf{A}_i := \begin{bmatrix} 0 & 1-\delta\lambda_i \\ -c & 1+c-q\lambda_i \end{bmatrix}$. In this work, we are particularly interested in analyzing the eigenvalues of $\mathbf{A}_i$, since the spectral norm of $\mathbf{A}_i$ determines the decay rate of the bias error in the subspace of λ_i.

Following the technique used extensively in previous works [4, 8, 10, 17, 18], we decompose the centered iterate $\boldsymbol{\eta}_t$ into the bias sequence $\boldsymbol{\eta}_t^{\text{bias}}$ and the variance sequence $\boldsymbol{\eta}_t^{\text{var}}$, defined recursively as

$$\boldsymbol{\eta}_t^{\text{bias}} = \widehat{\mathbf{A}}_t \boldsymbol{\eta}_{t-1}^{\text{bias}}, \quad \boldsymbol{\eta}_0^{\text{bias}} = \boldsymbol{\eta}_0; \tag{C.2}$$

$$\boldsymbol{\eta}_t^{\text{var}} = \widehat{\mathbf{A}}_t \boldsymbol{\eta}_{t-1}^{\text{var}} + \boldsymbol{\zeta}_t, \quad \boldsymbol{\eta}_0^{\text{var}} = \mathbf{0}. \tag{C.3}$$

The tail averaged iterate is then $\overline{\boldsymbol{\eta}}_{s:s+N} = \overline{\boldsymbol{\eta}}_{s:s+N}^{\text{bias}} + \overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}}$, where

$$\overline{\boldsymbol{\eta}}_{s:s+N}^{\text{bias}} := \frac{1}{N}\sum_{t=s}^{s+N-1} \boldsymbol{\eta}_t^{\text{bias}}, \quad \overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}} := \frac{1}{N}\sum_{t=s}^{s+N-1} \boldsymbol{\eta}_t^{\text{var}}. \tag{C.4}$$

The excess risk can be decomposed into bias and variance:

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) = \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}_{s:s+N} \otimes \overline{\boldsymbol{\eta}}_{s:s+N}] \right\rangle \le 2 \cdot \text{Bias} + 2 \cdot \text{Variance},$$

where

$$\text{Bias} := \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}^{\text{bias}}_{s:s+N} \otimes \overline{\boldsymbol{\eta}}^{\text{bias}}_{s:s+N}] \right\rangle, \ \text{Variance} := \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}^{\text{var}}_{s:s+N} \otimes \overline{\boldsymbol{\eta}}^{\text{var}}_{s:s+N}] \right\rangle.$$

Define the covariance matrices $\mathbf{B}_t := \mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}]$ and $\mathbf{C}_t := \mathbb{E}[\boldsymbol{\eta}_t^{\text{var}} \otimes \boldsymbol{\eta}_t^{\text{var}}]$. The recursive forms of $\mathbf{B}_t$ and $\mathbf{C}_t$ then satisfy

$$\mathbf{B}_t = \mathcal{B} \circ \mathbf{B}_{t-1}, \quad \mathbf{B}_0 = \boldsymbol{\eta}_0 \otimes \boldsymbol{\eta}_0; \tag{C.5}$$

$$\mathbf{C}_t = \mathcal{B} \circ \mathbf{C}_{t-1} + \widehat{\boldsymbol{\Sigma}}, \quad \mathbf{C}_0 = \mathbf{0}. \tag{C.6}$$

Appendix D. Parameter Choice

D.1. Derivation of Parameter Choice

Following the optimization literature [12], we first fix the relationship between α and β as

$$\alpha = \frac{1}{1+\beta}. \tag{D.1}$$

We then fix

$$\delta = \psi\widetilde{\kappa}\beta\gamma, \tag{D.2}$$

following [8]. We remark that introducing $\widetilde{\kappa}$ prevents the effect of fourth moment from blowing up (see proof of Lemma H.5). Furthermore, we require $\gamma \geq \delta$ to enforce acceleration. Then, from the requirement $\psi l < 1$, we require

$$\frac{\delta\psi \operatorname{tr}(\mathbf{H})}{2} + \frac{1}{2} + \frac{\psi\gamma}{4}\sum_{i>\widetilde{\kappa}} \lambda_i < 1.$$

Therefore, it suffices to take

$$\delta \leq \frac{1}{2\psi \operatorname{tr}(\mathbf{H})}, \quad \gamma \leq \frac{1}{2\psi \sum_{i>\widetilde{\kappa}} \lambda_i}. \tag{D.3}$$

Combining (D.1), (D.2) and (D.3), we derive the choice of parameters in (3.1).

We remark that we get rid of dimension dependency by merit of the term $\psi\gamma/4 \cdot \sum_{i>\widetilde{\kappa}} \lambda_i$. Without this term, $\widetilde{\kappa}$ should be chosen as the model dimension d (as in Jain et al. [8]).

D.2. Discussion of Parameters

In the parameter choice (3.1), $\widetilde{\kappa}$ is a free parameter. In this section, we discuss how the choice of $\widetilde{\kappa}$ affects the excess risk bound. Suppose that both equalities are attained in (D.3). We focus the impact of $\widetilde{\kappa}$ on (i) eigenvalue cutoff $\widehat{k}$, and (ii) bias decay rate.

Note that

$$\gamma = \frac{1}{2\psi \sum_{i>\widetilde{\kappa}} \lambda_i},$$

so γ increases as κ increases. Furthermore,

$$\beta = \frac{\delta}{\psi\widetilde{\kappa}\gamma},$$

so β decreases as $\widetilde{\kappa}$ increases. We also have

$$c = \alpha(1-\beta) = \frac{1-\beta}{1+\beta},$$

so c increases as $\widetilde{\kappa}$ increases.

$\widehat{k}$ is defined as $\widehat{k} = \max\{k : \lambda_k \geq (1-c)/\delta\}$, so $\widehat{k}$ increases as $\widetilde{\kappa}$ increases; The bias decay rate in the subspace of the smallest eigenvalues (i.e., $i > k^\dagger$) is $1-(\gamma+\delta)\lambda_i/2$, so the decay rate accelerates for larger $\widetilde{\kappa}$. However, for the subspace of λ_i where $k^\dagger < i \leq k^\ddagger$, the bias decay rate is $[c(1-\delta\lambda_i)]^{s/2}$, so the decay rate slows down for larger $\widetilde{\kappa}$.

Combining all the above, we conclude that the choice of $\widetilde{\kappa}$ is subject to the eigenvalue spectrum of the data covariance matrix. Additionally, choosing a small $\widetilde{\kappa}$ will make the algorithm perform more like SGD.

Appendix E. Implication for Stochastic Heavy Ball Method

In this section, we extend the results we obtained for ASGD to By taking $\delta = 0$ in (2.3) and eliminating $\mathbf{v}_t$ and $\mathbf{u}_t$ using (2.2) and (2.4), we get

$$\mathbf{w}_{t+1} = \mathbf{w}_t - (1-\alpha)\gamma \cdot \widehat{\nabla} L(\mathbf{w}_t) + \alpha(1-\beta)\cdot(\mathbf{w}_t - \mathbf{w}_{t-1}),$$

which is exactly the form of the stochastic heavy ball (SHB) update. Therefore, the excess risk bound we presented in Theorem 3.1 can be directly applied to SHB.

As there are three free parameters but only two combinations $(1-\alpha)\gamma$ and $\alpha(1-\beta)$ are used, we enforce that $\beta = (1-\alpha)/\alpha$ and define $c = \alpha(1-\beta)$ and $q = (1-\alpha)\gamma$, similar to ASGD. By (3.2) and the definition of $\widehat{k}$, we have $k^\ddagger = \widehat{k} = 0$. Therefore, the following corollary gives the excess risk bound of SHB:

Corollary E.1 *Consider stochastic heavy ball (SHB) method, given by the update rule*

$$\mathbf{w}_{t+1} = \mathbf{w}_t - q\widehat{\nabla} L(\mathbf{w}_t) + c(\mathbf{w}_t - \mathbf{w}_{t-1}),$$

where the hyperparameters satisfy $c \in (0, 1-2/N]$ and $q = (1-c)\gamma/2$ with

$$\gamma \in \left(0, \frac{4}{\psi\,\mathrm{tr}(\mathbf{H})}\right).$$

Define $r_{SHB} := (1-\psi\gamma\,\mathrm{tr}(\mathbf{H})/4)^{-1}$, $k^* := \max\{k : \lambda_k \geq 1/(\gamma N)\}$, and define $k^\dagger$ as in (3.2). Then we have the following upper bound for the excess risk:

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \leq 2\cdot \mathit{EffectiveVar} + 2\cdot \mathit{EffectiveBias},$$

where effective variance is bounded by

$$\begin{aligned}\mathit{EffectiveVar} \leq \sigma^2 r_{SHB}\left[\frac{27k^*}{2N} + 18(s+N)\gamma^2\sum_{i>k^*}\lambda_i^2\right] + \frac{\psi r_{SHB}}{N}\left[\frac{9k^*}{N} + 36N\gamma^2\sum_{i>k^*}\lambda_i^2\right]\\ \cdot\left[\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{0:k^\dagger}} + \frac{2}{\gamma}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{k^\dagger:k^*}} + 4(s+N)\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right],\end{aligned}$$

and effective bias is bounded by

$$\textit{EffectiveBias} \le c^s \cdot \left(4s^2 + \frac{100}{(1-c)^2}\right) \cdot \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{0:k^\dagger}}}{N^2} + \frac{18}{N^2\gamma^2} \left\| \left(\mathbf{I} - \frac{\gamma \mathbf{H}}{2}\right)^s (\mathbf{w}_0 - \mathbf{w}^*) \right\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}} + 18 \left\| \left(\mathbf{I} - \frac{\gamma \mathbf{H}}{2}\right)^s (\mathbf{w}_0 - \mathbf{w}^*) \right\|^2_{\mathbf{H}_{k^*:\infty}}.$$

Remark E.2 *In the eigen-subspace of λ_i, the exponential decay rate of effective bias of SHB is $\max(c^s, (1-\gamma\lambda_i)^{2s})$, which is never faster than that of SGD. This happens because for SHB, γ has to be smaller than that of ASGD to control the effect of stochastic gradient. We can thus demonstrate that ASGD is superior to SHB in terms of the exponitial decay rate of the bias error, which extends a similar result given by Kidambi et al. [9] to the instance-dependent case.*

Appendix F. Proof of Main Results

In this section we prove Theorems 3.1 and 4.1.

F.1. Proof of Theorem 3.1

We start with the basic bias-variance decomposition lemma.

Lemma F.1 (Bias-variance decomposition, Jain et al. [8]) *The excess risk can be decomposed into bias and variance as*

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) = \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}_{s:s+N} \otimes \overline{\boldsymbol{\eta}}_{s:s+N}] \right\rangle \le 2 \cdot \textit{Bias} + 2 \cdot \textit{Variance}, \quad \text{(F.1)}$$

where

$$\textit{Bias} := \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}^{\textit{bias}}_{s:s+N} \otimes \overline{\boldsymbol{\eta}}^{\textit{bias}}_{s:s+N}] \right\rangle$$
$$\textit{Variance} := \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}^{\textit{var}}_{s:s+N} \otimes \overline{\boldsymbol{\eta}}^{\textit{var}}_{s:s+N}] \right\rangle.$$

This indicates that the generalization error could be bounded respectively by analyzing the bias and variance. We then further decompose bias and variance.

Lemma F.2 *Bias and Variance can be decomposed as*

$$\textit{Variance} = \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_1 + \mathbf{M}_2 \right\rangle, \quad \textit{Bias} = \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_3 + \mathbf{M}_4 \right\rangle,$$

where

$$\mathbf{M}_1 := \frac{1}{N^2}\left[\sum_{k=0}^{N-1} \mathbf{A}^k\right] \mathbf{C}_s \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right]^\top, \quad \text{(F.2)}$$

$$\mathbf{M}_2 := \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right](\mathbf{C}_{s+t} - \widetilde{\mathcal{B}}\circ\mathbf{C}_{s+t-1})\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top, \tag{F.3}$$

$$\mathbf{M}_3 := \frac{1}{N^2}\left[\sum_{k=0}^{N-1}\mathbf{A}^k\right]\mathbf{B}_s\left[\sum_{k=0}^{N-1}\mathbf{A}^k\right]^\top, \tag{F.4}$$

$$\mathbf{M}_4 := \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right](\mathbf{B}_{s+t} - \widetilde{\mathcal{B}}\circ\mathbf{B}_{s+t-1})\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top. \tag{F.5}$$

Proof The proof largely follows Zou et al. [18]. From the definitions of $\boldsymbol{\eta}_t^{\text{bias}}$, we have the following

$$\mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}}|\boldsymbol{\eta}_{t-1}^{\text{bias}}] = \mathbb{E}[\widehat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\text{bias}}|\boldsymbol{\eta}_{t-1}^{\text{bias}}] = \mathbf{A}\boldsymbol{\eta}_{t-1}^{\text{bias}}, \tag{F.6}$$

and for $\boldsymbol{\eta}_t^{\text{variance}}$, we have

$$\mathbb{E}[\boldsymbol{\eta}_t^{\text{var}}|\boldsymbol{\eta}_{t-1}^{\text{var}}] = \mathbb{E}[\widehat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\text{var}} + \boldsymbol{\zeta}_t|\boldsymbol{\eta}_{t-1}^{\text{var}}] = \mathbf{A}\boldsymbol{\eta}_{t-1}^{\text{var}}. \tag{F.7}$$

Then, regarding the term $\mathbb{E}[\overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}}\otimes\overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}}]$, we have

$$\begin{aligned}
&\mathbb{E}[\overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}}\otimes\overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}}]\\
&= \frac{1}{N^2}\sum_{t=s}^{s+N-1}\left(\mathbb{E}[\boldsymbol{\eta}_t^{\text{var}}\otimes\boldsymbol{\eta}_t^{\text{var}}] + \sum_{k=t+1}^{s+N-1}\mathbb{E}[\boldsymbol{\eta}_k^{\text{var}}\otimes\boldsymbol{\eta}_t^{\text{var}}] + \sum_{k=t+1}^{s+N-1}\mathbb{E}[\boldsymbol{\eta}_t^{\text{var}}\otimes\boldsymbol{\eta}_k^{\text{var}}]\right)\\
&= \frac{1}{N^2}\sum_{t=s}^{s+N-1}\left[\mathbf{C}_t + \sum_{k=t+1}^{s+N-1}\mathbf{A}^{k-t}\mathbf{C}_t + \sum_{k=t+1}^{s+N-1}\mathbf{C}_t(\mathbf{A}^{k-t})^\top\right]\\
&= \frac{1}{N^2}\left[\sum_{k=0}^{N-1}\mathbf{A}^k\right]\mathbf{C}_s\left[\sum_{k=0}^{N-1}\mathbf{A}^k\right]^\top\\
&\quad + \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right](\mathbf{C}_{s+t} - \widetilde{\mathcal{B}}\circ\mathbf{C}_{s+t-1})\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top,
\end{aligned}$$

where the second equality holds by applying (F.7) $k-t$ times, and the last inequality holds due to Lemma M.4. The decomposition of bias into $\mathbf{M}_3$ and $\mathbf{M}_4$ can be proven in exactly the same manner. ■

From Lemma F.2, we can further bound the variance and bias terms as follows.

We have the following bound for variance, whose detailed proof can be found in Appendix I.

Lemma F.3 *Under Assumptions 2.1, 2.2 and 2.3, with our choice of parameters as in (3.1), we have*

$$\textit{Variance} \le \sigma^2 r\left[\frac{27k^*}{2N} + \frac{18(s+N)(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right].$$

where $k^* = \max\{k : \lambda_k \ge 2N(q-c\delta)/(1-c)\}$.

The following lemma provides an upper bound for the bias error, whose detailed proof can be found in Appendix J.

Lemma F.4 *Under Assumptions 2.1, 2.2 and 2.3, and with our choice of parameters as in (3.1), we have*

$$\textit{Bias} \leq \text{Effective Bias} + \frac{\psi r}{N}\left[\frac{9k^*}{N} + \frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\cdot\left[\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:\widehat{k}}}^2 + \frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{\widehat{k}:k^\dagger}}^2 + \frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{k^\dagger:k^*}}^2 + 4(s+N)\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\right],$$

where

$$\begin{aligned}\text{Effective Bias} \leq{}& \frac{8(c\delta/q)^{2s}}{N^2\delta^2}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{0:k^\ddagger}^{-1}}^2 + \frac{4s^2}{N^2}c^s\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|_{\mathbf{H}_{k^\ddagger:k^\dagger}}^2\\ &+ \frac{16c^s}{N^2\delta^2}\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|_{\mathbf{H}_{k^\ddagger:\widehat{k}}^{-1}}^2 + \frac{100c^s}{N^2(1-c)^2}\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|_{\mathbf{H}_{\widehat{k}:k^\dagger}}^2\\ &+ \frac{9(1-c)^2}{2N^2(q-c\delta)^2}\left\|\left(\mathbf{I}-\frac{q-c\delta}{1-c}\mathbf{H}\right)^s(\mathbf{w}_0-\mathbf{w}^*)\right\|_{\mathbf{H}_{k^\dagger:k^*}^{-1}}^2\\ &+ 18\left\|\left(\mathbf{I}-\frac{q-c\delta}{1-c}\mathbf{H}\right)^s(\mathbf{w}_0-\mathbf{w}^*)\right\|_{\mathbf{H}_{k^*:\infty}}^2.\end{aligned}$$

Substituting Lemma F.3 and Lemma F.4 into (F.1) in Lemma F.1 yields our final result presented in Theorem 3.1.

F.2. Proof of Theorem 4.1

We consider the linear regression instance where the samples $\mathbf{x}_t$ follow the Gaussian distribution $\mathcal{N}(\mathbf{0},\mathbf{H})$ where $\lambda_i = i^{-2}$, so $\psi = 3$ in Assumption 2.2. The hyperparameters of ASGD are chosen as $\delta = 0.1$, $\alpha = 0.9875$, $\beta = (1-\alpha)/\alpha$, $\widetilde{\kappa} = 5$, $\gamma = \delta/(\psi\widetilde{\kappa}\beta) = 79/150$ and $N = 500$. Finally, we require $(\mathbf{w}_0-\mathbf{w}^*)_i = 0$ for $i \geq 8$.

We now present a formal expression of Theorem 4.1:

Theorem F.5 (Restatement of Theorem 4.1) *When applied to the aforementioned class of problem instances and initialization such that $\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}}^2 = \mathcal{O}(\sigma^2)$, the excess risk of SGD satisfies*

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N}^{SGD})] - L(\mathbf{w}^*) = \Omega(\sigma^2(N^{-1/2} + N^{-2}\cdot 0.996^s)),$$

and the excess risk of ASGD satisfies

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) = \mathcal{O}(\sigma^2(N^{-1/2} + N^{-2}\cdot 0.9873^s)).$$

Proof We first recall the excess risk lower bound for SGD given by Theorem 5.2 of Zou et al. [18]:

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N}^{\text{SGD}})] - L(\mathbf{w}^*) \geq \underbrace{\frac{\sigma^2}{600}\left[\frac{k_{\text{SGD}}^*}{N} + (s+N)\delta^2\sum_{i>k_{\text{SGD}}^*}\lambda_i^2\right]}_{\text{Variance}}$$

$$+ \underbrace{\frac{1}{100\delta^2 N^2} \cdot \|(\mathbf{I} - \delta\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \frac{1}{100} \cdot \|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}_{k^*:\infty}}}_{\text{EffectiveBias}},$$

As $c = 2\alpha - 1$ and $q = \alpha\delta + (1-\alpha)\gamma$, we have $c = 0.975$ and $q = 79/750$. By definition of $\widehat{k}$, $k^\ddagger$, $k^\dagger$ in (3.2), we have

$$k^\ddagger = 0, \ \widehat{k} = 2, \ k^\dagger = 6.$$

For the EffectiveBias term, note that all coefficients are absolute constants, so it suffices to consider the exponential decay rate in the eigen-subspace of λ_7. For SGD, the exponential decay rate is $(1 - \delta\lambda_i) = 0.996^s$, and for ASGD, the exponential decay rate is $(1 - (\gamma + \delta)\lambda_i/2)^s = 0.9873^s$. ■

Appendix G. Properties of $\mathbf{A}_i$

G.1. Segmentation of Eigen-subspaces

Recall that $\mathbf{A}_i$ is defined as

$$\mathbf{A}_i := \begin{bmatrix} 0 & 1 - \delta\lambda_i \\ -c & 1 + c - q\lambda_i \end{bmatrix}, \tag{G.1}$$

so the eigenvalues of $\mathbf{A}_i$ are

$$x_1 = \frac{1 + c - q\lambda_i}{2} - \frac{\sqrt{(1 + c - q\lambda_i)^2 - 4c(1 - \delta\lambda_i)}}{2}, \tag{G.2}$$

$$x_2 = \frac{1 + c - q\lambda_i}{2} + \frac{\sqrt{(1 + c - q\lambda_i)^2 - 4c(1 - \delta\lambda_i)}}{2}. \tag{G.3}$$

From (G.2) and (G.3), we see that whether $\mathbf{A}_i$ has complex or real eigenvalues depends on whether the following holds:

$$(1 + c - q\lambda_i)^2 - 4c(1 - \delta\lambda_i) < 0. \tag{G.4}$$

Directly solving (G.4), we have

$$(\sqrt{q - c\delta} - \sqrt{c(q - \delta)})^2/q^2 < \lambda_i < (\sqrt{q - c\delta} + \sqrt{c(q - \delta)})^2/q^2.$$

Define the eigenvalue cutoffs as

$$k^\dagger := \max\{i : \lambda_i > (\sqrt{q - c\delta} - \sqrt{c(q - \delta)})^2/q^2\}, \tag{G.5}$$

$$k^\ddagger := \max\{i : \lambda_i \geq (\sqrt{q - c\delta} + \sqrt{c(q - \delta)})^2/q^2\}, \tag{G.6}$$

and we note that

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

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

Thus, if $i \le k^\ddagger$ or $i > k^\dagger$, then $\mathbf{A}_i$ has real eigenvalues; If $k^\ddagger < i \le k^\dagger$, then $\mathbf{A}_i$ has complex eigenvalues. We also define two other important eigenvalue cutoffs

$$\widehat{k} := \max\{i : \lambda_i \ge (1-c)/\delta\} \tag{G.7}$$

and

$$k^* := \max\left\{i : \lambda_i \ge \frac{1-c}{2N(q-c\delta)}\right\}.$$

We have the following lemma concerning the cutoff of eigenvalues:

Lemma G.1 *Let $k^\dagger$ and $k^\ddagger$ be defined in (G.5) and (G.6). Then we have*

- *For all $i > k^\dagger$, we have*
$$\lambda_i \le \frac{1-c}{q} \le \frac{1-c}{\delta};$$
- *For all $i \le k^\ddagger$, we have*
$$\lambda_i \ge \frac{1-c}{\delta}.$$

Proof For all $i > k^\dagger$, according to (G.5), we have

$$\lambda_i \le \frac{1-c}{q} \cdot \frac{\sqrt{q-c\delta} - \sqrt{c(q-\delta)}}{\sqrt{q-c\delta} + \sqrt{c(q-\delta)}} \le \frac{1-c}{q} \le \frac{1-c}{\delta},$$

where the second inequality holds because $\frac{\sqrt{q-c\delta} - \sqrt{c(q-\delta)}}{\sqrt{q-c\delta} + \sqrt{c(q-\delta)}} \le 1$, and the last inequality holds because $q \ge \delta$.

For all $i > k^\ddagger$, we have

$$\begin{aligned}
\lambda_i - \frac{1-c}{\delta} &\ge \frac{1-c}{q} \cdot \frac{\sqrt{q-c\delta} + \sqrt{c(q-\delta)}}{\sqrt{q-c\delta} - \sqrt{c(q-\delta)}} - \frac{1-c}{\delta} \\
&= (1-c)\sqrt{c(q-\delta)} \cdot \frac{(q+\delta) - \sqrt{(q-\delta)(q-c\delta)/c}}{\delta q(\sqrt{q-c\delta} - \sqrt{c(q-\delta)})} \\
&\ge \frac{(1-c)\sqrt{c(q-\delta)}}{\delta q(\sqrt{q-c\delta} - \sqrt{c(q-\delta)})} \cdot [(q+\delta) - (q-c\delta)] \\
&= \frac{(1-c^2)\sqrt{c(q-\delta)}}{q(\sqrt{q-c\delta} - \sqrt{c(q-\delta)})} \ge 0,
\end{aligned}$$

where the first inequality holds due to (G.6), and the second inequality holds because $q - \delta \le c(q-c\delta)$. ■

With Lemma G.1, we immediately know that $k^\ddagger \le \widehat{k} \le k^\dagger$. If we also assume that $N(1-c) \ge 2$, then

$$\frac{1-c}{2N(q-c\delta)} \le \frac{(1-c)^2}{4(q-c\delta)} \le \frac{(1-c)^2}{(\sqrt{q-c\delta} + \sqrt{c(q-\delta)})^2},$$

where the inequality holds because $c(q-\delta) \le q - c\delta$. We thus have $k^* \ge k^\dagger$.

We then provide bounds for the spectral norm of $\mathbf{A}_i$. The bounds are accurate in the sense that when x_1, x_2 are real, the upper bound of $1 - x_2$ is at most the multiply of a constant of its lower bound.

Lemma G.2 *Let x_1 and x_2 be defined in (G.2) and (G.3). Then we have*

- *If $i \le k^\ddagger$, then x_1, x_2 are real, x_2 is an increasing function in λ_i, and*

$$\frac{c\delta - \sqrt{c(q-\delta)(q-c\delta)}}{q} \le x_2 \le \frac{c\delta}{q};$$

- *If $k^\ddagger < i \le k^\dagger$, then x_1, x_2 are complex, and*

$$|x_1| = |x_2| = \sqrt{c(1-\delta\lambda_i)};$$

- *If $k > k^\dagger$, then x_1, x_2 are real, and*

$$1 - \frac{\sqrt{q-c\delta}(\sqrt{q-c\delta} + \sqrt{c(q-\delta)})}{1-c}\lambda_i \le x_2 \le 1 - \frac{q-c\delta}{1-c}\lambda_i$$

Proof If $i \le k^\ddagger$, then by definition of x_2, we have

$$\begin{aligned}
c - x_2 &= \frac{q\lambda_i + c - 1 - \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}}{2} \\
&= \frac{2c(q-\delta)\lambda_i}{q\lambda_i + c - 1 + \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}} \\
&= \frac{2c(q-\delta)}{q} \cdot \frac{1}{1 - \frac{1-c}{q\lambda_i} + \sqrt{\left(1 - \frac{1-c}{q\lambda_i} \cdot \frac{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}}{\sqrt{q-c\delta}-\sqrt{c(q-\delta)}}\right)\left(1 - \frac{1-c}{q\lambda_i} \cdot \frac{\sqrt{q-c\delta}-\sqrt{c(q-\delta)}}{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}}\right)}}
\end{aligned}$$

Note that the denominator is decreasing as a function of $(1-c)/(q\lambda_i)$, so we have

$$\begin{aligned}
&1 - \frac{1-c}{q\lambda_i} + \sqrt{\left(1 - \frac{1-c}{q\lambda_i} \cdot \frac{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}}{\sqrt{q-c\delta}-\sqrt{c(q-\delta)}}\right)\left(1 - \frac{1-c}{q\lambda_i} \cdot \frac{\sqrt{q-c\delta}-\sqrt{c(q-\delta)}}{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}}\right)} \\
&\le 1 - 0 + 1 = 2;
\end{aligned}$$

we also have

$$\begin{aligned}
&1 - \frac{1-c}{q\lambda_i} + \sqrt{\left(1 - \frac{1-c}{q\lambda_i} \cdot \frac{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}}{\sqrt{q-c\delta}-\sqrt{c(q-\delta)}}\right)\left(1 - \frac{1-c}{q\lambda_i} \cdot \frac{\sqrt{q-c\delta}-\sqrt{c(q-\delta)}}{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}}\right)} \\
&\ge 1 - \frac{\sqrt{q-c\delta}-\sqrt{c(q-\delta)}}{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}} = \frac{2\sqrt{c(q-\delta)}}{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}}.
\end{aligned}$$

Therefore, we have

$$x_2 \le c - \frac{2c(q-\delta)}{2q} = \frac{c\delta}{q},$$

and

$$x_2 \ge c - \frac{2c(q-\delta)}{q} \cdot \frac{\sqrt{q-c\delta} + \sqrt{c(q-\delta)}}{2\sqrt{c(q-\delta)}} = \frac{c\delta - \sqrt{c(q-\delta)(q-c\delta)}}{q}$$

If $k^\ddagger < i \le k^\dagger$, then we have

$$x_1 x_2 = c(1-\delta\lambda_i),$$

where $x_1 = \bar{x}_2$. Thus, $|x_1| = |x_2| = \sqrt{c(1-\delta\lambda_i)}$.

If $i > k^\dagger$, then we have

$$\begin{aligned} 1 - x_2 &= \frac{1 - c + q\lambda_i - \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}}{2} \\ &= \frac{2(q-c\delta)\lambda_i}{1 - c + q\lambda_i + \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}}. \end{aligned} \tag{G.8}$$

Note that

$$\begin{aligned} &\frac{\partial}{\partial\lambda_i}\left(1 - c + q\lambda_i + \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}\right) = q + \frac{2c\delta - q(1+c-q\lambda_i)}{\sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}} \\ &= \frac{-4c(q-\delta)(q-c\delta)}{\sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}\{q\sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)} + [(1+c)q - 2c\delta - q^2\lambda_i]\}}. \end{aligned}$$

We also note that

$$\begin{aligned} &q\sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)} + [(1+c)q - 2c\delta - q^2\lambda_i] \\ &\ge (1+c)q - 2c\delta - q^2\lambda_i \\ &\ge (1+c)q - 2c\delta - (\sqrt{c(q-\delta)} - \sqrt{c(q-\delta)})^2 \\ &= 2\sqrt{c(q-\delta)(q-c\delta)} \ge 0, \end{aligned}$$

where the first inequality holds because $\sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)} \ge 0$, the second inequality holds because due to (G.5). Therefore, we have

$$\frac{\partial}{\partial\lambda_i}\left(1 - c + q\lambda_i + \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}\right) \le 0,$$

so $1 - c + q\lambda_i + \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}$ is a function decreasing in λ_i. We thus have

$$1 - c + q\lambda_i + \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)} \le 1 - c + \sqrt{(1+c)^2 - 4c} = 2(1-c),$$

and

$$\begin{aligned} &1 - c + q\lambda_i + \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)} \\ &\ge 1 - c + (1-c)\frac{\sqrt{q-c\delta} - \sqrt{c(q-\delta)}}{\sqrt{q-c\delta} + \sqrt{c(q-\delta)}} = \frac{2(1-c)\sqrt{q-c\delta}}{\sqrt{q-c\delta} + \sqrt{c(q-\delta)}}. \end{aligned}$$

Therefore, we have

$$x_2 \leq 1 - \frac{2(q-c\delta)\lambda_i}{2(1-c)} = 1 - \frac{q-c\delta}{1-c}\lambda_i,$$

and

$$x_2 \geq 1 - \frac{2(q-c\delta)\lambda_i}{\frac{2(1-c)\sqrt{q-c\delta}}{\sqrt{q-c\delta}+\sqrt{c(q-\delta)}}} = 1 - \frac{\sqrt{q-c\delta}(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})}{1-c}\lambda_i.$$

■

G.2. Characterization of $\mathbf{A}_i^k$

Before we prove the upper bound for variance and bias, we first characterize the property of $\mathbf{A}_i^k$ for $k \geq 1$ and $i \in [1, d]$, i.e., each block of matrix $\mathbf{A}$ corresponding to each eigenvalue λ_i of $\mathbf{H}$.

Lemma G.3 *Let $\mathbf{A}_i$ be defined as in (G.1), write $\mathbf{A}_i^k$ as*

$$\mathbf{A}_i^k = \begin{bmatrix} (\mathbf{A}_i^k)_{11} & (\mathbf{A}_i^k)_{12} \\ (\mathbf{A}_i^k)_{21} & (\mathbf{A}_i^k)_{22} \end{bmatrix}.$$

Let the eigenvalues of $\mathbf{A}_i$ be x_1 and x_2 as defined in (G.2) and (G.3). Then, for any integer $k \geq 1$, we have

$$(\mathbf{A}_i^k)_{11} = -c(1-\delta\lambda_i)\frac{x_2^{k-1} - x_1^{k-1}}{x_2 - x_1},$$
$$(\mathbf{A}_i^k)_{12} = (1-\delta\lambda_i)\frac{x_2^{k} - x_1^{k}}{x_2 - x_1},$$
$$(\mathbf{A}_i^k)_{21} = -c\frac{x_2^{k} - x_1^{k}}{x_2 - x_1},$$
$$(\mathbf{A}_i^k)_{22} = \frac{x_2^{k+1} - x_1^{k+1}}{x_2 - x_1}.$$

Proof We prove Lemma G.3 by induction. For $k = 1$, we trivially have

$$-c(1-\delta\lambda_i)\frac{x_2^0 - x_1^0}{x_2 - x_1} = 0, \quad (1-\delta\lambda_i)\frac{x_2^1 - x_1^1}{x_2 - x_1} = 1 - \delta\lambda_i, \quad -c\frac{x_2^1 - x_1^1}{x_2 - x_1} = -c.$$

We also have

$$\frac{x_2^2 - x_1^2}{x_2 - x_1} = x_1 + x_2 = 1 + c - q\lambda_i.$$

Therefore, Lemma G.3 holds for $k = 1$. Suppose that the lemma holds for k. Note that $\mathbf{A}_i^{k+1} = \mathbf{A}_i \cdot \mathbf{A}_i^k$, so by induction hypothesis, we have

$$(\mathbf{A}_i^{k+1})_{11} = (1-\delta\lambda_i)(\mathbf{A}_i^k)_{21} = -c(1-\delta\lambda_i)\frac{x_2^k - x_1^k}{x_2 - x_1},$$
$$(\mathbf{A}_i^{k+1})_{12} = (1-\delta\lambda_i)(\mathbf{A}_i^k)_{22} = (1-\delta\lambda_i)\frac{x_2^{k+1} - x_1^{k+1}}{x_2 - x_1},$$

$$
\begin{aligned}
(\mathbf{A}_i^{k+1})_{21} &= -c(\mathbf{A}_i^k)_{11} + (1+c-q\lambda_i)(\mathbf{A}_i^k)_{21} \\
&= c^2(1-\delta\lambda_i)\frac{x_2^{k-1}-x_1^{k-1}}{x_2-x_1} - c(1+c-q\lambda_i)\frac{x_2^k-x_1^k}{x_2-x_1} \\
&= c\cdot x_1x_2\cdot\frac{x_2^{k-1}-x_1^{k-1}}{x_2-x_1} - c(x_1+x_2)\frac{x_2^k-x_1^k}{x_2-x_1} \\
&= -c\frac{x_2^{k+1}-x_1^{k+1}}{x_2-x_1}, \\
(\mathbf{A}_i^{k+1})_{22} &= -c(\mathbf{A}_i^k)_{12} + (1+c-q\lambda_i)(\mathbf{A}_i^k)_{22} \\
&= -c(1-\delta\lambda_i)\frac{x_2^k-x_1^k}{x_2-x_1} + (1+c-q\lambda_i)\frac{x_2^{k+1}-x_1^{k+1}}{x_2-x_1} \\
&= -x_1x_2\cdot\frac{x_2^k-x_1^k}{x_2-x_1} + (x_1+x_2)\cdot\frac{x_2^{k+1}-x_1^{k+1}}{x_2-x_1} \\
&= \frac{x_2^{k+2}-x_1^{k+2}}{x_2-x_1},
\end{aligned}
$$

where we used the property that $x_1+x_2 = 1+c-q\lambda_i$ and $x_1x_2 = c(1-\delta\lambda_i)$. Therefore, Lemma G.3 holds for $k+1$, and induction is completed. ■

Appendix H. Linear Operators and Effect of Fourth Moment

H.1. Properties of Linear Operators

In this section, we introduce linear operators on matrices as well as their properties. We first give the following definitions of linear operators:

$$
\begin{gathered}
\mathcal{I} \coloneqq \mathbf{I}\otimes\mathbf{I}, \quad \mathcal{M} \coloneqq \mathbb{E}[\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}], \quad \widetilde{\mathcal{M}} \coloneqq \mathbf{H}\otimes\mathbf{H}, \\
\mathcal{B} \coloneqq \mathbb{E}[\widehat{\mathbf{A}}_t\otimes\widehat{\mathbf{A}}_t], \quad \widetilde{\mathcal{B}} \coloneqq \mathbf{A}\otimes\mathbf{A}.
\end{gathered}
\tag{H.1}
$$

$\widehat{\mathbf{A}}_t$ can be defined as the sum of deterministic component $\mathbf{V}_1$ and stochastic component $\widehat{\mathbf{V}}_2$:

$$
\mathbf{V}_1 = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -c\mathbf{I} & (1+c)\mathbf{I} \end{bmatrix}, \quad \widehat{\mathbf{V}}_2 = \begin{bmatrix} \mathbf{0} & -\delta\mathbf{x}_t\mathbf{x}_t^\top \\ \mathbf{0} & -q\mathbf{x}_t\mathbf{x}_t^\top \end{bmatrix}. \tag{H.2}
$$

Define

$$
\mathbf{V}_2 \coloneqq \mathbb{E}[\widehat{\mathbf{V}}_2] = \begin{bmatrix} \mathbf{0} & -\delta\mathbf{H} \\ \mathbf{0} & -q\mathbf{H} \end{bmatrix}, \tag{H.3}
$$

then $\mathbf{A} = \mathbf{V}_1 + \mathbf{V}_2$. We are also interested in linear operators $\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]$ and $\mathbf{V}_2\otimes\mathbf{V}_2$.

We introduce the concept of PSD operators:

Definition H.1 (PSD operator) *An operator $\mathcal{O}$ defined on symmetric matrices is called a PSD operator if $\mathbf{M}\succeq\mathbf{0}$ implies $\mathcal{O}\circ\mathbf{M}\succeq\mathbf{0}$.*

The following lemma summarizes some basic properties of the linear operators.

Lemma H.2 *The operators defined in (H.1) have the following properties:*

(a) *$\mathcal{M}$, $\widetilde{\mathcal{M}}$, and $\mathcal{M} - \widetilde{\mathcal{M}}$ are PSD operators.*

(b) *For any PSD matrix $\mathbf{M} \in \mathbb{R}^{2d\times 2d}$, let*

$$\mathbf{M} := \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}, \tag{H.4}$$

where $\mathbf{M}_{11}, \mathbf{M}_{12}, \mathbf{M}_{21}$ and $\mathbf{M}_{22}$ are d-by-d blocks. We then have

$$\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{M} = \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\mathcal{M} \circ \mathbf{M}_{22}),$$
$$(\mathbf{V}_2 \otimes \mathbf{V}_2) \circ \mathbf{M} = \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\widetilde{\mathcal{M}} \circ \mathbf{M}_{22}).$$

Thus, $\mathbb{E}\big[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2\big]$ and $\mathbf{V}_2 \otimes \mathbf{V}_2$ are both PSD operators.

(c) *$\mathcal{B}$ and $\widetilde{\mathcal{B}}$ are both PSD operators.*

(d) *$\mathcal{B} - \widetilde{\mathcal{B}} = \mathbb{E}\big[\widehat{\mathbf{V}}_2 \otimes \mathbf{V}_2\big] - \mathbf{V}_2 \otimes \mathbf{V}_2$ is a PSD operator.*

Proof The proof follows those of Jain et al. [8], Zou et al. [18], and Wu et al. [17].

(a) For any PSD matrix $\mathbf{M}$, we have

$$\mathcal{M} \circ \mathbf{M} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{M}\mathbf{x}\mathbf{x}^\top] = \mathbb{E}[(\mathbf{x}^\top \mathbf{M}\mathbf{x})\mathbf{x}\mathbf{x}^\top] \succeq \mathbf{0},$$

where the inequality holds because $\mathbf{x}^\top\mathbf{M}\mathbf{x} \geq 0$ and $\mathbf{x}\mathbf{x}^\top \succeq \mathbf{0}$. Furthermore,

$$\widetilde{\mathcal{M}} \circ \mathbf{M} = \mathbf{HMH} \succeq \mathbf{0},$$

where the inequality holds because $\mathbf{M} \succeq \mathbf{0}$ and $\mathbf{H}$ is symmetric. Lastly,

$$(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{M} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{M}\mathbf{x}\mathbf{x}^\top] - \mathbf{HMH} = \mathbb{E}[(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{M}(\mathbf{x}\mathbf{x}^\top - \mathbf{H})] \succeq \mathbf{0},$$

where the inequality holds because $\mathbf{M} \succeq \mathbf{0}$ and $\mathbf{x}\mathbf{x}^\top - \mathbf{H}$ is symmetric.

(b) Note that $\mathbf{M}_{22} \succeq \mathbf{0}$ because $\mathbf{M} \succeq \mathbf{0}$. We thus have

$$\begin{aligned}
\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{M} &= \mathbb{E}\left[\widehat{\mathbf{V}}_2 \mathbf{M} \widehat{\mathbf{V}}_2^\top\right] \\
&= \mathbb{E}\left[\begin{bmatrix} \mathbf{0} & -\delta\mathbf{x}_t\mathbf{x}_t^\top \\ \mathbf{0} & -q\mathbf{x}_t\mathbf{x}_t^\top \end{bmatrix} \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ -\delta\mathbf{x}_t\mathbf{x}_t^\top & -q\mathbf{x}_t\mathbf{x}_t^\top \end{bmatrix}\right] \\
&= \mathbb{E}\left[\begin{bmatrix} \delta^2\mathbf{x}_t\mathbf{x}_t^\top\mathbf{M}_{22}\mathbf{x}_t\mathbf{x}_t^\top & \delta q\mathbf{x}_t\mathbf{x}_t^\top\mathbf{M}_{22}\mathbf{x}_t\mathbf{x}_t^\top \\ \delta q\mathbf{x}_t\mathbf{x}_t^\top\mathbf{M}_{22}\mathbf{x}_t\mathbf{x}_t^\top & q^2\mathbf{x}_t\mathbf{x}_t^\top\mathbf{M}_{22}\mathbf{x}_t\mathbf{x}_t^\top \end{bmatrix}\right] \\
&= \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\mathbf{M}_{22}\mathbf{x}_t\mathbf{x}_t^\top\right] \\
&= \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\mathcal{M} \circ \mathbf{M}_{22}) \succeq \mathbf{0},
\end{aligned}$$

where the last inequality holds because $\mathbf{M}_{22} \succeq \mathbf{0}$, $\mathcal{M}$ is a PSD operator and $\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \succeq \mathbf{0}$. In a similar way,

$$
\begin{aligned}
(\mathbf{V}_2 \otimes \mathbf{V}_2) \circ \mathbf{M} &= \mathbf{V}_2 \mathbf{M} \mathbf{V}_2^\top \\
&= \begin{bmatrix} \mathbf{0} & -\delta\mathbf{H} \\ \mathbf{0} & -q\mathbf{H} \end{bmatrix} \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ -\delta\mathbf{H} & -q\mathbf{H} \end{bmatrix} \\
&= \begin{bmatrix} \delta^2 \mathbf{H}\mathbf{M}_{22}\mathbf{H} & \delta q \mathbf{H}\mathbf{M}_{22}\mathbf{H} \\ \delta q \mathbf{H}\mathbf{M}_{22}\mathbf{H} & q^2 \mathbf{H}\mathbf{M}_{22}\mathbf{H} \end{bmatrix} \\
&= \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}\mathbf{M}_{22}\mathbf{H} \\
&= \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\widetilde{\mathcal{M}} \circ \mathbf{M}_{22}) \succeq \mathbf{0},
\end{aligned}
$$

where the inequality holds because $\mathbf{M}_{22} \succeq \mathbf{0}$, $\widetilde{\mathcal{M}}$ is a PSD operator, and $\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \succeq \mathbf{0}$.

(c) We have

$$
\mathcal{B} \circ \mathbf{M} = \mathbb{E}[\widehat{\mathbf{A}}_t \mathbf{M} \widehat{\mathbf{A}}_t^\top], \quad \widetilde{\mathcal{B}} \circ \mathbf{M} = \mathbf{A}\mathbf{M}\mathbf{A}^\top,
$$

so both $\mathcal{B}$ and $\widetilde{\mathcal{B}}$ are PSD operators.

(d) Note that $\widehat{\mathbf{A}}_t = \mathbf{V}_1 + \widehat{\mathbf{V}}_2$, and $\mathbf{A} = \mathbf{V}_1 + \mathbf{V}_2$, so

$$
\begin{aligned}
(\mathcal{B} - \widetilde{\mathcal{B}}) \circ \mathbf{M} &= (\mathbb{E}[(\mathbf{V}_1 + \widehat{\mathbf{V}}_2) \otimes (\mathbf{V}_1 + \widehat{\mathbf{V}}_2)] - (\mathbf{V}_1 + \mathbf{V}_2) \otimes (\mathbf{V}_1 + \mathbf{V}_2)) \circ \mathbf{M} \\
&= (\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] - \mathbf{V}_2 \otimes \mathbf{V}_2) \circ \mathbf{M} \\
&= \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes ((\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{M}_{22}) \succeq \mathbf{0},
\end{aligned}
$$

where the second inequality holds because because $\mathbb{E}[\mathbf{V}_1 \otimes \widehat{\mathbf{V}}_2] = \mathbf{V}_1 \otimes \mathbf{V}_2$ and $\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \mathbf{V}_1] = \mathbf{V}_2 \otimes \mathbf{V}_1$, the third inequality follows from part (b), and the inequality holds because $\mathbf{M}_{22} \succeq \mathbf{0}$, $\mathcal{M} - \widetilde{\mathcal{M}}$ is a PSD operator, and $\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \succeq \mathbf{0}$.

■

H.2. Analysis of Fourth Moment

In this section, we study the difference of operators $\mathcal{B}$ and $\widetilde{\mathcal{B}}$ (due to the fourth moment) when they are operated on PSD matrix $\mathbf{M}$. Specifically, we are interested in bounding the inner product

$$
\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \sum_{j=0}^{t-1} \mathcal{B}^j \circ \mathbf{M} \right\rangle. \tag{H.5}
$$

The following lemma is the starting point of the analysis of fourth moment:

Lemma H.3 *For any PSD matrix* $\mathbf{M}$*, we have*

$$(\mathcal{B} - \widetilde{\mathcal{B}}) \circ \mathbf{M} \preceq \mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{M},$$

where

$$\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{M} \preceq \psi \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{M} \right\rangle \cdot \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}.$$

Proof By Lemma H.2(d), we have

$$(\mathcal{B} - \widetilde{\mathcal{B}}) \circ \mathbf{M} = \Big(\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] - \mathbf{V}_2 \otimes \mathbf{V}_2\Big) \circ \mathbf{M} \preceq \mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{M},$$

where the inequality holds due to Lemma H.2(b).

Let $\mathbf{M}_{22}$ be the matrix that contains the last d rows and d columns of $\mathbf{M}$. By definition of $\widehat{\mathbf{V}}_2$ in (H.2), we have

$$\begin{aligned} \mathbb{E}\left[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2\right] \circ \mathbf{M} &= \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\mathcal{M} \circ \mathbf{M}) \\ &\preceq \psi \operatorname{tr}(\mathbf{H}\mathbf{M}_{22}) \cdot \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H} \\ &= \psi \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{M} \right\rangle \cdot \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}, \end{aligned}$$

where first eqaulity holds due to Lemma H.2(b), and the inequality holds due to Assumption 2.2. ■

The operators $(\mathcal{I} - \mathcal{B})^{-1}$ and $(\mathcal{I} - \mathcal{B})^{-1}$ are of special interest in the analysis of fourth moment. We first show the existence $(\mathcal{I} - \widetilde{\mathcal{B}})^{-1}$.

Lemma H.4 *With the parameters in (3.1),* $(\mathcal{I} - \widetilde{\mathcal{B}})^{-1}$ *exists and is a PSD operator.*

Proof

It suffices to show that the property holds for any rank-one matrix $\mathbf{x}\mathbf{x}^\top$. We have

$$(\mathcal{I} - \widetilde{\mathcal{B}})^{-1} \circ (\mathbf{x}\mathbf{x}^\top) = \sum_{k=0}^{\infty} \widetilde{\mathcal{B}}^k \circ (\mathbf{x}\mathbf{x}^\top) = \sum_{k=0}^{\infty} \mathbf{A}^k(\mathbf{x}\mathbf{x}^\top)(\mathbf{A}^k)^\top = \sum_{k=0}^{\infty} (\mathbf{A}^k\mathbf{x})(\mathbf{A}^k\mathbf{x})^\top.$$

Thus, the ij-entry of $(\mathcal{I} - \mathcal{B})^{-1} \circ (\mathbf{x}\mathbf{x}^\top)$ is

$$\sum_{k=0}^{\infty} (\mathbf{A}^k\mathbf{x})_i (\mathbf{A}^k\mathbf{x})_j \leq \sum_{k=0}^{\infty} |(\mathbf{A}^k\mathbf{x})_i| \cdot |(\mathbf{A}^k\mathbf{x})_j| \infty.$$

The series converges because all eigenvalues of $\mathbf{A}$, i.e., eigenvalues of all $\mathbf{A}_i$, have magnitudes strictly smaller than 1. ■

We then define operator $\mathcal{T}$ as

$$\mathcal{T} := \mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1 = \mathcal{I} - \widetilde{\mathcal{B}} + \mathbf{V}_2 \otimes \mathbf{V}_2. \tag{H.6}$$

Since $\mathcal{I} - \widetilde{\mathcal{B}}$ is invertible and $(\mathcal{I} - \widetilde{\mathcal{B}})^{-1}$ is a PSD operator, $\mathcal{T}$ is also invertible, and $\mathcal{T}^{-1}$ is a PSD operator. We can thus define matrix $\mathbf{U}$ as

$$\mathbf{U} := \mathcal{T}^{-1} \circ \left(\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H} \right). \tag{H.7}$$

The following lemma charanterizes a key property of $\mathbf{U}$:

Lemma H.5 (Modified from Jain et al. [8]) *With the choice of parameters in* (3.1)*, the inner product* $\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{U} \right\rangle$ *is upper bounded by* l*, where*

$$l := \frac{\delta \operatorname{tr}(\mathbf{H})}{2} + \frac{1}{2\psi} + \frac{\gamma}{4} \sum_{i > \widetilde{\kappa}} \lambda_i. \tag{H.8}$$

Specifically for SHB where $\delta = 0$*, we have*

$$\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{U} \right\rangle \le \frac{q \operatorname{tr}(\mathbf{H})}{2(1-c)}.$$

Proof Denote $\mathbf{U}_i \in \mathbb{R}^{2\times 2}$ as the i-th block of the block-diagonal matrix $\mathbf{U}$. By Equation (56) of Jain et al. [8], we have

$$(\mathbf{U}_i)_{22} = \frac{(1 + c - c\delta\lambda_i)(q - c\delta) + 2cq\delta\lambda_i}{2(1 - c^2 + c\lambda_i(q + c\delta))} = \frac{\delta}{2} + \frac{(1+c)(q-\delta)}{2(1 - c^2 + c\lambda_i(q + c\delta))}. \tag{H.9}$$

On the one hand, $(\mathbf{U}_i)_{22}$ is bounded by

$$\begin{aligned} (\mathbf{U}_i)_{22} &\le \frac{\delta}{2} + \frac{(1+c)(q-\delta)}{2((1-c^2)\delta\lambda_i + c\lambda_i(q + c\delta))} = \frac{\delta}{2} + \frac{(1+c)(q-\delta)}{2(cq+\delta)\lambda_i} \\ &\le \frac{\delta}{2} + \frac{(1+c)(q-\delta)}{2(1+c)\delta\lambda_i} = \frac{\delta}{2} + \frac{q-\delta}{1-c} \cdot \frac{1-c}{2\delta\lambda_i}, \\ &= \frac{\delta}{2} + \frac{\gamma - \delta}{2} \cdot \frac{2\alpha\beta}{2\delta\lambda_i} \le \frac{\delta}{2} + \frac{\gamma}{2} \cdot \frac{\beta}{\delta\lambda_i} = \frac{\delta}{2} + \frac{1}{2\psi\widetilde{\kappa}\lambda_i}, \end{aligned} \tag{H.10}$$

where the first inequality holds because $\delta\lambda_i \le 1$, and the second inequality holds because $q \ge \delta$, and the third inequality holds because $\gamma - \delta \le \gamma$ and $\alpha\beta \le \beta$. On the other hand, $(\mathbf{U}_i)_{22}$ can also be bounded by

$$(\mathbf{U}_i)_{22} \le \frac{\delta}{2} + \frac{(1+c)(q-\delta)}{2(1-c^2)} = \frac{\delta}{2} + \frac{q-\delta}{2(1-c)} = \frac{\delta}{2} + \frac{\gamma-\delta}{4} \le \frac{\delta}{2} + \frac{\gamma}{4}, \tag{H.11}$$

where the first inequality holds because $1 - c^2 + (q - c\delta)\lambda_i \ge 1 - c^2$, and the second inequality holds because $(\gamma - \delta)/4 \le \gamma/4$. Thus, we have

$$\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{U} \right\rangle = \sum_{i=1}^{d} \lambda_i (\mathbf{U}_i)_{22} \le \frac{\delta}{2} \sum_{i=1}^{d} \lambda_i + \sum_{i=1}^{\widetilde{\kappa}} \frac{1}{2\psi\widetilde{\kappa}} + \sum_{i > \widetilde{\kappa}} \frac{\gamma\lambda_i}{4} = \frac{\delta \operatorname{tr}(\mathbf{H})}{2} + \frac{1}{2\psi} + \frac{\gamma}{4} \sum_{i > \widetilde{\kappa}} \lambda_i,$$

where the inequality holds due to (H.10) for $i \le \widetilde{\kappa}$ and (H.11) for $i > \widetilde{\kappa}$.

Specifically for SHB, we have

$$(\mathbf{U}_i)_{22} = \frac{(1+c)q}{2((1-c^2)+cq\lambda_i)} \le \frac{(1+c)q}{2(1-c^2)} = \frac{q}{2(1-c)},$$

where the inequality holds because $2((1-c^2)+cq\lambda_i) \ge 2(1-c^2)$. We thus have

$$\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{U} \right\rangle = \sum_{i=1}^{d} \lambda_i (\mathbf{U}_i)_{22} \le \frac{q \operatorname{tr}(\mathbf{H})}{2(1-c)}.$$

■

The following lemma charaterizes $(\mathcal{I}-\mathcal{B})^{-1}$ in terms of $\mathcal{T}$ and $\widehat{\mathbf{V}}_2$:

Lemma H.6 *The operator $(\mathcal{I}-\mathcal{B})^{-1}$ can be written in the form of geometric series*

$$(\mathcal{I}-\mathcal{B})^{-1} = \sum_{k=0}^{\infty} (\mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2])^k \circ \mathcal{T}^{-1}.$$

Proof According to the definition of $\mathcal{B}$,

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

where the last equality holds because $\mathbb{E}[\widehat{\mathbf{V}}_2] = \mathbf{V}_2$. We thus have

$$\begin{aligned} (\mathcal{I}-\mathcal{B})^{-1} &= \left(\mathcal{T} - \mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2]\right)^{-1} \\ &= \left\{\mathcal{T}\left[\mathcal{I} - \mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2]\right]\right\}^{-1} \\ &= \left[\mathcal{I} - \mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2]\right]^{-1} \mathcal{T}^{-1} \\ &= \sum_{k=0}^{\infty} (\mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2])^k \circ \mathcal{T}^{-1}, \end{aligned}$$

where the last inequality holds due to geometric series of linear operators. ■

We now show that $(\mathcal{I}-\mathcal{B})^{-1}$ exists and is a PSD operator.

Lemma H.7 *With the parameters in (3.1), for any PSD matrix $\mathbf{M}$, $(\mathcal{I}-\mathcal{B})^{-1} \circ \mathbf{M}$ exists and is a PSD matrix. Moreover, if we define $\mathbf{Q} := \mathcal{T}^{-1} \circ \mathbf{M}$, then we have*

$$(\mathcal{I}-\mathcal{B})^{-1} \circ \mathbf{M} = \mathbf{Q} + \frac{\psi}{1-\psi l} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{Q} \right\rangle \cdot \mathbf{U}.$$

Proof With Lemma H.6, we have

$$(\mathcal{I}-\mathcal{B})^{-1}\circ\mathbf{M}=\sum_{k=0}^{\infty}(\mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2])^k\circ\mathbf{Q}.$$

Note that by Lemma H.3

$$\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]\circ\mathbf{Q}\preceq\psi\left\langle\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{H}\end{bmatrix},\mathbf{Q}\right\rangle\cdot\begin{bmatrix}\delta^2&\delta q\\\delta q&q^2\end{bmatrix}\otimes\mathbf{H},$$

and by definition of $\mathbf{U}$ in (H.7), we have

$$\mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]\circ\mathbf{Q}\preceq\psi\left\langle\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{H}\end{bmatrix},\mathbf{Q}\right\rangle\cdot\mathbf{U}.$$

Then, applying Lemma H.5 and the definition of $\mathbf{U}$ recursively, we have for all $k\geq 1$,

$$(\mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2])^k\circ\mathbf{Q}\preceq\psi^k l^{k-1}\left\langle\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{H}\end{bmatrix},\mathbf{Q}\right\rangle\cdot\mathbf{U}.\tag{H.12}$$

Summing (H.12), considering the special case of $k=0$, we have

$$(\mathcal{I}-\mathcal{B})^{-1}\circ\mathbf{M}\preceq\mathbf{Q}+\sum_{k=1}^{\infty}\psi^k l^{k-1}\left\langle\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{H}\end{bmatrix},\mathbf{Q}\right\rangle\cdot\mathbf{U}=\mathbf{Q}+\frac{\psi}{1-\psi l}\left\langle\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{H}\end{bmatrix},\mathbf{Q}\right\rangle\cdot\mathbf{U}.$$

Therefore, $(\mathcal{I}-\mathcal{B})^{-1}$ exists and is a PSD operator. ■

The following result shows that the inner product (H.5) is different by only a constant if all $\mathcal{B}$ operators are replaced with $\widetilde{\mathcal{B}}$.

Lemma H.8 *For any PSD matrix $\mathbf{M}\in\mathbb{R}^{2d\times 2d}$, define the partial sum*

$$\mathbf{R}_t=\sum_{k=0}^{t-1}\mathcal{B}^k\circ\mathbf{M}.$$

Then we have

$$\mathbf{R}_t\preceq\sum_{k=0}^{t-1}\widetilde{\mathcal{B}}^k\circ\mathbf{M}+\frac{\psi}{1-\psi l}\left\langle\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{H}\end{bmatrix},\sum_{k=0}^{t-1}\widetilde{\mathcal{B}}^k\circ\mathbf{M}\right\rangle\cdot\mathbf{U}$$

and

$$\left\langle\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{H}\end{bmatrix},\mathbf{R}_t\right\rangle\leq r\left\langle\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{H}\end{bmatrix},\sum_{j=0}^{t-1}\widetilde{\mathcal{B}}^j\circ\mathbf{M}\right\rangle,$$

where $r=(1-\psi l)^{-1}$.

Proof By definition of $\mathbf{R}_t$, we have

$$\mathbf{R}_t=(\mathcal{I}-\mathcal{B})^{-1}(\mathcal{I}-\mathcal{B}^t)\circ\mathbf{M}$$

$$
\begin{aligned}
&\preceq (\mathcal{I} - \mathcal{B})^{-1}(\mathcal{I} - \widetilde{\mathcal{B}}^t) \circ \mathbf{M} \\
&= (\mathcal{I} - \mathcal{B})^{-1}(\mathcal{I} - \widetilde{\mathcal{B}}) \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M},
\end{aligned} \tag{H.13}
$$

where the inequality holds because $\widetilde{\mathcal{B}} \preceq \mathcal{B}$. Note that by definition of $\widetilde{\mathcal{B}}$, we have

$$
\mathcal{I} - \widetilde{\mathcal{B}} = \mathcal{I} - (\mathbf{V}_1 + \mathbf{V}_2) \otimes (\mathbf{V}_1 + \mathbf{V}_2) \preceq \mathcal{I} - \mathbf{V}_1 \otimes \mathbf{V}_1 - \mathbf{V}_1 \otimes \mathbf{V}_2 - \mathbf{V}_2 \otimes \mathbf{V}_1 = \mathcal{T}, \tag{H.14}
$$

where the inequality holds because $\mathbf{V}_2 \otimes \mathbf{V}_2$ is a PSD operator. $\mathbf{R}_t$ can thus be further bounded as

$$
\begin{aligned}
\mathbf{R}_t &\preceq (\mathcal{I} - \mathcal{B})^{-1}\mathcal{T} \circ \left(\sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M} \right) \\
&\preceq \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M} + \frac{\psi}{1 - \psi l} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M} \right\rangle \cdot \mathbf{U},
\end{aligned}
$$

where the first inequality holds due to (H.14), and the second inequality holds due to Lemma H.6. Therefore, taking inner product with $\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}$, we have

$$
\begin{aligned}
&\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{R}_t \right\rangle \\
&\leq \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M} \right\rangle + \frac{\psi}{1 - \psi l} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M} \right\rangle \cdot \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{U} \right\rangle \\
&\leq \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M} \right\rangle + \frac{\psi l}{1 - \psi l} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M} \right\rangle \\
&= \frac{1}{1 - \psi l} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \mathbf{M} \right\rangle,
\end{aligned}
$$

where the second inequality holds due to Lemma H.5. ■

Appendix I. Variance Upper Bound

I.1. Proof of Lemma F.3

In this subsection, we prove Lemma F.3. We need the following lemmas. The first lemma characterizes the recursive formula of $\mathbf{C}_t$:

Lemma I.1 (Section E.2 of Jain et al. [8]) *Define*

$$
\widehat{\mathbf{\Sigma}} := \mathbb{E}[\boldsymbol{\zeta}_t \otimes \boldsymbol{\zeta}_t], \tag{I.1}
$$

then the covariance matrix $\mathbf{C}_t$ *satisfies*

$$
\mathbf{C}_t = \mathcal{B} \circ \mathbf{C}_{t-1} + \widehat{\mathbf{\Sigma}}.
$$

Combining Lemma I.1 with Lemma M.3, we immediate know that $\mathbf{C}_t$ is an increasing sequence with

$$\mathbf{C}_t = \sum_{k=0}^{t-1} \mathcal{B}^k \circ \widehat{\mathbf{\Sigma}}. \tag{I.2}$$

The following lemmas provide upper bounds for $\mathbf{M}_1$ and $\mathbf{M}_2$, respectively:

Lemma I.2 *With the choice of parameters as in (3.1), we have*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_1 \right\rangle \leq \sigma^2 r \left[\frac{9k^*}{N} + \frac{36N(q-c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2 \right].$$

Lemma I.3 *With the choice of parameters as in (3.1), we have*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_2 \right\rangle \leq \sigma^2 r \left[\frac{18k^*}{N} + \frac{36s(q-c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2 \right].$$

We now prove Lemma F.3.
Proof [Proof of Lemma F.3] By Lemma F.2,

$$\begin{aligned}
\text{Variance} &\leq \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_1 + \mathbf{M}_2 \right\rangle \\
&\leq \frac{\sigma^2 r}{2} \left[\frac{9k^*}{N} + \frac{36N(q-c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2 \right] + \frac{\sigma^2 r}{2} \left[\frac{18k^*}{N} + \frac{36s(q-c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2 \right] \\
&= \sigma^2 r \left[\frac{27k^*}{2N} + \frac{18(s+N)(q-c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2 \right],
\end{aligned}$$

where the second inequality holds due to Lemma I.2 and Lemma I.3. ■

We remark that due to Lemma M.1, we have $\frac{q-c\delta}{1-c} = \frac{\gamma+\delta}{2} \leq \gamma$. Additionally, the constants in this proof are smaller than those given in Theorem 3.1. Therefore, the variance bound in Theorem 3.1 can be fully covered by the result provided in this proof.

I.2. Proof of Lemma I.3

We start with an upper bound for $\widehat{\mathbf{\Sigma}}$:

Lemma I.4 *Let $\widehat{\mathbf{\Sigma}}$ be defined in (I.1). Then*

$$\widehat{\mathbf{\Sigma}} \preceq \sigma^2 \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}.$$

Proof By definition of $\widehat{\mathbf{\Sigma}}$ in (I.1), we have

$$\widehat{\mathbf{\Sigma}} = \mathbb{E}[\boldsymbol{\zeta}_t \otimes \boldsymbol{\zeta}_t] = \mathbb{E}\left[\begin{bmatrix} \delta^2 \cdot \epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^\top & \delta q \cdot \epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^\top \\ \delta q \cdot \epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^\top & q^2 \cdot \epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^\top \end{bmatrix} \right] = \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{\Sigma}. \tag{I.3}$$

By Assumption 2.3, we have $\sigma^2 = \|\mathbf{H}^{-1/2}\boldsymbol{\Sigma}\mathbf{H}^{-1/2}\|$, so $\mathbf{H}^{-1/2}\boldsymbol{\Sigma}\mathbf{H}^{-1/2} \preceq \sigma^2\mathbf{I}$, and

$$\boldsymbol{\Sigma} \preceq \sigma^2\mathbf{H}. \tag{I.4}$$

Combining (I.3) with (I.4), we complete the proof. ■

We then provide an upper bound for the limiting matrix $\mathbf{C}_\infty$.

Lemma I.5 *Let $\mathbf{C}_\infty$ be defined as*

$$\mathbf{C}_\infty := (\mathcal{I} - \mathcal{B})^{-1} \circ \widehat{\boldsymbol{\Sigma}} = \sum_{k=0}^{\infty} \mathcal{B}^k \circ \widehat{\boldsymbol{\Sigma}}, \tag{I.5}$$

Then

$$\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{C}_\infty \right\rangle \le \frac{\sigma^2 l}{1 - \psi l},$$

where l is defined in (H.8).

Proof By definition of $\mathbf{C}_\infty$, we have

$$\begin{aligned}
\mathbf{C}_\infty &= (\mathcal{I} - \mathcal{B})^{-1} \circ \widehat{\boldsymbol{\Sigma}} \preceq \sigma^2 (\mathcal{I} - \mathcal{B})^{-1} \circ \left(\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H} \right) \\
&\preceq \sigma^2 \left(\mathbf{U} + \frac{\psi}{1 - \psi l} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{U} \right\rangle \cdot \mathbf{U} \right) \\
&\preceq \sigma^2 \left(\mathbf{U} + \frac{\psi l}{1 - \psi l} \cdot \mathbf{U} \right) = \frac{\sigma^2}{1 - \psi l} \cdot \mathbf{U},
\end{aligned}$$

where the first inequality holds due to Lemma H.6, and the second inequality holds due to Lemma H.5. Therefore, the inner product is bounded by

$$\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{C}_\infty \right\rangle \le \frac{\sigma^2}{1 - \psi l} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{U} \right\rangle \le \frac{\sigma^2 l}{1 - \psi l},$$

where the second inequality holds due to Lemma H.5. ■

We now prove Lemma I.3. For the matrix $\mathbf{M}_2$, we have the following upper bound:

$$\begin{aligned}
\mathbf{M}_2 &= \frac{1}{N^2} \sum_{t=1}^{N-1} \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k \right] \left[(\mathcal{B} - \widetilde{\mathcal{B}}) \circ \mathbf{C}_{s+t-1} + \widehat{\boldsymbol{\Sigma}} \right] \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k \right]^\top \\
&\preceq \frac{1}{N^2} \sum_{t=1}^{N-1} \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k \right] \left[\left(\psi \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{C}_{s+t-1} \right\rangle + \sigma^2 \right) \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H} \right] \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k \right]^\top \\
&\preceq \frac{1}{N^2} \sum_{t=1}^{N-1} \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k \right] \left[\left(\psi \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{C}_\infty \right\rangle + \sigma^2 \right) \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H} \right] \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k \right]^\top
\end{aligned}$$

$$
\begin{aligned}
&\preceq \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left[\left(\frac{\sigma^2\psi l}{1-\psi l}+\sigma^2\right)\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\right]\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top\\
&= \frac{\sigma^2 r}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left(\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\right)\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top, \qquad \text{(I.6)}
\end{aligned}
$$

where the first equality holds due to Lemma I.1, the first inequality holds due to Lemma I.4 and Lemma H.3, the second inequality holds because $\mathbf{C}_t$ is increasing, and the third inequality holds due to Lemma I.5. As $\mathbf{A}$ is block diagonal and $\mathbf{H}$ is diagonal, plugging (I.6) into the inner product, we have

$$
\begin{aligned}
\left\langle\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \mathbf{M}_2\right\rangle &\le \frac{\sigma^2 r}{N^2}\sum_{i=1}^{d}\lambda_i^2\sum_{t=1}^{N-1}\left(\left[\sum_{k=0}^{N-t-1}\mathbf{A}_i^k\right]\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\\
&= \frac{\sigma^2 r}{N^2}\sum_{t=0}^{N-1}\sum_{i=1}^{d}\lambda_i^2\left(\left[\sum_{k=0}^{t-1}\mathbf{A}_i^k\right]\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\\
&\le \frac{\sigma^2 r}{N^2}\sum_{t=0}^{N-1}\left[9k^* + \frac{36N^2(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\\
&= \sigma^2 r\left[\frac{9k^*}{N} + \frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right],
\end{aligned}
$$

where the second inequality holds due to Corollary M.7.

I.3. Proof of Lemma I.2

The following lemma provides an upper bound on $\mathbf{C}_t$ by its update rule.

Lemma I.6 *For any $t > 0$, $\mathbf{C}_t$ can be upper bounded by*

$$
\mathbf{C}_t \preceq \sigma^2 r\sum_{k=0}^{t-1}\widetilde{\mathcal{B}}^k\circ\left(\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\right).
$$

Proof By the recursive formula (C.6), we have the following,

$$
\begin{aligned}
\mathbf{C}_t &= \mathcal{B}\circ\mathbf{C}_{t-1} + \widehat{\mathbf{\Sigma}} = \widetilde{\mathcal{B}}\circ\mathbf{C}_{t-1} + (\mathcal{B}-\widetilde{\mathcal{B}})\circ\mathbf{C}_{t-1} + \widehat{\mathbf{\Sigma}}\\
&\preceq \widetilde{\mathcal{B}}\circ\mathbf{C}_{t-1} + \psi\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{C}_{t-1}\right\rangle\cdot\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H} + \sigma^2\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\\
&\preceq \widetilde{\mathcal{B}}\circ\mathbf{C}_{t-1} + \psi\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{C}_{\infty}\right\rangle\cdot\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H} + \sigma^2\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\\
&\preceq \widetilde{\mathcal{B}}\circ\mathbf{C}_{t-1} + \frac{\sigma^2\psi l}{1-\psi l}\cdot\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H} + \sigma^2\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\\
&= \widetilde{\mathcal{B}}\circ\mathbf{C}_{t-1} + \sigma^2 r\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}, \qquad \text{(I.7)}
\end{aligned}
$$

where the first inequality holds due to Lemma H.3 and Lemma I.4, the second inequality holds due to holds because $\mathbf{C}_t$ is increasing, and the last inequality holds due to Lemma I.5. Applying (I.7) recursively, we have for all $t > 0$,

$$\mathbf{C}_t \preceq \sigma^2 r \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}.$$

∎

We are now ready to prove Lemma I.2. With Lemma I.6, we have

$$\begin{aligned} \mathbf{M}_1 &= \frac{1}{N^2} \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right] \mathbf{C}_s \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right]^\top \\ &\preceq \frac{\sigma^2 r}{N^2} \sum_{j=0}^{s-1} \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right] \left[\widetilde{\mathcal{B}}^j \circ \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}\right] \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right]^\top \\ &= \frac{\sigma^2 r}{N^2} \sum_{j=0}^{s-1} \left[\sum_{k=0}^{N-1} \mathbf{A}^{j+k}\right] \left(\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}\right) \left[\sum_{k=0}^{N-1} \mathbf{A}^{j+k}\right]^\top. \end{aligned} \tag{I.8}$$

As $\mathbf{A}$ is block-diagonal and $\mathbf{H}$ is diagonal, plugging (I.8) into the inner product, we have

$$\begin{aligned} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_1 \right\rangle &\le \frac{\sigma^2 r}{N^2} \sum_{i=1}^{d} \lambda_i^2 \sum_{j=0}^{s-1} \left(\sum_{k=0}^{N-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1^2 \qquad \text{(I.9)} \\ &\le \frac{\sigma^2 r}{N^2} \left[18 N k^* + \frac{36 s N^2 (q - c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2\right] \\ &= \sigma^2 r \left[\frac{18 k^*}{N} + \frac{36 s (q - c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2\right], \end{aligned}$$

where the second inequality holds due to Corollary M.6.

Appendix J. Bias Upper Bound

J.1. Proof of Lemma F.4

In this subsection, we prove Lemma F.4. We first need the following lemma for $\mathbf{B}_t$:

Lemma J.1 *With $\mathbf{B}_t$ as defined in (C.5), we have*

$$\mathbf{B}_t = \mathcal{B} \circ \mathbf{B}_{t-1},$$

and

$$\mathbf{B}_t \preceq \widetilde{\mathcal{B}}^t \circ \mathbf{B}_0 + \psi \sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_{t-1-k} \right\rangle \cdot \widetilde{\mathcal{B}}^k \circ \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}.$$

We also have the following lemma for the partial sum of $\mathbf{B}_t$:

Lemma J.2 *Let* $\mathbf{B}_t$ *defined in* (C.5). *Then we have*

$$\sum_{k=0}^{t-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix},\mathbf{B}_k\right\rangle \le r\Bigg[\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:\widehat{k}}}+\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}} + \frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{k^\dagger:k^*}}+4t\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\Bigg].$$

We are now ready prove Lemma F.4.

Proof [Proof of Lemma F.4] By Lemma F.2, it suffices to bound the inner produce of $\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ with $\mathbf{M}_3$ and $\mathbf{M}_4$ separately. For $\mathbf{M}_3$, by Lemma J.1, we have

$$\begin{aligned}\mathbf{M}_3 &= \frac{1}{N^2}\left[\sum_{k=0}^{N-1}\mathbf{A}^k\right]\mathbf{B}_s\left[\sum_{k=0}^{N-1}\mathbf{A}^k\right]^\top\\ &\preceq \frac{1}{N^2}\left[\sum_{k=0}^{N-1}\mathbf{A}^{k+s}\right]\mathbf{B}_0\left[\sum_{k=0}^{N-1}\mathbf{A}^{k+s}\right]^\top\\ &\quad+\frac{\psi}{N^2}\sum_{t=0}^{s-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix},\mathbf{B}_{s-1-t}\right\rangle\left[\sum_{k=0}^{N-1}\mathbf{A}^{k+t}\right]\left(\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\right)\left[\sum_{k=0}^{N-1}\mathbf{A}^{k+t}\right]^\top. \end{aligned}\tag{J.1}$$

We also note that

$$\mathbf{B}_0=\begin{bmatrix}(\mathbf{w}_0-\mathbf{w}^*)(\mathbf{w}_0-\mathbf{w}^*)^\top & (\mathbf{w}_0-\mathbf{w}^*)(\mathbf{w}_0-\mathbf{w}^*)^\top\\ (\mathbf{w}_0-\mathbf{w}^*)(\mathbf{w}_0-\mathbf{w}^*)^\top & (\mathbf{w}_0-\mathbf{w}^*)(\mathbf{w}_0-\mathbf{w}^*)^\top\end{bmatrix}=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\otimes[(\mathbf{w}_0-\mathbf{w}^*)(\mathbf{w}_0-\mathbf{w}^*)^\top].\tag{J.2}$$

$\mathbf{H}$ is diagonal and $\mathbf{A}$ is block diagonal, so plugging (J.1) and (J.2) into the inner product, we have

$$\begin{aligned}\left\langle\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix},\mathbf{M}_3\right\rangle &\le \frac{1}{N^2}\sum_{i=1}^d\lambda_i w_i^2\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{k+s}\begin{bmatrix}1\\1\end{bmatrix}\right)_1^2\\ &\quad+\underbrace{\frac{\psi}{N^2}\sum_{t=0}^{s-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix},\mathbf{B}_{s-1-t}\right\rangle\cdot\sum_{i=1}^d\lambda_i^2\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{k+t}\begin{bmatrix}\delta\\q\end{bmatrix}\right)_1^2}_{\mathrm{K}}.\end{aligned}\tag{J.3}$$

By Corollary M.9, we have

$$\begin{aligned}&\text{Effective Bias} := \frac{1}{2N^2}\sum_{i=1}^d\lambda_i w_i^2\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{k+s}\begin{bmatrix}1\\1\end{bmatrix}\right)_1^2\\ &\le \frac{8(c\delta/q)^{2s}}{N^2\delta^2}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^\ddagger}}+\frac{4s^2}{N^2}c^s\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^\ddagger:k^\dagger}}\\ &\quad+\frac{16c^s}{N^2\delta^2}\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{k^\dagger:\widehat{k}}}+\frac{100c^s}{N^2(1-c)^2}\|(\mathbf{I}-\delta\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^\ddagger:\widehat{k}}}\end{aligned}$$

$$
+ \frac{9(1-c)^2}{2N^2(q-c\delta)^2}\left\|\left(\mathbf{I}-\frac{q-c\delta}{1-c}\mathbf{H}\right)^s(\mathbf{w}_0-\mathbf{w}^*)\right\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}}
+ 18\left\|\left(\mathbf{I}-\frac{q-c\delta}{1-c}\mathbf{H}\right)^s(\mathbf{w}_0-\mathbf{w}^*)\right\|^2_{\mathbf{H}_{k^*:\infty}}.
$$

K can be bounded as

$$
\begin{aligned}
\mathrm{K} &= \frac{\psi}{N^2}\sum_{t=0}^{s-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{B}_{s-1-t}\right\rangle\cdot\sum_{i=1}^{d}\lambda_i^2\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{k+t}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\\
&\le \frac{\psi}{N^2}\left[9k^*+\frac{36(q-c\delta)^2N^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\sum_{t=0}^{s-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{B}_{s-1-t}\right\rangle\\
&= \frac{\psi}{N^2}\left[9k^*+\frac{36(q-c\delta)^2N^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\sum_{t=0}^{s-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{B}_{t}\right\rangle\\
&\le \frac{\psi r}{N}\left[\frac{9k^*}{N}+\frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\cdot\left[\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:\widehat{k}}}\right.\\
&\quad\left.+\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}}+\frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{k^\dagger:k^*}}+4s\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right],
\end{aligned}
\tag{J.4}
$$

where the first inequality holds due to Corollary M.6, and the second inequality holds due to Lemma J.2.

For $\mathbf{M}_4$, we have

$$
\begin{aligned}
\mathbf{M}_4 &= \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]((\mathcal{B}-\widetilde{\mathcal{B}})\circ\mathbf{B}_{s+t-1})\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top\\
&\preceq \frac{\psi}{N^2}\sum_{t=1}^{N-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{B}_{s+t-1}\right\rangle\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left(\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\right)\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top,
\end{aligned}
\tag{J.5}
$$

where the first equality holds beause $\mathcal{B}\circ\mathbf{B}_{s+t-1}=\mathbf{B}_{s+t}$, and the inequality holds due to Lemma H.3. $\mathbf{H}$ is diagonal and $\mathbf{A}$ is block-diagonal, so plugging (J.5) into the inner product, we have

$$
\begin{aligned}
\left\langle\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \mathbf{M}_4\right\rangle &\le \frac{\psi}{N^2}\sum_{t=1}^{N-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{B}_{s+t-1}\right\rangle\cdot\sum_{i=1}^{d}\lambda_i^2\left(\sum_{k=0}^{N-t-1}\mathbf{A}_i^k\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\\
&= \frac{\psi}{N^2}\sum_{t=0}^{N-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{B}_{s+N-t-1}\right\rangle\sum_{i=1}^{d}\lambda_i^2\left(\sum_{k=0}^{t-1}\mathbf{A}_i^k\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\\
&\le \frac{\psi}{N}\left[\frac{9k^*}{N}+\frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\sum_{t=0}^{N-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{B}_{s+N-t-1}\right\rangle\\
&= \frac{\psi}{N}\left[\frac{9k^*}{N}+\frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\sum_{t=0}^{N-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}, \mathbf{B}_{s+t}\right\rangle,
\end{aligned}
\tag{J.6}
$$

where the second inequality holds due to Corollary M.7. Note that

$$
\begin{aligned}
\sum_{t=0}^{N-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix},\mathbf{B}_{s+t}\right\rangle &\leq \sum_{t=0}^{s+N-1}\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix},\mathbf{B}_{t}\right\rangle\\
&\leq r\Bigg[\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:\widehat{k}}}^2+\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{\widehat{k}:k^\dagger}}^2\\
&\quad+\frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{k^\dagger:k^*}}^2+4(s+N)\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\Bigg],
\end{aligned}
\tag{J.7}
$$

where the first inequality holds because $\mathbf{B}_t \succeq 0$, and the second inequality holds due to Lemma J.2. Plugging (J.7) into (J.6), combining the result with (J.4), we have

$$
\begin{aligned}
&\text{Bias} = \frac{1}{2}\left\langle\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix},\mathbf{M}_3+\mathbf{M}_4\right\rangle\\
&\leq \text{Effective Bias}+\frac{\psi r}{2N}\left[\frac{9k^*}{N}+\frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\cdot\Bigg[\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:\widehat{k}}}^2\\
&\quad+\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{\widehat{k}:k^\dagger}}^2+\frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{k^\dagger:k^*}}^2+4s\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\Bigg]\\
&\quad+\frac{\psi r}{2N}\left[\frac{9k^*}{N}+\frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\cdot\Bigg[\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:\widehat{k}}}^2+\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{\widehat{k}:k^\dagger}}^2\\
&\quad+\frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{k^\dagger:k^*}}^2+4(s+N)\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\Bigg]\\
&\leq \text{Effective Bias}+\frac{\psi r}{N}\left[\frac{9k^*}{N}+\frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right]\cdot\Bigg[\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:\widehat{k}}}^2\\
&\quad+\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{\widehat{k}:k^\dagger}}^2+\frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{k^\dagger:k^*}}^2+4(s+N)\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\Bigg],
\end{aligned}
$$

where the second inequality holds because $4s\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k:\infty}}^2 \leq 4(s+N)\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k:\infty}}^2$. ■

We remark that due to Lemma M.1, we have $\frac{q-c\delta}{1-c}=\frac{\gamma+\delta}{2}\leq\gamma$. Additionally, the constants in this proof are smaller than those given in Theorem 3.1. Therefore, the bias bound in Theorem 3.1 can be fully covered by the result provided in this proof.

J.2. Proof of Lemma J.1

The recursive formula $\mathbf{B}_t=\mathcal{B}\circ\mathbf{B}_{t-1}$ is proven in Section B.2 of Jain et al. [8]. We further have

$$
\begin{aligned}
\mathbf{B}_t &= \mathcal{B}\circ\mathbf{B}_{t-1}=\widetilde{\mathcal{B}}\circ\mathbf{B}_{t-1}+(\mathcal{B}-\widetilde{\mathcal{B}})\circ\mathbf{B}_{t-1}\\
&\preceq \widetilde{\mathcal{B}}\circ\mathbf{B}_{t-1}+\psi\left\langle\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix},\mathbf{B}_{t-1}\right\rangle\cdot\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}
\end{aligned}
$$

$$\preceq \widetilde{\mathcal{B}}^t \circ \mathbf{B}_0 + \psi \sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_k \right\rangle \cdot \widetilde{\mathcal{B}}^{t-1-k} \circ \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H}$$
$$= \widetilde{\mathcal{B}}^t \circ \mathbf{B}_0 + \psi \sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_{t-1-k} \right\rangle \cdot \widetilde{\mathcal{B}}^{k} \circ \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H},$$

where the first inequality holds due to Lemma H.3, and the second inequality holds by recursively applying the bound.

J.3. Proof of Lemma J.2

Note that $\mathbf{B}_t = \mathcal{B}^t \circ \mathbf{B}_0$ by (C.5). By Lemma H.8, we have

$$\sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_k \right\rangle \le r \sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \widetilde{\mathcal{B}}^k \circ \mathbf{B}_0 \right\rangle = r \sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{A}^k \mathbf{B}_0 (\mathbf{A}^k)^\top \right\rangle. \quad \text{(J.8)}$$

$\mathbf{H}$ is a diagonal matrix, and $\mathbf{A}$ is a block-diagonal matrix with each block being $\mathbf{A}_i$, so (J.8) can be further bounded by

$$\sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_k \right\rangle \le r \sum_i \lambda_i w_i^2 \sum_{k=0}^{t-1} \left(\mathbf{A}_i^k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_2^2$$
$$\le r \Bigg[\frac{14}{\delta} \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:\widehat{k}}} + \frac{10}{1-c} \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}}$$
$$+ \frac{1-c}{q - c\delta} \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{k^\dagger:k^*}} + 4t \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \Bigg],$$

where the second inequality holds due to Corollary M.11.

Appendix K. Implication in the Classical Setting

In this subsection, we show that Theorem 3.1 implies the excess risk bound in the strongly convex setting and can recover a similar result as Jain et al. [8].

We remark that our result is consistent with the acceleration of bias decay presented in Jain et al. [8]. Without instance-specific analysis, the exponential decay rate of bias is determined by the decay rate in subspace of the smallest eigenvalue. As the effective bias of ASGD decays faster than that of SGD in the eigen-subspace of small eigenvalues, the worst-case decay rate of the bias error of ASGD enjoys acceleration compared to SGD.

The hyperparameters of ASGD are chosen to be

$$\delta = \frac{1}{2\psi \operatorname{tr}(\mathbf{H})}, \quad \gamma = \sqrt{\frac{2\delta}{\psi \mu d}}, \quad \beta = \sqrt{\frac{\mu \delta}{2\psi d}}, \quad \alpha = \frac{1}{1+\beta}, \quad \text{(K.1)}$$

where $\mu := \lambda_d$ is the smallest eigenvalue of $\mathbf{H}$. We remark that the parameter choice in (K.1) is different from the choice under the overparameterized setting given in (3.1) because $\widetilde{\kappa}$ is chosen as the model dimension d, and the upper bound of γ in (3.1), which is $1/(2\psi \sum_{i>\widetilde{\kappa}} \lambda_i)$, becomes vacuous. Instead, we require $\gamma = 2\beta/\mu$ to guarantee that no eigenvalue falls in the region of small eigenvalues such that $\mathbf{A}_i$ has real eigenvalues (i.e., when $i > k^\dagger$, see the rest of this section for detailed proof). The following corollary provides the excess risk bound in the strongly convex setting:

Corollary K.1 *Under Assumptions 2.1, 2.2 and 2.3, and with the parameter choice in (K.1), the excess risk of tail-averaged iterate from ASGD in the classical regime satisfies:*

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \leq \underbrace{\frac{100}{N^2\beta^2}\exp\Big(-\frac{\beta s}{2}\Big)[L(\mathbf{w}_0) - L(\mathbf{w}^*)]}_{\textit{Effective Bias}}$$
$$+ \underbrace{\frac{1008\psi d}{N^2\beta}[L(\mathbf{w}_0) - L(\mathbf{w}^*)] + \frac{36\sigma^2 d}{N} + \frac{128\sigma^2 d}{N^2\beta}}_{\textit{Effective Variance}}.$$

Denote $\kappa := \mathrm{tr}(\mathbf{H})/\mu$, then $\beta = \Theta(1/\sqrt{\kappa\widetilde{\kappa}})$. Assuming that $L(\mathbf{w}_0) - L(\mathbf{w}^*) = \mathcal{O}(\sigma^2)$, then the bound given in Corollary K.1 fully recovers the excess risk upper bound given in Theorem 1 of [8] in terms of exponential decay rate, leading-order variance and lower-order variance. Moreover, the coefficient of effective bias is $\mathcal{O}(\kappa\widetilde{\kappa}/N^2)$, which significantly improves upon $\mathcal{O}(\kappa^{13/4}\widetilde{\kappa}^{9/4}d/N^2)$ given in [8]. It is worth noting that Liu and Belkin [11] proved $\mathcal{O}(1)$ coefficient for effective bias of ASGD. Our result can also recover the constant coefficient when $N(1-c) \geq 2$, because $1 - c = 2\alpha\beta \leq 2\beta$ and $1/(N^2\beta^2) \leq 1$. The difference in this coefficient between the bound in Liu and Belkin [11] and ours is mainly due to slightly different treatments of terms in the form of $N^{-1}\sum_{i=0}^{N-1}(1-\beta)^i$, which is not essential.

We now prove the result in the classical setting. Before we prove the theorem, we first note that with the parameter choice in (K.1) and $\widetilde{\kappa} = d$,

$$1 - \psi l = 1 - \frac{\psi\delta\,\mathrm{tr}(\mathbf{H})}{2} - \frac{1}{2} = \frac{1}{4},$$

so $r = 4$. We also note that with $\gamma\mu = 2\beta$, we have

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

where the first equality holds because $2\sqrt{c(q-\delta)(q-c\delta)} \geq 0$, the second equality holds because $c = 2\alpha - 1$ and $q = \alpha\delta + (1-\alpha)\gamma$, and the second inequality holds because $(1-\alpha)\delta \geq 0$. That is to say, there is no eigenvalue in the region of $i > k^\dagger$.

The main idea of the proof is similar to that of Theorem 3.1. We decompose the excess risk into variance and bias, and then characterize $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ and $\mathbf{M}_4$. The following lemmas provide upper bounds for the inner product of $\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ with $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$ and $\mathbf{M}_4$.

Lemma K.2 (Modified from Lemma I.2) *With $\mathbf{M}_1$ defined in (F.2), we have*

$$\Big\langle \begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \mathbf{M}_1 \Big\rangle \leq \frac{128\sigma^2 d}{N^2(1-c)}.$$

Lemma K.3 (Modified from Lemma I.3) *With* $\mathbf{M}_2$ *defined in* (F.3), *we have*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_2 \right\rangle \leq \frac{9\sigma^2 r d}{N} = \frac{36\sigma^2 d}{N}.$$

Lemma K.4 *With* $\mathbf{M}_3$ *defined in* (F.4), *we have*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_3 \right\rangle \leq \frac{100}{N^2(1-c)^2} \exp\left(-\frac{(1-c)s}{2}\right) \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 + \frac{504\psi d}{N^2(1-c)} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2.$$

Lemma K.5 *With* $\mathbf{M}_4$ *defined in* (F.5), *we have*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_4 \right\rangle \leq \frac{504\psi d}{N^2(1-c)} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2.$$

With the lemmas above, we can prove the upper bound of excess risk in the classical setting.

Theorem K.6 (Restatement of Corollary K.1) *Under Assumptions 2.1, 2.2 and 2.3, with the parameter choice in* (K.1), *we have*

$$\begin{aligned}\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \leq{}& \frac{100}{N^2(1-c)^2} \exp\left(-\frac{(1-c)s}{2}\right) \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 \\ &+ \frac{36\sigma^2 d}{N} + \frac{128}{N^2(1-c)} + \frac{1008\psi d}{N^2(1-c)} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2.\end{aligned}$$

Proof By Lemma F.1 and Lemma F.2, we have

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \leq \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3 + \mathbf{M}_4 \right\rangle. \tag{K.2}$$

Substituting the results of Lemma K.2, Lemma K.3, Lemma K.4 and Lemma K.5 into (K.2), we get the desired result. ∎

We remark that due to Lemma M.1, we have $1 - c \geq \beta$. Additionally, the constants in Theorem K.6 are smaller than those in Corollary K.1. Therefore, Theorem K.6 can fully recover Corollary K.1.

K.1. Variance Upper Bound

The proof for Lemma K.3 is straightforward given Lemma I.3 and the fact that there is no eigenvalue in the region of $i > k^\dagger$. Below we provide the proof for Lemma K.2.

Proof [Proof of Lemma K.2] According to (I.9) in the proof of Lemma I.2, we have

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_1 \right\rangle \leq \frac{\sigma^2 r}{N^2} \sum_{i=1}^{d} \lambda_i^2 \sum_{j=0}^{s-1} \left(\sum_{k=0}^{N-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix} \right)_1^2. \tag{K.3}$$

Similar to the proof of Corollary M.6, we have

(a) When $i \leq k^\ddagger$,

$$\sum_{j=0}^{s-1} \left(\sum_{k=0}^{N-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix} \right)_1^2 \leq \frac{4}{(1-c)\lambda_i^2}.$$

(b) By (M.24) and (M.26), when $k^\ddagger < i \le k^\dagger$,

$$\sum_{j=0}^{s-1}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \le \frac{32}{(1-c)\lambda_i^2}.$$

(K.3) can thus be bounded by

$$\begin{aligned}
\left\langle \begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \mathbf{M}_1\right\rangle &\le \frac{4\sigma^2}{N}\sum_{i=1}^{d}\lambda_i^2\sum_{j=0}^{s-1}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\\
&\le \frac{4\sigma^2}{N^2}\left[\sum_{i\le k^\ddagger}\lambda_i^2\cdot\frac{4}{(1-c)\lambda_i^2}+\sum_{i>k^\ddagger}\lambda_i^2\cdot\frac{32}{(1-c)\lambda_i^2}\right]\\
&= \frac{\sigma^2}{N^2(1-c)}\left[16k^\ddagger + 128(d-k^\ddagger)\right]\\
&\le \frac{128\sigma^2 d}{N^2(1-c)},
\end{aligned}$$

where first inequality holds because $r=4$ and due to (K.3), and the last inequality holds because the coefficient $16<128$. ■

K.2. Bias Upper Bound

We first provide a list of lemmas modified by considering only eigenvalues λ_i with $i \le k^\dagger$:

Lemma K.7 (Modified from Corollary M.6) *Let $\mathbf{A}_i$ be defined in (G.1). Then for all $j \ge 0$,*

$$\sum_{i=1}^{d}\lambda_i^2\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \le 9d.$$

Lemma K.7 follows directly from the corresponding results in the overparameterized regime, and we do not provide the proof here.

Lemma K.8 (Modified from Corollary M.9) *Let $\mathbf{A}_i$ be defined in (G.1). Then we have*

$$\sum_{i=1}^{d}\lambda_i w_i^2\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{s+k}\begin{bmatrix}1\\ 1\end{bmatrix}\right)_1^2 \le \frac{100}{(1-c)^2}\exp\left(-\frac{(1-c)s}{2}\right)\cdot\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}}^2.$$

Proof By Lemma M.8,

(a) For all $i \le k^\ddagger$, we have

$$\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{s+k}\begin{bmatrix}1\\ 1\end{bmatrix}\right)_1^2 \le \frac{4}{\lambda_i^2}(c\delta/q)^{2s} \le \frac{4}{\lambda_i^2}c^{s/2},$$

where the second inequality holds because $(c\delta/q)^2 \le c^2 \le \sqrt{c}$.

(b) For all $k^\ddagger < i \le \widehat{k}$, we have

$$\begin{aligned}
&2s[c(1-\delta\lambda_i)]^{s/2} + \frac{4}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{s/2} \\
&\le 2\sum_{j=0}^{s-1}[c(1-\delta\lambda_i)]^{(s+j)/4} + \frac{4}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{s/2} \\
&= 2\frac{[c(1-\delta\lambda_i)]^{s/4} - [c(1-\delta\lambda_i)]^{s/2}}{1-[c(1-\delta\lambda_i)]^{1/4}} + \frac{4}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{s/2} \\
&\le 2\frac{[c(1-\delta\lambda_i)]^{s/4} - [c(1-\delta\lambda_i)]^{s/2}}{\delta\lambda_i/4} + \frac{4}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{s/2} \\
&= \frac{8[c(1-\delta\lambda_i)]^{s/4} - 4[c(1-\delta\lambda_i)]^{s/2}}{\delta\lambda_i} \le \frac{8}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{s/4}, \qquad \text{(K.4)}
\end{aligned}$$

where the first inequality holds because $[c(1-\delta\lambda_i)]^{s/2} \le [c(1-\delta\lambda_i)]^{(s+j)/4}$, the second inequality holds because $1-[c(1-\delta\lambda_i)]^{1/4} \ge 1-(1-\delta\lambda_i)^{1/4} \ge \delta\lambda_i/4$, and the last inequality holds because $[c(1-\delta\lambda_i)]^{s/2} \ge 0$. We thus have

$$\begin{aligned}
\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{s+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1^2 &\le \left(2s[c(1-\delta\lambda_i)]^{s/2} + \frac{4}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{s/2}\right)^2 \\
&\le \frac{64}{\delta^2\lambda_i^2}[c(1-\delta\lambda_i)]^{s/2} \le \frac{64}{\delta^2\lambda_i^2}\cdot c^{s/2},
\end{aligned}$$

where the first inequality holds due to Lemma M.8, the second inequality holds due to (K.4), and the last inequality holds because $c(1-\delta\lambda_i) \le c$.

(c) For $\widehat{k} < i \le k^\dagger$, we have

$$\begin{aligned}
&2s[c(1-\delta\lambda_i)]^{s/2} + \frac{10}{1-c}[c(1-\delta\lambda_i)]^{s/2} \\
&\le 2\sum_{j=0}^{s-1}[c(1-\delta\lambda_i)]^{(s+j)/4} + \frac{10}{1-c}[c(1-\delta\lambda_i)]^{s/2} \\
&= 2\frac{[c(1-\delta\lambda_i)]^{s/4} - [c(1-\delta\lambda_i)]^{s/2}}{1-[c(1-\delta\lambda_i)]^{1/4}} + \frac{10}{1-c}[c(1-\delta\lambda_i)]^{s/2} \\
&\le 2\frac{[c(1-\delta\lambda_i)]^{s/4} - [c(1-\delta\lambda_i)]^{s/2}}{(1-c)/4} + \frac{10}{1-c}[c(1-\delta\lambda_i)]^{s/2} \\
&= \frac{8[c(1-\delta\lambda_i)]^{s/4} + 2[c(1-\delta\lambda_i)]^{s/2}}{1-c} \le \frac{10}{1-c}[c(1-\delta\lambda_i)]^{s/4}, \qquad \text{(K.5)}
\end{aligned}$$

where the first inequality holds because $[c(1-\delta\lambda_i)]^{s/2} \le [c(1-\delta\lambda_i)]^{(s+j)/4}$, the second inequality holds because $1-[c(1-\delta\lambda_i)]^{1/4} \ge 1-c^{1/4} \ge (1-c)/4$, and the last inequality holds because $[c(1-\delta\lambda_i)]^{s/2} \le [c(1-\delta\lambda_i)]^{s/4}$. We thus have

$$\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{s+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1^2 \le \left(2s[c(1-\delta\lambda_i)]^{s/2} + \frac{10}{1-c}[c(1-\delta\lambda_i)]^{s/2}\right)^2$$

$$\leq \frac{100}{(1-c)^2}[c(1-\delta\lambda_i)]^{s/2} \leq \frac{100}{(1-c)^2} \cdot c^{s/2},$$

where the first inequality holds due to Lemma M.8, the second inequality holds due to (K.5), and the last inequality holds because $c(1-\delta\lambda_i) \leq c$.

Concluding all the above,

$$\begin{aligned}
&\sum_{i=1}^{d} \lambda_i w_i^2 \left(\sum_{k=0}^{N-1} \mathbf{A}_i^{s+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_1^2 \\
&\leq \sum_{i \leq k^\ddagger} \lambda_i w_i^2 \cdot \frac{4}{\delta^2 \lambda_i^2} \cdot c^{s/2} + \sum_{k^\ddagger < i \leq \widehat{k}} \lambda_i w_i^2 \cdot \frac{64}{\delta^2 \lambda_i^2} \cdot c^{s/2} + \sum_{i > \widehat{k}} \lambda_i w_i^2 \cdot \frac{100}{(1-c)^2} \cdot c^{s/2} \\
&\leq \sum_{i \leq k^\ddagger} \lambda_i w_i^2 \cdot \frac{4}{(1-c)^2} \cdot c^{s/2} + \sum_{k^\ddagger < i \leq \widehat{k}} \lambda_i w_i^2 \cdot \frac{64}{(1-c)^2} \cdot c^{s/2} + \sum_{i > \widehat{k}} \lambda_i w_i^2 \cdot \frac{100}{(1-c)^2} \cdot c^{s/2} \\
&\leq \frac{100 c^{s/2}}{(1-c)^2} \sum_{i=1}^{d} \lambda_i w_i^2 = \frac{100 c^{s/2}}{(1-c)^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 \\
&\leq \frac{100}{(1-c)^2} \exp\left(-\frac{(1-c)s}{2} \right) \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2,
\end{aligned}$$

where the second inequality holds because $\delta\lambda_i \geq 1-c$ for $i \leq \widehat{k}$, the third inequality holds because the coefficients $4, 64, 100$ are bounded by 100, and the last inequality holds because $c^{s/2} \leq \exp(-(1-c)s/2)$. ■

Lemma K.9 *With $\mathbf{B}_t$ defined in (C.5), we have*

$$\sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_k \right\rangle \leq \frac{56}{1-c} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2.$$

Proof By Lemma J.2, taking $r = 4$, we have

$$\begin{aligned}
\sum_{k=0}^{t-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_k \right\rangle &\leq \frac{56}{\delta\lambda_i} \sum_{i \leq k^\ddagger} \lambda_i w_i^2 + \frac{40}{1-c} \sum_{k^\ddagger < i \leq k^\dagger} \lambda_i w_i^2 \\
&\leq \frac{56}{1-c} \sum_{i \leq k^\ddagger} \cdot \lambda_i w_i^2 + \frac{40}{1-c} \sum_{k^\ddagger < i \leq k^\dagger} \lambda_i w_i^2 \\
&\leq \frac{56}{1-c} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2,
\end{aligned}$$

where the first inequality holds because $\delta\lambda_i \geq 1-c$ for $i \leq k^\ddagger$, and the second inequality holds because the coefficients $40, 50$ can be bounded by 56. ■

We are now ready to bound the inner product of $\begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ with $\mathbf{M}_3$ and $\mathbf{M}_4$.

Proof [Proof of Lemma K.4] Similar to the proof of Lemma F.4, we have

$$
\begin{aligned}
&\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_3 \right\rangle \\
&\leq \frac{1}{N^2} \sum_{i=1}^{d} \lambda_i w_i^2 \left(\sum_{k=0}^{N-1} \mathbf{A}_i^{j+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_1^2 \\
&\quad + \frac{\psi}{N^2} \sum_{t=0}^{s-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_{s-1-t} \right\rangle \sum_{i=1}^{d} \lambda_i^2 \left(\sum_{k=0}^{N-1} \mathbf{A}_i^{k+t} \begin{bmatrix} \delta \\ q \end{bmatrix} \right)_1^2 \\
&\leq \frac{100}{N^2(1-c)^2} \exp\left(-\frac{(1-c)s}{2} \right) \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 + \frac{\psi}{N^2} \sum_{t=0}^{s-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_{s-1-t} \right\rangle \cdot 9d \\
&\leq \frac{100}{N^2(1-c)^2} \exp\left(-\frac{(1-c)s}{2} \right) \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 + \frac{9\psi d}{N^2} \cdot \frac{56}{1-c} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 \\
&= \frac{100}{N^2(1-c)^2} \exp\left(-\frac{(1-c)s}{2} \right) \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 + \frac{504\psi d}{N^2(1-c)} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2,
\end{aligned}
$$

where the second inequality holds due to Lemma K.8 and Lemma K.7, and the third inequality holds due to Lemma K.9. ∎

Proof [Proof of Lemma K.5] Similar to the proof of Lemma F.4, we have

$$
\begin{aligned}
\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_4 \right\rangle &\leq \frac{\psi}{N^2} \sum_{t=0}^{N-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_{s+N-t-1} \right\rangle \sum_{i=1}^{d} \lambda_i^2 \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{k} \begin{bmatrix} \delta \\ q \end{bmatrix} \right)_1^2 \\
&\leq \frac{9d\psi}{N^2} \sum_{t=0}^{N-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_{s+N-t-1} \right\rangle \\
&\leq \frac{9\psi d}{N^2} \cdot \sum_{t=0}^{s+N-1} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \mathbf{B}_{t} \right\rangle \\
&\leq \frac{9\psi d}{N^2} \cdot \frac{56}{1-c} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 = \frac{504\psi d}{N^2(1-c)} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2,
\end{aligned}
$$

where the second inequality holds due to Lemma K.7, the second inequality holds because $\mathbf{B}_t \succeq 0$, and the last inequality holds due to Lemma K.9. ∎

Appendix L. Proof for the One-hot Distribution Setting

The choice of parameters is as follows:

$$
\gamma \in (0, 1),\ \delta \in (0, \gamma],\ \beta \in (0, 1),\ \alpha = \frac{1}{1+\beta}. \tag{L.1}
$$

We now present the excess risk bound:

Theorem L.1 *Under Assumptions 2.1, 2.3 and 2.2, with the parameter choice of* (L.1)*, assuming* $N(1-c) \geq 2$*, we have the following upper bound for the excess risk:*

$$\mathbb{E}[L(\bar{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \leq 2 \cdot \mathrm{EffectiveVar} + 2 \cdot \mathrm{EffectiveBias},$$

where effective variance is bounded by

$$\begin{aligned}\mathrm{EffectiveVar} \leq \sigma^2 r \left[\frac{27k^*}{2N} + 18(s+N)\gamma^2 \sum_{i>k^*} \lambda_i^2\right] + \frac{r}{N^2}\Bigg[\frac{126}{\delta}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:\widehat{k}}} \\ + \frac{90}{1-c}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{\widehat{k}:k^\dagger}} + \frac{18}{\gamma}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}} + 36\gamma^2 N^2 s \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^2_{k^*:\infty}}\Bigg],\end{aligned}$$

and effective bias is bounded in the same way as Theorem 3.1.

The constant r is formally defined as

$$r = \frac{1}{1 - \max_{1\leq i\leq d} (\mathbf{U}_i)_{22}}, \tag{L.2}$$

Note that

$$(\mathbf{U}_i)_{22} \leq \frac{q - c\delta}{2(1-c)} \leq \frac{\gamma}{2},$$

so the upper bound of r is given by

$$r \leq \frac{1}{1-\gamma/2}.$$

The proof of Theorem L.1 depends on the following lemmas:

Lemma L.2 (Modified from Lemma I.2) *Let r be defined in* (L.2)*. Then we have*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_1 \right\rangle \leq \sigma^2 r \left[\frac{18k^*}{N} + \frac{36s(q-c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2\right].$$

Lemma L.3 (Modified from Lemma I.3) *Let r be defined in* (L.2)*. Then we have*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_2 \right\rangle \leq \sigma^2 r \left[\frac{9k^*}{N} + \frac{36N(q-c\delta)^2}{(1-c)^2} \sum_{i>k^*} \lambda_i^2\right].$$

Lemma L.4 *Let r be defined in* (L.2)*. Then we have*

$$\begin{aligned}\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_3 \right\rangle \leq \frac{r}{N^2}\Bigg[\frac{126}{\delta}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:\widehat{k}}} + \frac{90}{1-c}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{\widehat{k}:k^\dagger}} + \frac{9(1-c)}{q-c\delta}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}} \\ + \frac{36(q-c\delta)^2 N^2 s}{(1-c)^2}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^2_{k^*:\infty}}\Bigg] + \mathrm{EffectiveBias},\end{aligned}$$

where EffectiveBias *is the same as one in Theorem 3.1.*

Lemma L.5 *Let r be defined in (L.2). Then we have*

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_4 \right\rangle \le \frac{r}{N^2}\left[\frac{126}{\delta}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:\widehat{k}}} + \frac{90}{1-c}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{\widehat{k}:k^\dagger}} + \frac{9(1-c)}{q-c\delta}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}} + \frac{36(q-c\delta)^2 N^2 (s+N)}{(1-c)^2}\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^2_{k^*:\infty}}\right].$$

Proof [Proof of Theorem L.1] Note that the excess risk is

$$\left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3 + \mathbf{M}_4 \right\rangle,$$

so the upper bound can be obtained by combining Lemmas L.2, L.3, L.4 and L.5. ■

Notations. In this section, for any matrix $\mathbf{M} \in \mathbb{R}^{2d\times 2d}$, denote

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \in \mathbb{R}^{2d\times 2d},$$

where $\mathbf{M}_{ij} \in \mathbb{R}^{d\times d}$.

L.1. Analysis of Fourth Moment

In this setting, for any matrix $\mathbf{M} \in \mathbb{R}^{2d\times 2d}$, we have

$$\mathbb{E}[\widehat{\mathbf{V}}_2 \times \widehat{\mathbf{V}}_2] \circ \mathbf{M} = \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\mathbf{H} \odot \mathbf{M}_{22}) = \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathrm{diag}(\lambda_1(\mathbf{M}_{22})_{11}, \dots, \lambda_d(\mathbf{M}_{22})_{dd}).$$

Lemma L.6 (Modified from Lemma H.7) *For any PSD matrix $\mathbf{M} \in \mathbb{R}^{2d\times 2d}$ define $\mathbf{Q} := \mathcal{T}^{-1} \circ \mathbf{M}$. Then*

$$(\mathcal{I} - \mathcal{B})^{-1} \circ \mathbf{M} = \mathbf{Q} + \mathrm{diag}\left(\frac{(\mathbf{Q}_{22})_{11}}{1-(\mathbf{U}_1)_{22}}\mathbf{U}_1, \dots, \frac{(\mathbf{Q}_{22})_{dd}}{1-(\mathbf{U}_d)_{22}}\mathbf{U}_d\right).$$

Proof By Lemma H.6, we have

$$(\mathcal{I} - \mathcal{B})^{-1} \circ \mathbf{M} = \sum_{k=0}^{\infty} (\mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2])^k \circ \mathbf{Q}.$$

Note that

$$\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{Q} = \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathrm{diag}(\lambda_1(\mathbf{Q}_{22})_{11}, \dots, \lambda_d(\mathbf{Q}_{22})_{dd}),$$

so by definition of $\mathcal{T}$,

$$\mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{Q} = \mathrm{diag}((\mathbf{Q}_{22})_{11}\mathbf{U}_1, \dots, (\mathbf{Q}_{22})_{dd}\mathbf{U}_d).$$

We can similarly prove that for all $k \ge 1$,

$$(\mathcal{T}^{-1}\mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2])^k \circ \mathbf{Q} = \mathrm{diag}((\mathbf{Q}_{22})_{11}(\mathbf{U}_1)_{22}^{k-1}\mathbf{U}_1, \dots, (\mathbf{Q}_{22})_{11}(\mathbf{U}_d)_{22}^{k-1}\mathbf{U}_d).$$

Summing the above, we have

$$(\mathcal{I}-\mathcal{B})^{-1}\circ\mathbf{M}=\mathbf{Q}+\mathrm{diag}\left(\frac{(\mathbf{Q}_{22})_{11}}{1-(\mathbf{U}_1)_{22}}\mathbf{U}_1,\dots,\frac{(\mathbf{Q}_{22})_{dd}}{1-(\mathbf{U}_d)_{22}}\mathbf{U}_d\right).$$

■

Lemma L.7 (Modified from Lemma H.8) *For any PSD matrix* $\mathbf{M}\in\mathbb{R}^{2d\times 2d}$*, define the partial sum*

$$\mathbf{R}_t=\sum_{k=0}^{t-1}\mathcal{B}^k\circ\mathbf{M}.$$

Then we have

$$\mathbf{R}_t\preceq\sum_{k=0}^{t-1}\widetilde{\mathcal{B}}^k\circ\mathbf{M}+\sum_{k=0}^{t-1}\mathrm{diag}\left(\frac{((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{11}}{1-(\mathbf{U}_1)_{22}}\mathbf{U}_1,\dots,\frac{((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{dd}}{1-(\mathbf{U}_d)_{22}}\mathbf{U}_d\right).$$

and

$$\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]\preceq\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathrm{diag}\left(\frac{\lambda_1((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{11}}{1-(\mathbf{U}_1)_{22}},\dots,\frac{\lambda_d((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{dd}}{1-(\mathbf{U}_d)_{22}}\right).$$

Proof Similar to the proof of Lemma H.8, we have

$$\begin{aligned}\sum_{k=0}^{t-1}\mathcal{B}^k\circ\mathbf{M}&\preceq(\mathcal{I}-\mathcal{B})^{-1}\mathcal{T}\circ\left(\sum_{k=0}^{t-1}\widetilde{\mathcal{B}}^k\circ\mathbf{M}\right)\\&=\sum_{k=0}^{t-1}\widetilde{\mathcal{B}}^k\circ\mathbf{M}+\sum_{k=0}^{t-1}\mathrm{diag}\left(\frac{((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{11}}{1-(\mathbf{U}_1)_{22}}\mathbf{U}_1,\dots,\frac{((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{dd}}{1-(\mathbf{U}_d)_{22}}\mathbf{U}_d\right),\end{aligned}$$

where the equality holds due to Lemma L.6. We thus have

$$\begin{aligned}\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]\circ\left(\sum_{k=0}^{t-1}\mathcal{B}^k\circ\mathbf{M}\right)&\preceq\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\left[\sum_{k=0}^{t-1}\mathrm{diag}\left(\lambda_1((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{11},\dots,\lambda_d((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{dd}\right)\right.\\&\left.+\sum_{k=0}^{t-1}\mathrm{diag}\left(\frac{\lambda_1(\mathbf{U}_1)_{22}}{1-(\mathbf{U}_1)_{22}}((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{11},\dots,\frac{\lambda_d(\mathbf{U}_d)_{22}}{1-(\mathbf{U}_d)_{22}}((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{dd}\right)\right]\\&=\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathrm{diag}\left(\frac{\lambda_1((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{11}}{1-(\mathbf{U}_1)_{22}},\dots,\frac{\lambda_d((\widetilde{\mathcal{B}}^k\circ\mathbf{M})_{22})_{dd}}{1-(\mathbf{U}_d)_{22}}\right).\end{aligned}$$

■

L.2. Variance Upper Bound

We now provide the proof of Lemma L.3.
Proof [Proof of Lemma L.3] Note that

$$\begin{aligned}
\mathbf{C}_\infty &= (\mathcal{I}-\mathcal{B})^{-1}\circ\widehat{\boldsymbol{\Sigma}} \preceq \sigma^2(\mathcal{I}-\mathcal{B})^{-1}\circ\left(\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\right)\\
&= \sigma^2\left[\mathbf{U}+\mathrm{diag}\left(\frac{(\mathbf{U}_1)_{22}}{1-(\mathbf{U}_1)_{22}}\mathbf{U}_1,\dots,\frac{(\mathbf{U}_d)_{22}}{1-(\mathbf{U}_d)_{22}}\mathbf{U}_d\right)\right]\\
&= \sigma^2\mathrm{diag}\left(\frac{\mathbf{U}_1}{1-(\mathbf{U}_1)_{22}},\dots,\frac{\mathbf{U}_d}{1-(\mathbf{U}_d)_{22}}\right),
\end{aligned}$$

where the second equality holds due to Lemma L.6. We thus have

$$\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]\circ\mathbf{C}_\infty \preceq \sigma^2\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathrm{diag}\left(\frac{\lambda_1(\mathbf{U}_1)_{22}}{1-(\mathbf{U}_1)_{22}},\dots,\frac{\lambda_d(\mathbf{U}_d)_{22}}{1-(\mathbf{U}_d)_{22}}\right). \tag{L.3}$$

Therefore, $\mathbf{M}_2$ can be bounded by

$$\begin{aligned}
\mathbf{M}_2 &= \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left[(\mathcal{B}-\widetilde{\mathcal{B}})\circ\mathbf{C}_{s+t-1}+\widehat{\boldsymbol{\Sigma}}\right]\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top\\
&\preceq \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left[\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]\circ\mathbf{C}_\infty+\widehat{\boldsymbol{\Sigma}}\right]\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top\\
&\preceq \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left[\sigma^2\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\left(\mathrm{diag}\left(\frac{\lambda_1(\mathbf{U}_1)_{22}}{1-(\mathbf{U}_1)_{22}},\dots,\frac{\lambda_d(\mathbf{U}_d)_{22}}{1-(\mathbf{U}_d)_{22}}\right)+\mathbf{H}\right)\right]\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top\\
&= \frac{\sigma^2}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left[\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathrm{diag}\left(\frac{\lambda_1}{1-(\mathbf{U}_1)_{22}},\dots,\frac{\lambda_d}{1-(\mathbf{U}_d)_{22}}\right)\right]\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top\\
&\preceq \frac{\sigma^2 r}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left[\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes\mathbf{H}\right]\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top,
\end{aligned}\tag{L.4}$$

where the first inequality holds because $\mathcal{B}-\widetilde{\mathcal{B}} = \mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]-\mathbf{V}_2\otimes\mathbf{V}_2 \preceq \mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]$ and $\mathcal{B}_{s+t-1}\preceq\mathbf{C}_\infty$, the second inequality holds due to (L.3), and the last inequality holds due to definition of r. The inner product of $\mathbf{M}_2$ and $\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ can thus be bounded by

$$\begin{aligned}
\left\langle\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix},\mathbf{M}_2\right\rangle &\le \frac{\sigma^2 r}{N^2}\sum_{i=1}^{d}\lambda_i^2\sum_{t=1}^{N-1}\left(\sum_{k=0}^{N-t-1}\mathbf{A}_i^k\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_2^2\\
&\le \sigma^2 r\left[\frac{9k^*}{N}+\frac{36N(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\right],
\end{aligned}$$

where the second inequality holds by deduction similar to that of Lemma I.3. ∎

Lemma L.8 (Modified from Lemma I.6) *For any $t > 0$, $\mathbf{C}_t$ can be upper bounded by*

$$\mathbf{C}_t \preceq \sigma^2 r \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \left(\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H} \right).$$

Proof By the iteration formula $\mathbf{C}_t = \mathcal{B} \circ \mathbf{C}_{t-1} + \widehat{\mathbf{\Sigma}}$, we have

$$\begin{aligned}
\mathbf{C}_t &= \widetilde{\mathcal{B}} \circ \mathbf{C}_{t-1} + (\mathcal{B} - \widetilde{\mathcal{B}}) \circ \mathbf{C}_{t-1} + \widehat{\mathbf{\Sigma}} \\
&\preceq \widetilde{\mathcal{B}} \circ \mathbf{C}_{t-1} + \mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{C}_{t-1} + \widehat{\mathbf{\Sigma}} \\
&\preceq \widetilde{\mathcal{B}} \circ \mathbf{C}_{t-1} + \mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{C}_{\infty} + \sigma^2 \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H} \\
&\preceq \widetilde{\mathcal{B}} \circ \mathbf{C}_{t-1} + \sigma^2 \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \operatorname{diag}\left(\frac{\lambda_1}{1 - (\mathbf{U}_1)_{22}}, \ldots, \frac{\lambda_d}{1 - (\mathbf{U}_d)_{22}} \right) \\
&\leq \widetilde{\mathcal{B}} \circ \mathbf{C}_{t-1} + \sigma^2 r \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H},
\end{aligned}$$

where the first inequality holds because $\mathcal{B} - \widetilde{\mathcal{B}} \preceq \mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2]$, the second inequality holds because $\mathbf{C}_{t-1} \preceq \mathbf{C}_{\infty}$ and Lemma I.4, the third inequality holds due to (L.3), and the last inequality holds due to the definition of r. Iterating the inequality above, we have

$$\mathbf{C}_t \preceq \sigma^2 r \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \left(\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes \mathbf{H} \right).$$

■

As the bound for $\mathbf{C}_t$ is exactly the same as the bound given in Lemma I.6, we can prove the Lemma L.2 in exactly the same way as Lemma I.2.

L.3. Bias Upper Bound

Lemma L.9 (Modified from Lemma J.1) *For any $t \geq 0$, $\mathbf{B}_t$ can be upper bounded by*

$$\mathbf{B}_t \preceq \widetilde{\mathcal{B}}^t \circ \mathbf{B}_0 + \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \left(\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\mathbf{H} \odot (\mathbf{B}_{t-1-k})_{22}) \right).$$

Proof By the iterative formula $\mathbf{B}_t = \mathcal{B} \circ \mathbf{B}_{t-1}$, we have

$$\begin{aligned}
\mathbf{B}_t &= \widetilde{\mathcal{B}} \circ \mathbf{B}_{t-1} + (\mathcal{B} - \widetilde{\mathcal{B}}) \circ \mathbf{B}_{t-1} \\
&\preceq \widetilde{\mathcal{B}} \circ \mathbf{B}_{t-1} + \mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2] \circ \mathbf{B}_{t-1} \\
&= \widetilde{\mathcal{B}} \circ \mathbf{B}_{t-1} + \begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\mathbf{H} \odot (\mathbf{B}_{t-1})_{22}) \\
&\preceq \widetilde{\mathcal{B}}^t \circ \mathbf{B}_0 + \sum_{k=0}^{t-1} \widetilde{\mathcal{B}}^k \circ \left(\begin{bmatrix} \delta^2 & \delta q \\ \delta q & q^2 \end{bmatrix} \otimes (\mathbf{H} \odot (\mathbf{B}_{t-1-k})_{22}) \right),
\end{aligned}$$

where the first inequality holds because $\mathcal{B} - \widetilde{\mathcal{B}} \preceq \mathbb{E}[\widehat{\mathbf{V}}_2 \otimes \widehat{\mathbf{V}}_2]$, and the second inequality holds by iteratively applying the previous inequality. ■

Lemma L.10 *We have*

$$\sum_{t=0}^{s-1}((\mathbf{B}_t)_{22})_{ii} \le rw_i^2 \sum_{t=0}^{s-1}\left(\mathbf{A}_i^t \begin{bmatrix}1\\1\end{bmatrix}\right)_2^2.$$

Proof Note that

$$\begin{aligned}
\sum_{t=0}^{s-1}((\mathbf{B}_t)_{22})_{ii} &= \sum_{t=0}^{s-1}\left((\mathcal{B}^t \circ \mathbf{B}_0)_{22}\right)_{ii} \\
&\le \sum_{t=0}^{s-1}\left(\left(\widetilde{\mathcal{B}}^t \circ \mathbf{B}_0\right)_{22}\right)_{ii} + \sum_{t=0}^{s-1}\frac{((\widetilde{\mathcal{B}}^t \circ \mathbf{B}_0)_{22})_{ii}}{1-(\mathbf{U}_i)_{22}}(\mathbf{U}_i)_{22} \\
&= \frac{w_i^2}{1-(\mathbf{U}_i)_{22}} \sum_{t=0}^{s-1}\left(\mathbf{A}_i^t \begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \\
&\le rw_i^2 \sum_{t=0}^{s-1}\left(\mathbf{A}_i^t \begin{bmatrix}1\\1\end{bmatrix}\right)_2^2,
\end{aligned}$$

where the first inequality holds due to Lemma L.7, and the second inequality holds due to the definition of r. ■

Proof [Proof of Lemma L.4] By the bound for $\mathbf{B}_s$, we have

$$\begin{aligned}
\mathbf{M}_3 = \frac{1}{N^2}\left[\sum_{k=0}^{N-1}\mathbf{A}^k\right]\mathbf{B}_s\left[\sum_{k=0}^{N-1}\mathbf{A}^k\right]^\top &\preceq \frac{1}{N^2}\left[\sum_{k=0}^{N-1}\mathbf{A}^{k+s}\right]\mathbf{B}_0\left[\sum_{k=0}^{N-1}\mathbf{A}^{k+s}\right]^\top \\
+\frac{1}{N^2}\sum_{t=0}^{s-1}\left[\sum_{k=0}^{N-1}\mathbf{A}^{k+t}\right]&\left(\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes(\mathbf{H}\odot(\mathbf{B}_{s-1-t})_{22})\right)\left[\sum_{k=0}^{N-1}\mathbf{A}^{k+t}\right]^\top,
\end{aligned}$$

so its inner product with $\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ is

$$\begin{aligned}
\left\langle\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \mathbf{M}_3\right\rangle &\le \underbrace{\sum_{i=1}^{d}\lambda_i w_i^2\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{k+s}\begin{bmatrix}1\\1\end{bmatrix}\right)_1^2}_{\text{Effective Bias}} \\
&\quad + \underbrace{\frac{1}{N^2}\sum_{i=1}^{d}\lambda_i^2\sum_{t=0}^{s-1}((\mathbf{B}_{s-1-t})_{22})_{ii}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{k+t}\begin{bmatrix}\delta\\q\end{bmatrix}\right)_1^2}_{\mathrm{K}}.
\end{aligned}$$

The Effective Bias is the same as the standard case. K can be bounded by

$$\mathrm{K} \le \frac{1}{N^2}\left[\sum_{i=1}^{k^*}\lambda_i^2\sum_{t=0}^{s-1}((\mathbf{B}_{s-1-t})_{22})_{ii}\cdot\frac{9}{\lambda_i^2} + \sum_{i=k^*+1}^{d}\lambda_i^2\sum_{t=0}^{t-1}((\mathbf{B}_{s-1-t})_{22})_{ii}\cdot\frac{36(q-c\delta)^2N^2}{(1-c)^2}\right]$$

$$
\begin{aligned}
&= \frac{1}{N^2}\left[9\sum_{i=1}^{k^*}\sum_{t=0}^{s-1}((\mathbf{B}_{s-1-t})_{22})_{ii} + \sum_{i=k^*+1}^{d}\frac{36(q-c\delta)^2N^2\lambda_i^2}{(1-c)^2}\sum_{t=0}^{s-1}((\mathbf{B}_{s-1-t})_{22})_{ii}\right]\\
&\le \frac{r}{N^2}\left[\sum_{i\le\widehat{k}}\frac{126w_i^2}{\delta\lambda_i} + \sum_{\widehat{k}<i\le k^\dagger}\frac{90w_i^2}{1-c} + \sum_{k^\dagger<i\le k^*}\frac{9(1-c)w_i^2}{(q-c\delta)\lambda_i} + \sum_{i>k^*}\frac{36(q-c\delta)^2N^2s\lambda_i^2w_i^2}{(1-c)^2}\right]\\
&= \frac{r}{N^2}\Bigg[\frac{126}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:\widehat{k}}} + \frac{90}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{\widehat{k}:k^\dagger}} + \frac{9(1-c)}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}}\\
&\quad + \frac{36(q-c\delta)^2N^2s}{(1-c)^2}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^2_{k^*:\infty}}\Bigg],
\end{aligned}
$$

where the first inequality holds due to Corollary M.7, and the second inequality holds due to Lemma L.10. ■

Proof [Proof of Lemma L.5] For $\mathbf{M}_4$, we have

$$
\begin{aligned}
\mathbf{M}_4 &= \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]((\mathcal{B}-\widetilde{\mathcal{B}})\circ\mathbf{B}_{s+t-1})\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top\\
&\preceq \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right](\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]\circ\mathbf{B}_{s+t-1})\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top\\
&= \frac{1}{N^2}\sum_{t=1}^{N-1}\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]\left(\begin{bmatrix}\delta^2 & \delta q\\ \delta q & q^2\end{bmatrix}\otimes(\mathbf{H}\odot(\mathbf{B}_{s+t-1})_{22})\right)\left[\sum_{k=0}^{N-t-1}\mathbf{A}^k\right]^\top,
\end{aligned}
$$

where the inequality holds because $\mathcal{B}-\widetilde{\mathcal{B}}\preceq\mathbb{E}[\widehat{\mathbf{V}}_2\otimes\widehat{\mathbf{V}}_2]$. The inner produce of $\mathbf{M}_4$ and $\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ is thus bounded by

$$
\begin{aligned}
&\left\langle\mathbf{M}_4, \begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right\rangle\\
&\le \frac{1}{N^2}\sum_{i=1}^{d}\lambda_i^2\sum_{t=1}^{N-1}((\mathbf{B}_{s+t-1})_{22})_{ii}\left(\sum_{k=0}^{N-t-1}\mathbf{A}_i^k\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\\
&\le \frac{1}{N^2}\left[9\sum_{i=1}^{k^*}\sum_{t=1}^{N-1}((\mathbf{B}_{s+t-1})_{22})_{ii} + \sum_{i=k^*+1}^{d}\frac{36(q-c\delta)^2N^2\lambda_i^2}{(1-c)^2}\sum_{t=1}^{N-1}((\mathbf{B}_{s+t-1})_{22})_{ii}\right]\\
&\le \frac{1}{N^2}\left[9\sum_{i=1}^{k^*}\sum_{t=0}^{s+N-1}((\mathbf{B}_{t})_{22})_{ii} + \sum_{i=k^*+1}^{d}\frac{36(q-c\delta)^2N^2\lambda_i^2}{(1-c)^2}\sum_{t=0}^{s+N-1}((\mathbf{B}_{t})_{22})_{ii}\right]\\
&\le \frac{r}{N^2}\Bigg[\frac{126}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:\widehat{k}}} + \frac{90}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{\widehat{k}:k^\dagger}} + \frac{9(1-c)}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}}\\
&\quad + \frac{36(q-c\delta)^2N^2(s+N)}{(1-c)^2}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^2_{k^*:\infty}}\Bigg],
\end{aligned}
$$

where the second inequality holds due to Corollary M.7, the second inequality holds due to Corollary M.7, the third inequality holds because $\sum_{t=1}^{N-1}((\mathbf{B}_{s+t-1})_{22})_{ii} \leq \sum_{t=0}^{s+N-1}((\mathbf{B}_t)_{22})_{ii}$, and the last inequality holds due to Lemma L.10. ■

Appendix M. Auxiliary Lemmas

The following lemma summarizes properties of auxiliary parameters q and c in relation to model parameters α, β, γ and δ.

Lemma M.1 *We have the following properties regarding q and c:*

(a) We have $c = 2\alpha - 1$, and $0 < c < 1$. Moreover, $\beta \leq 1 - c = 2\alpha\beta \leq 2\beta$.

(b) We have $\delta \leq q \leq (1+c)\delta$. Thus, $q - \delta \leq c(q - c\delta)$.

(c) We have

$$\frac{q - c\delta}{1-c} = \frac{\gamma+\delta}{2}, \quad \frac{q-\delta}{1-c} = \frac{\gamma-\delta}{2}.$$

Thus,

$$\delta \leq \frac{q-c\delta}{1-c} \leq \gamma.$$

Proof We first recall that $c = \alpha(1-\beta)$ and $q = \alpha\delta + (1-\alpha)\gamma$.

(a) Substituting $\beta = (1-\alpha)/\alpha$ into the definition of c, we have

$$c = \alpha\left(1 - \frac{1-\alpha}{\alpha}\right) = 2\alpha - 1.$$

Note that $\beta \in (0,1)$, so $\alpha = 1/(1+\beta) \in (1/2, 1)$. Therefore, $c = 2\alpha - 1 \in (0,1)$. Moreover,

$$1 - c = 1 - \alpha(1-\beta) = 1 - \alpha + \alpha\beta \geq (1-\alpha)\beta + \alpha\beta = \beta,$$

where the equality holds because $\beta < 1$. We also have

$$1 - c = 2(1-\alpha) = 2\alpha\beta \leq 2\beta,$$

where the inequality holds because $\alpha < 1$.

(b) we have

$$q - \delta = \alpha\delta + (1-\alpha)\gamma - \delta = (1-\alpha)(\gamma-\delta) \geq 0, \tag{M.1}$$

where the inequality holds because $\gamma \geq \delta$ and $\alpha \in (0,1)$. We also have

$$q - (1+c)\delta = \alpha\delta + (1-\alpha)\gamma - 2\alpha\delta = (1-\alpha)\gamma - \alpha\delta = \alpha(\beta\gamma - \delta) = \alpha\left(\frac{\delta}{\psi\widetilde{\kappa}} - \delta\right) \leq 0,$$

where the third equality holds because $1 - \alpha = \alpha\beta$, the fourth equality holds because $\beta = \delta/(\psi\widetilde{\kappa}\gamma)$, and the last inequality holds because $\psi\widetilde{\kappa} \geq 1$. We thus have

$$(q-\delta) - c(q - c\delta) = (1-c)[q - (1+c)\delta] \leq 0.$$

(c) We have
$$q - c\delta = \alpha\delta + (1-\alpha)\gamma - (2\alpha - 1)\delta = (1-\alpha)(\gamma + \delta). \tag{M.2}$$
Combining (M.1) and (M.2) with the fact that $1 - c = 2(1-\alpha)$, we have
$$\frac{q - c\delta}{1 - c} = \frac{\gamma + \delta}{2}, \quad \frac{q - \delta}{1 - c} = \frac{\gamma - \delta}{2}.$$
Note that $\delta \le \gamma$, so
$$\delta \le \frac{q - c\delta}{1 - c} \le \gamma.$$

■

Lemma M.2 *Let x_1, x_2 be defined in (G.2) and (G.3). Then we have*

(a) $(1 - x_1)(1 - x_2) = (q - c\delta)\lambda_i$.

(b) $(c - x_1)(c - x_2) = c(q - \delta)\lambda_i$.

(c) $(1 + x_1)(1 + x_2) = 2(1 + c) - (q + c\delta)\lambda_i$.

(d) $(c\delta - qx_1)(c\delta - qx_2) = c(q - \delta)(q - c\delta)$.

Proof In the proof, we will use the properties $x_1 + x_2 = 1 + c - q\lambda_i$ and $x_1 x_2 = c(1 - \delta\lambda_i)$ extensively, which follows from Veda's Theorem.

(a) We have
$$(1 - x_1)(1 - x_2) = 1 - (x_1 + x_2) - x_1 x_2 = 1 - (1 + c - q\lambda_i) - c(1 - \delta\lambda_i) = (q - c\delta)\lambda_i.$$

(b) We have
$$(c - x_1)(c - x_2) = c^2 - c(x_1 + x_2) + x_1 x_2 = c^2 - c(1 + c - q\lambda_i) + c(1 - \delta\lambda_i) = c(q - \delta)\lambda_i.$$

(c) We have
$$\begin{aligned}(1 + x_1)(1 + x_2) &= 1 + (x_1 + x_2) + x_1 x_2 = 1 + (1 + c - q\lambda_i) + c(1 - \delta\lambda_i) \\ &= 2(1 + c) - (q + c\delta)\lambda_i.\end{aligned}$$

(d) We have
$$\begin{aligned}(c\delta - qx_1)(c\delta - qx_2) &= c^2\delta^2 - c\delta q(x_1 + x_2) + q^2 x_1 x_2 \\ &= c^2\delta^2 - c\delta q(1 + c - q\lambda_i) + q^2 \cdot c(1 - \delta\lambda_i) \\ &= c(q - \delta)(q - c\delta)\end{aligned}$$

■

Lemma M.3 *For a given PSD matrix $\mathbf{M}$, we define the following sequence of matrices recursively: $\mathbf{R}_0 = \mathbf{0}$, and*

$$\mathbf{R}_{t+1} = \mathcal{B} \circ \mathbf{R}_t + \mathbf{M}, \quad t \geq 0. \tag{M.3}$$

Then for all $t \geq 0$, we have

$$\mathbf{R}_t = \sum_{k=0}^{t-1} \mathcal{B}^k \circ \mathbf{M}. \tag{M.4}$$

Thus, $\mathbf{R}_t$ is an increasing sequence:

$$\mathbf{R}_0 \preceq \mathbf{R}_1 \preceq \cdots \preceq \mathbf{R}_\infty. \tag{M.5}$$

Proof We prove (M.4) by induction. When $t = 0$, (M.4) holds trivially. Suppose that (M.4) holds for t. By the recursive formula M.3, we have

$$\mathbf{R}_{t+1} = \mathcal{B} \circ \mathbf{R}_t + \mathbf{M} = \mathcal{B} \circ \left(\sum_{k=0}^{t-1} \mathcal{B}^k \circ \mathbf{M} \right) + \mathbf{M} = \sum_{k=1}^{t} \mathcal{B}^k \circ \mathbf{M} + \mathbf{M} = \sum_{k=0}^{t} \mathcal{B}^k \circ \mathbf{M},$$

where the second equality holds due to the induction hypothesis. Thus, (M.4) holds for $t + 1$.

By (M.4), note that

$$\mathbf{R}_{t+1} - \mathbf{R}_t = \sum_{k=0}^{t} \mathcal{B}^k \circ \mathbf{M} - \sum_{k=0}^{t-1} \mathcal{B}^k \circ \mathbf{M} = \mathcal{B}^t \circ \mathbf{M} \succeq 0,$$

where the inequality holds due to Lemma H.2(c). Therefore, $\mathbf{R}_t \preceq \mathbf{R}_{t+1}$. ∎

Lemma M.4 *Let $\{\mathbf{M}_t\}_{t \geq 1}$ be a sequence of PSD matrices and s, N be positive integers. Then*

$$\begin{aligned}
&\sum_{t=s}^{s+N-1} \left[\sum_{k=t+1}^{s+N-1} \mathbf{A}^{k-t} \mathbf{M}_t + \mathbf{M}_t + \sum_{k=t+1}^{s+N-1} \mathbf{M}_t (\mathbf{A}^\top)^{k-t} \right] \\
&= \left[\sum_{k=0}^{N-1} \mathbf{A}^k \right] \mathbf{M}_s \left[\sum_{k=0}^{N-1} \mathbf{A}^k \right]^\top + \sum_{t=1}^{N-1} \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k \right] (\mathbf{M}_{s+t} - \widetilde{\mathcal{B}} \circ \mathbf{M}_{s+t-1}) \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k \right]^\top .
\end{aligned}$$

Proof For $t = s, s+1, \ldots, s+N-2$, we have

$$\begin{aligned}
&\left[\sum_{j=0}^{s+N-t-1} \mathbf{A}^j \right] \mathbf{M}_t \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k \right]^\top - \left[\sum_{j=0}^{s+N-t-2} \mathbf{A}^j \right] (\mathbf{A} \mathbf{M}_t \mathbf{A}^\top) \left[\sum_{k=0}^{s+N-t-2} \mathbf{A}^k \right]^\top \\
&= \sum_{j,k=0}^{s+N-t-1} \mathbf{A}^j \mathbf{M}_t (\mathbf{A}^k)^\top - \sum_{j,k=1}^{s+N-t-1} \mathbf{A}^j \mathbf{M}_t (\mathbf{A}^k)^\top \\
&= \sum_{j=1}^{s+N-t-1} \mathbf{A}^j \mathbf{M}_t + \mathbf{M}_t + \sum_{k=1}^{s+N-t-1} (\mathbf{A}^\top)^k
\end{aligned}$$

$$= \sum_{k=t+1}^{s+N-1} \mathbf{A}^{k-t}\mathbf{M}_t + \mathbf{M}_t + \sum_{k=t+1}^{s+N-1} (\mathbf{A}^\top)^{k-t}.$$

Take the sum over t, and we have

$$\sum_{t=s}^{s+N-1} \left[\sum_{k=t+1}^{s+N-1} \mathbf{A}^{k-t}\mathbf{M}_t + \mathbf{M}_t + \sum_{k=t+1}^{s+N-1} \mathbf{M}_t(\mathbf{A}^\top)^{k-t}\right]$$
$$= \mathbf{M}_{s+N-1} + \sum_{t=s}^{s+N-2} \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k\right] \mathbf{M}_t \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k\right]^\top$$
$$- \sum_{t=s}^{s+N-2} \left[\sum_{k=0}^{s+N-t-2} \mathbf{A}^k\right] (\mathbf{A}\mathbf{M}_t\mathbf{A}^\top) \left[\sum_{k=0}^{s+N-t-2} \mathbf{A}^k\right]^\top$$
$$= \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right] \mathbf{M}_s \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right]^\top + \sum_{t=s+1}^{s+N-1} \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k\right] \mathbf{M}_t \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k\right]^\top$$
$$- \sum_{t=s+1}^{s+N-1} \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k\right] (\mathbf{A}\mathbf{M}_{t-1}\mathbf{A}^\top) \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k\right]^\top$$
$$= \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right] \mathbf{M}_s \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right]^\top + \sum_{t=s+1}^{s+N-1} \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k\right] (\mathbf{M}_t - \widetilde{\mathcal{B}} \circ \mathbf{M}_{t-1}) \left[\sum_{k=0}^{s+N-t-1} \mathbf{A}^k\right]^\top$$
$$= \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right] \mathbf{M}_s \left[\sum_{k=0}^{N-1} \mathbf{A}^k\right]^\top + \sum_{t=1}^{N-1} \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k\right] (\mathbf{M}_{s+t} - \widetilde{\mathcal{B}} \circ \mathbf{M}_{s+t-1}) \left[\sum_{k=0}^{N-t-1} \mathbf{A}^k\right]^\top,$$

where the second equality holds due to change of index, the third equality holds due to the definition of $\widetilde{\mathcal{B}}$, and the fourth equality holds also due to change of index. ■

Lemma M.5 *With $\mathbf{A}_i$ defined in (G.1), let x_1 and x_2 be the eigenvalues of $\mathbf{A}_i$ defined in (G.2) and (G.3). Then*

- *For all $i \le k^\ddagger$, we have*

$$-\frac{2}{\lambda_i}(c\delta/q)^j \le \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1 \le \frac{2}{\lambda_i}(c\delta/q)^j.$$

- *For all $k^\ddagger < i \le \widehat{k}$, we have*

$$\left|\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1\right| \le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \delta j[c(1-\delta\lambda_i)]^{(j-1)/2}$$

- *For all $\widehat{k} < i \le k^\dagger$, we have*

$$\left|\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1\right| \le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{1-c}{\lambda_i} \cdot j[c(1-\delta\lambda_i)]^{(j-1)/2}$$

- *For all $i > k^\dagger$, we have*

$$0 \le \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1 \le \frac{3}{\lambda_i}\left[1 - \left(1 - 2\frac{q-c\delta}{1-c}\lambda_i\right)^t\right]\left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^j$$

Proof Note that

$$\begin{aligned}\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix} &= (\mathbf{A}_i^j - \mathbf{A}_i^{j+t})(\mathbf{I} - \mathbf{A}_i)^{-1} \begin{bmatrix} \delta \\ q \end{bmatrix} = (\mathbf{A}_i^j - \mathbf{A}_i^{j+t}) \cdot \frac{1}{\lambda_i} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{\lambda_i} \begin{bmatrix} (\mathbf{A}_i^j)_{11} + (\mathbf{A}_i^j)_{12} - (\mathbf{A}_i^{j+t})_{11} - (\mathbf{A}_i^{j+t})_{12} \\ (\mathbf{A}_i^j)_{11} + (\mathbf{A}_i^j)_{12} - (\mathbf{A}_i^{j+t})_{11} - (\mathbf{A}_i^{j+t})_{12} \end{bmatrix}. \end{aligned} \tag{M.6}$$

Combining Lemma G.3 with (M.6), we have

$$\begin{aligned} &\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1 = \frac{1}{\lambda_i}((\mathbf{A}_i^j)_{11} + (\mathbf{A}_i^j)_{12} - (\mathbf{A}_i^{j+t})_{11} - (\mathbf{A}_i^{j+t})_{12}) \\ &= \frac{1}{\lambda_i}\left[-\frac{x_1 x_2^j - x_2 x_1^j}{x_2 - x_1} + (1-\delta\lambda_i)\frac{x_2^j - x_1^j}{x_2 - x_1} + \frac{x_1 x_2^{j+t} - x_2 x_1^{j+t}}{x_2 - x_1} - (1-\delta\lambda_i)\frac{x_2^{j+t} - x_1^{j+t}}{x_2 - x_1}\right] \\ &= \frac{1}{\lambda_i} \cdot \frac{(1-\delta\lambda_i - x_1)x_2^j(1-x_2^t) - (1-\delta\lambda_i - x_2)x_1^j(1-x_1^t)}{x_2 - x_1}. \end{aligned} \tag{M.7}$$

For $i \le k^\ddagger$, note that $x_1 x_2 = c(1-\delta\lambda_i)$ and $x_2 \le c$ by Lemma G.2, so we have

$$1 - \delta\lambda_i \le x_1 \le x_2 \le c. \tag{M.8}$$

Thus, the upper bound of (M.7) is given by

$$\begin{aligned} \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+t} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1 &\le \frac{1}{\lambda_i} \cdot \frac{-(\delta\lambda_i + x_1 - 1)x_1^j(1-x_2^t) + (\delta\lambda_i + x_2 - 1)x_1^j(1-x_1^t)}{x_2 - x_1} \\ &= \frac{x_1^j}{\lambda_i}\left[(\delta\lambda_i + x_1 - 1)\frac{x_2^t - x_1^t}{x_2 - x_1} + (1 - x_1^t)\right] \\ &\le \frac{x_1^j}{\lambda_i}\left[(\delta\lambda_i + x_1 - 1)\frac{1 - x_1^t}{1 - x_1} + (1 - x_1^t)\right] = \frac{\delta x_1^j(1-x_1^t)}{1-x_1} \le \frac{\delta x_1^j}{1-x_1}, \end{aligned} \tag{M.9}$$

where the first inequality holds because $\delta\lambda_i + x_1 - 1 \ge 0$ and $x_2 \ge x_1$, and the second inequality holds due to Lemma M.12. Note that

$$\begin{aligned} \frac{1}{1-x_1} &= \frac{1-x_2}{(1-x_2)(1-x_1)} = \frac{1-x_2}{(q-c\delta)\lambda_i} \\ &\le \frac{1 - \frac{c\delta - \sqrt{c(q-\delta)(q-c\delta)}}{q}}{(q-c\delta)\lambda_i} = \frac{1}{q\lambda_i}\left(1 + \sqrt{\frac{c(q-\delta)}{q-c\delta}}\right) \le \frac{2}{\delta\lambda_i}, \end{aligned} \tag{M.10}$$

where the second equality holds by Lemma M.2(a), the first inequality holds due to Lemma G.2, and the second inequality holds because $c(q-\delta) \le q - c\delta$ and $q \ge \delta$. Note that $x_1 \le x_2 \le c\delta/q$, so (M.9) can be further bounded by

$$\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+t} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1 \le \frac{\delta}{1-x_1}(c\delta/q)^j \le \frac{2}{\lambda_i}(c\delta/q)^j,$$

where the second inequality holds due to (M.10).

The lower bound of (M.7) is given by

$$\begin{aligned}
\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1 &\ge \frac{1}{\lambda_i} \cdot \frac{-(\delta\lambda_i + x_1 - 1)x_2^j(1-x_2^t) + (\delta\lambda_i + x_2 - 1)x_1^j(1-x_2^t)}{x_2 - x_1} \\
&= \frac{1-x_2^t}{\lambda_i} \cdot \frac{-(\delta\lambda_i + x_1 - 1)x_2^j + (\delta\lambda_i + x_2 - 1)x_1^j}{x_2 - x_1},
\end{aligned} \tag{M.11}$$

where the first inequality holds because $\delta\lambda_i + x_1 - 1 \ge 0$ and $x_1 \le x_2$. If $j \ge 1$, then

$$\begin{aligned}
&\frac{-(\delta\lambda_i + x_1 - 1)x_2^j + (\delta\lambda_i + x_2 - 1)x_1^j}{x_2 - x_1} \\
&= -(\delta\lambda_i + x_1 - 1)x_2 \frac{x_2^{j-1} - x_1^{j-1}}{x_2 - x_1} + (1-\delta\lambda_i)x_1^{j-1} \\
&\ge -(\delta\lambda_i + x_1 - 1)x_2 \cdot \frac{(c\delta/q)^{j-1}}{c\delta/q - x_1} \\
&= -\frac{x_1 x_2 - (1-\delta\lambda_i)x_2}{c\delta/q - x_1} \cdot (c\delta/q)^{j-1} \\
&= -(1-\delta\lambda_i)\frac{c - x_2}{c\delta/q - x_1}(c\delta/q)^{j-1},
\end{aligned} \tag{M.12}$$

where the inequality holds due to Lemma M.12, and the last equality holds because $x_1 x_2 = c(1 - \delta\lambda_i)$. Note that

$$\begin{aligned}
&(1-\delta\lambda_i)\frac{c - x_2}{c\delta/q - x_1} = (1-\delta\lambda_i) \cdot \frac{(c-x_2)(c\delta/q - x_2)}{(c\delta/q - x_1)(c\delta/q - x_2)} = \frac{q^2(1-\delta\lambda_i)(c-x_2)(c\delta/q - x_2)}{c(q-\delta)(q-c\delta)} \\
&\le \frac{q^2}{c(q-\delta)(q-c\delta)} \cdot \left(1 - \delta \cdot \frac{(\sqrt{q-c\delta} + \sqrt{c(q-\delta)})^2}{q^2}\right) \cdot \left(c - \frac{c\delta - \sqrt{c(q-\delta)(q-c\delta)}}{q}\right) \\
&\quad \cdot \frac{\sqrt{c(q-\delta)(q-c\delta)}}{q} \\
&= \frac{(c\delta - \sqrt{c(q-\delta)(q-c\delta)})^2}{cq^2}\left(1 + \sqrt{\frac{c(q-\delta)}{q-c\delta}}\right) \\
&\le \frac{c^2\delta^2}{cq^2} \cdot 2 \le 2\frac{c\delta}{q},
\end{aligned} \tag{M.13}$$

where the second equality holds due to Lemma M.2(d), the first inequality holds due to (G.6) and Lemma G.2, the second inequality holds because $c\delta - \sqrt{c(q-\delta)(q-c\delta)} \le c\delta$ and $c(q-\delta) \le$

$q - c\delta$, and the last inequality holds because $\delta \leq q$. Therefore, substituting (M.12) and (M.13) into (M.11), we have

$$\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1 \geq -\frac{1 - x_2^t}{\lambda_i} \cdot 2(c\delta/q)^j \geq -\frac{2}{\lambda_i}(c\delta/q)^j,$$

where the second inequality holds because $1 - x_2^t \leq 1$. If $j = 0$, then

$$\frac{1 - x_2^t}{\lambda_i} \cdot \frac{-(\delta\lambda_i + x_1 - 1) + (\delta\lambda_i + x_2 - 1)}{x_2 - x_1} = \frac{1 - x_2^t}{\lambda_i} \geq 0,$$

so the upper bound holds trivially.

For $k^\ddagger < i \leq k^\dagger$, the upper bound of (M.7) is given by

$$\begin{aligned}
\left|\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1\right| &= \frac{1}{\lambda_i}\left|\frac{(1 - \delta\lambda_i - x_1)x_2^j(1 - x_2^t) - (1 - \delta\lambda_i - x_2)x_1^j(1 - x_1^t)}{x_2 - x_1}\right| \\
&= \frac{1}{\lambda_i}\Bigg|\frac{x_2^j(1 - x_2^t) + x_1^j(1 - x_1^t)}{2} \\
&\quad + \left(1 - \delta\lambda_i - \frac{x_1 + x_2}{2}\right) \cdot \frac{x_2^j(1 - x_2^t) - x_1^j(1 - x_1^t)}{x_2 - x_1}\Bigg| \\
&= \frac{1}{\lambda_i}\Bigg|\frac{x_2^j(1 - x_2^t) + x_1^j(1 - x_1^t)}{2} + \frac{1 - c - (2\delta - q)\lambda_i}{2} \\
&\quad \cdot \left[\frac{x_2^j - x_1^j}{x_2 - x_1} \cdot \frac{(1 - x_1^t) + (1 - x_2^t)}{2} - \frac{x_2^t - x_1^t}{x_2 - x_1} \cdot \frac{x_2^j + x_1^j}{2}\right]\Bigg| \\
&\leq \frac{|x_2^j||1 - x_2^t| + |x_1^j||1 - x_1^t|}{2\lambda_i} + \frac{|1 - c - (2\delta - q)\lambda_i|}{2\lambda_i} \\
&\quad \cdot \left[\left|\frac{x_2^j - x_1^j}{x_2 - x_1}\right| \cdot \frac{|1 - x_1^t| + |1 - x_2^t|}{2} + \left|\frac{x_2^t - x_1^t}{x_2 - x_1}\right| \cdot \frac{|x_2^j| + |x_1^j|}{2}\right], \quad \text{(M.14)}
\end{aligned}$$

where the second equality holds because $x_1 + x_2 = 1 + c - q\lambda_i$, and the inequality holds due to triangle inequality. Note that $|1 - x_1^t| = |1 - x_2^t| \leq 1 + |x_2^t| \leq 2$ because $|x_2^t| \leq 1$. We can thus bound (M.14) as

$$\begin{aligned}
&\left|\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1\right| \\
&\leq \frac{|x_1^j| + |x_2^j|}{\lambda_i} + \frac{|1 - c - (2\delta - q)\lambda_i|}{2\lambda_i} \cdot \left[2\left|\frac{x_2^j - x_1^j}{x_2 - x_1}\right| + \left|\frac{x_2^t - x_1^t}{x_2 - x_1}\right| \cdot \frac{|x_2^j| + |x_1^j|}{2}\right] \\
&\leq \frac{2}{\lambda_i}[c(1 - \delta\lambda_i)]^{j/2} + \frac{|1 - c - (2\delta - q)\lambda_i|}{2\lambda_i} \cdot \left[2j[c(1 - \delta\lambda_i)]^{(j-1)/2} + t[c(1 - \delta\lambda_i)]^{(j+t-1)/2}\right] \\
&= j \cdot \frac{|1 - c - (2\delta - q)\lambda_i|}{\lambda_i} \cdot [c(1 - \delta\lambda_i)]^{(j-1)/2}
\end{aligned}$$

$$+\left\{\frac{2}{\lambda_i}+\frac{|1-c-(2\delta-q)\lambda_i|}{2\lambda_i}\cdot t[c(1-\delta\lambda_i)]^{(t-1)/2}\right\}\cdot[c(1-\delta\lambda_i)]^{j/2}, \tag{M.15}$$

where the second inequality holds due to Lemma M.13. For $k^\ddagger < i \le \widehat{k}$, we have

$$1-c-(2\delta-q)\lambda_i \le \delta\lambda_i-(2\delta-q)\lambda_i=(q-\delta)\lambda_i\le\delta\lambda_i,$$

where the first inequality holds because $1-c\le\delta\lambda_i$, and the second inequality holds because $q\le(1+c)\delta\le 2\delta$. We also have

$$1-c-(2\delta-q)\lambda_i\ge(q-2\delta)\lambda_i\ge(\delta-2\delta)\lambda_i=-\delta\lambda_i,$$

where the first inequality holds because $1-c\ge 0$, and the second inequality holds because $q\ge\delta$. Therefore,

$$|1-c-(2\delta-q)\lambda_i|\le\delta\lambda_i. \tag{M.16}$$

(M.15) can thus be bounded by

$$\begin{aligned}\left|\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1\right| &\le \delta j[c(1-\delta\lambda_i)]^{(j-1/2)}+\left(\frac{2}{\lambda_i}+\frac{\delta}{2}\cdot\frac{2}{\delta\lambda_i}\right)\cdot[c(1-\delta\lambda_i)]^{j/2}\\ &= \delta j[c(1-\delta\lambda_i)]^{(j-1/2)}+\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2},\end{aligned}$$

where the inequality holds due to (M.16) and Lemma M.14. For $\widehat{k} < i \le k^\dagger$, we have

$$1-c-(2\delta-q)\lambda_i\ge 1-c-(1-c)(2\delta-q)/\delta=\frac{(1-c)(q-\delta)}{\delta}\ge 0,$$

where the first inequality holds because $\lambda_i\le(1-c)/\delta$, and the second inequality holds because $q\ge\delta$. We also have

$$1-c-(2\delta-q)\lambda_i\le 1-c,$$

where the inequality holds because $2\delta-q\ge 2\delta-(1+c)\delta=(1-c)\delta\ge 0$. Therefore,

$$|1-c-(2\delta-q)\lambda_i|\le 1-c. \tag{M.17}$$

(M.15) can thus be bounded as

$$\begin{aligned}\left|\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1\right| &\le \frac{1-c}{\lambda_i}\cdot j[c(1-\delta\lambda_i)]^{(j-1)/2}+\left(\frac{2}{\lambda_i}+\frac{1-c}{2\lambda_i}\cdot\frac{2}{1-c}\right)\cdot[c(1-\delta\lambda_i)]^{j/2}\\ &= \frac{1-c}{\lambda_i}\cdot j[c(1-\delta\lambda_i)]^{(j-1)/2}+\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2},\end{aligned}$$

where the inequality holds due to (M.17) and Lemma M.14.

For $i > k^\dagger$, note that

$$1-\delta\lambda_i-x_2\ge(1-\delta\lambda_i)-\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)=\frac{q-\delta}{1-c}\lambda_i\ge 0,$$

where the first inequality holds due to Lemma G.2, and the second inequality holds because $q \geq \delta$. The upper bound of (M.7) is thus given by

$$\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1 \leq \frac{1}{\lambda_i} \cdot \frac{(1-\delta\lambda_i - x_1)x_2^j(1-x_2^t) - (1-\delta\lambda_i - x_2)x_1^j(1-x_2^t)}{x_2 - x_1}$$
$$= \frac{1-x_2^t}{\lambda_i}\left[(1-\delta\lambda_i - x_2)\frac{x_2^j - x_1^j}{x_2 - x_1} + x_2^j\right] \tag{M.18}$$

where the inequality holds due to $x_1 < x_2$. If $j \geq 1$,

$$(1-\delta\lambda_i - x_2)\frac{x_2^j - x_1^j}{x_2 - x_1} + x_2^j = (1-\delta\lambda_i - x_2)x_1\frac{x_2^{j-1} - x_1^{j-1}}{x_2 - x_1} + (1-\delta\lambda_i)x_2^{j-1}$$
$$\leq \left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^{j-1}\left[\frac{(1-\delta\lambda_i - x_2)x_1}{1-\frac{q-c\delta}{1-c}\lambda_i - x_1} + (1-\delta\lambda_i)\right]$$
$$= \left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^{j-1} \cdot \frac{(1-\delta\lambda_i)\left(1-\frac{q-c\delta}{1-c}\lambda_i\right) - c(1-\delta\lambda_i)}{1-\frac{q-c\delta}{1-c}\lambda_i - x_1}$$
$$= \left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^{j} \frac{1-\delta\lambda_i}{1-\frac{q-c\delta}{1-c}\lambda_i} \cdot \frac{1-\frac{q-c\delta}{1-c}\lambda_i - c}{1-\frac{q-c\delta}{1-c}\lambda_i - x_1},$$

where the inequality holds due to Lemma M.12. Note that

$$\frac{1-\delta\lambda_i}{1-\frac{q-c\delta}{1-c}\lambda_i} \leq \frac{1-\delta\cdot\frac{(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}}{1-\frac{q-c\delta}{1-c}\cdot\frac{(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}}$$
$$= \frac{(1+\sqrt{c(q-c\delta)/(q-\delta)})^2}{(1+\sqrt{c(q-c\delta)/(q-\delta)})^2 - (1-c)} \leq \frac{(1+1)^2}{(1+1)^2-(1-c)} = \frac{4}{3+c}, \tag{M.19}$$

where the first inequality holds due to (G.6), and the second inequality holds because $q - \delta \leq c(q - c\delta)$. We also note that

$$\frac{1-\frac{q-c\delta}{1-c}\lambda_i - c}{1-\frac{q-c\delta}{1-c}\lambda_i - x_1} = \frac{1-\frac{q-c\delta}{1-c}\lambda_i - c}{1-\frac{q-c\delta}{1-c}\lambda_i - \frac{(1+c-q\lambda_i)-\sqrt{(1+c-q\lambda_i)^2-4c(1-\delta\lambda_i)}}{2}}$$
$$\leq 2\cdot\frac{1-c-\frac{q-c\delta}{1-c}\lambda_i}{1-c-\frac{(1+c)q-2c\delta}{1-c}\lambda_i} \leq 2\cdot\frac{1-c-\frac{q-c\delta}{1-c}\cdot\frac{(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}}{1-c-\frac{(1+c)q-2c\delta}{1-c}\cdot\frac{(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}}$$
$$= 2 + \sqrt{\frac{c(q-\delta)}{q-c\delta}} \leq 2 + c, \tag{M.20}$$

where the first inequality holds because $\sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)} \geq 0$, the second inequality holds due to (G.6), and the third inequality holds because $q - \delta \leq c(q - c\delta)$. Combining (M.19) and (M.20), we have

$$\frac{1-\delta\lambda_i}{1-\frac{q-c\delta}{1-c}\lambda_i}\cdot\frac{1-\frac{q-c\delta}{1-c}\lambda_i-c}{1-\frac{q-c\delta}{1-c}\lambda_i-x_1}\leq\frac{4(2+c)}{3+c}\leq 3, \tag{M.21}$$

where the second inequality holds because $c\leq 1$. where the second inequality holds because $x_2\geq 1-2\frac{q-c\delta}{1-c}\lambda_i$ due to Lemma G.2. Combining (M.18) with (M.21), we have

$$\begin{aligned}\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1 &\leq\frac{3}{\lambda_i}\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^j(1-x_2^t)\\ &\leq\frac{3}{\lambda_i}\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^j\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^t\right],\end{aligned}$$

where the second inequality holds due to Lemma G.2. If $j=0$, then by (M.18)

$$\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1\leq\frac{1-x_2^t}{\lambda_i}\leq\frac{1}{\lambda_i}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^t\right],$$

where the second inequality holds due to Lemma G.2. Thus, the upper bound also holds for $j=0$.

The lower bound of (M.7) is given by

$$\begin{aligned}\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1 &=\frac{1}{\lambda_i}\left[(1-\delta\lambda_i-x_2)\frac{x_2^j(1-x_2^t)-x_1^j(1-x_1^t)}{x_2-x_1}+x_2^j(1-x_2^t)\right]\\ &\geq\frac{1}{\lambda_i}\left[(1-\delta\lambda_i-x_2)\frac{x_2^j(1-x_2^t)-x_2^j(1-x_1^t)}{x_2-x_1}+x_2^j(1-x_2^t)\right]\\ &=\frac{x_2^j}{\lambda_i}\left[-(1-\delta\lambda_i-x_2)\frac{x_2^t-x_1^t}{x_2-x_1}+(1-x_2^t)\right]\\ &\geq\frac{x_2^j}{\lambda_i}\left[-(1-\delta\lambda_i-x_2)\frac{1-x_2^t}{1-x_2}+(1-x_2^t)\right]\\ &=\frac{\delta x_2^j(1-x_2^t)}{1-x_2}\geq 0,\end{aligned}$$

where the first inequality holds because $x_1\leq x_2$, the second inequality holds due to Lemma M.12, and the third inequality holds because $0<x_2<1$. ∎

The following corollaries follow from Lemma M.5.

Corollary M.6 *With $\mathbf{A}_i$ defined in (G.1), assuming that $N(1-c)\geq 2$, we have*

$$\sum_{i=1}^d\lambda_i^2\sum_{j=0}^{s-1}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\leq 18Nk^*+\frac{36sN^2(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2.$$

Proof By Lemma M.5, we have

(a) For $i \le k^\ddagger$,

$$\sum_{j=0}^{s-1}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \le \frac{4}{\lambda_i^2}\sum_{j=0}^{s-1}(c\delta/q)^{2j} \le \frac{4}{\lambda_i^2}\sum_{j=0}^{s-1}c^j = \frac{4}{\lambda_i^2}\cdot\frac{1-c^s}{1-c} \le \frac{4}{(1-c)\lambda_i^2} \le \frac{2N}{\lambda_i^2}, \tag{M.22}$$

where the second inequality holds because $(c\delta/q)^2 \le c^2 \le c$, the third inequality holds because $1-c^s \le 1$, and the last inequality holds due to the assumption that $N(1-c) \ge 2$.

(b) For $k^\ddagger < i \le \widehat{k}$,

$$\begin{aligned}
&\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \delta j[c(1-\delta\lambda_i)]^{(j-1)/2} \\
&\le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + 2\delta[c(1-\delta\lambda_i)]^{j/4}\sum_{t=0}^{j/2-1}[c(1-\delta\lambda_i)]^{t/2} \\
&= \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + 2\delta\frac{[c(1-\delta\lambda_i)]^{j/4} - [c(1-\delta\lambda_i)]^{j/2}}{1-\sqrt{c(1-\delta\lambda_i)}} \\
&\le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + 2\delta\frac{[c(1-\delta\lambda_i)]^{j/4} - [c(1-\delta\lambda_i)]^{j/2}}{\delta\lambda_i/2} \\
&= \frac{4[c(1-\delta\lambda_i)]^{j/4} - [c(1-\delta\lambda_i)]^{j/2}}{\lambda_i} \le \frac{4}{\lambda_i}[c(1-\delta\lambda_i)]^{j/4},
\end{aligned} \tag{M.23}$$

where the first inequality holds because $[c(1-\delta\lambda_i)]^{j/4-1/2} \le [c(1-\delta\lambda_i)]^{t/2}$ for all $t \le j/2-1$, the second inequality holds because $1-\sqrt{c(1-\delta\lambda_i)} \ge 1-\sqrt{1-\delta\lambda_i} \ge \delta\lambda_i/2$, and the last inequality holds because $[c(1-\delta\lambda_i)]^{j/2} \ge 0$. We thus have

$$\begin{aligned}
\sum_{j=0}^{s-1}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 &\le \sum_{j=0}^{s-1}\left(\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \delta j[c(1-\delta\lambda_i)]^{(j-1)/2}\right)^2 \\
&\le \frac{16}{\lambda_i^2}\sum_{j=0}^{s-1}[c(1-\delta\lambda_i)]^{j/2} = \frac{16}{\lambda_i^2}\cdot\frac{1-[c(1-\delta\lambda_i)]^{s/2}}{1-\sqrt{c(1-\delta\lambda_i)}} \\
&\le \frac{16}{\lambda_i^2}\cdot\frac{1}{1-\sqrt{c(1-\delta\lambda_i)}} \le \frac{16}{\lambda_i^2}\cdot\frac{1}{(1-c)/2} = \frac{32}{(1-c)\lambda_i^2} \\
&\le \frac{16N}{\lambda_i^2},
\end{aligned} \tag{M.24}$$

where the first inequality holds due to Lemma M.5, the second inequality holds due to (M.23), the third inequality holds because $1-[c(1-\delta\lambda_i)]^{s/2} \le 1$, the fourth inequality holds because $1-\sqrt{c(1-\delta\lambda_i)} \ge 1-\sqrt{c} \ge (1-c)/2$, and the last inequality holds due to the assumption that $N(1-c) \ge 2$.

(c) For $\widehat{k} < i \le k^\dagger$,

$$\begin{aligned}
&\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{1-c}{\lambda_i}\cdot j[c(1-\delta\lambda_i)]^{(j-1)/2} \\
&\le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{2(1-c)}{\lambda_i}[c(1-\delta\lambda_i)]^{j/4}\sum_{t=0}^{j/2-1}[c(1-\delta\lambda_i)]^{t/2} \\
&= \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{2(1-c)}{\lambda_i}\cdot\frac{[c(1-\delta\lambda_i)]^{j/4}-[c(1-\delta\lambda_i)]^{j/2}}{1-\sqrt{c(1-\delta\lambda_i)}} \\
&\le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{2(1-c)}{\lambda_i}\cdot\frac{[c(1-\delta\lambda_i)]^{j/4}-[c(1-\delta\lambda_i)]^{j/2}}{(1-c)/2} \\
&= \frac{4[c(1-\delta\lambda_i)]^{j/4}-[c(1-\delta\lambda_i)]^{j/2}}{\lambda_i} \le \frac{4}{\lambda_i}[c(1-\delta\lambda_i)]^{j/4}, \qquad \text{(M.25)}
\end{aligned}$$

where the first inequality holds because $[c(1-\delta\lambda_i)]^{j/4-1/2} \le [c(1-\delta\lambda_i)]^{t/2}$ for all $t \le j/2-1$, the second inequality holds because $1-\sqrt{c(1-\delta\lambda_i)} \ge 1-\sqrt{c} \ge (1-c)/2$, and the last inequality holds because $[c(1-\delta\lambda_i)]^{j/2} \ge 0$. Due to the same deduction as that in part (b), we have

$$\begin{aligned}
\sum_{j=0}^{s-1}\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 &\le \frac{32}{(1-c)\lambda_i} \qquad \text{(M.26)}\\
&\le \frac{16N}{\lambda_i^2}.
\end{aligned}$$

(d) For $k^\dagger < i \le k^*$,

$$\begin{aligned}
&\sum_{j=0}^{s-1}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \\
&\le \sum_{j=0}^{s-1}\frac{9}{\lambda_i^2}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^N\right]^2\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{2j} \\
&\le \frac{9}{\lambda_i^2}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^N\right]^2\sum_{j=0}^{s-1}\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{j} \\
&= \frac{9(1-c)}{(q-c\delta)\lambda_i^3}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^N\right]^2\cdot\left[1-\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{s}\right] \\
&\le \frac{9(1-c)}{(q-c\delta)\lambda_i^3} \\
&\le \frac{9(1-c)}{(q-c\delta)\lambda_i^2}\cdot\frac{2N(q-c\delta)}{1-c} = \frac{18N}{\lambda_i^2},
\end{aligned}$$

where the second inequality holds because $\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{2j} \leq \left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{j}$, the third inequality holds because $1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^{N} \leq 1$ and $1-\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{s} \leq 1$, and the last inequality holds because $\lambda_i \geq \frac{1-c}{2N(q-c\delta)}$ due to definition of k^*.

(e) For $i > k^*$,

$$\sum_{j=0}^{s-1}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \leq \sum_{j=0}^{s-1}\frac{9}{\lambda_i^2}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^N\right]^2\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{2j}$$
$$\leq \frac{9}{\lambda_i^2}\cdot\left(2N\frac{q-c\delta}{1-c}\lambda_i\right)^2\sum_{j=0}^{s-1}1 = \frac{36sN^2(q-c\delta)^2}{(1-c)^2},$$

where the second inequality holds because $1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^{N} \leq 2N\frac{q-c\delta}{1-c}\lambda_i$ and $\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{2j} \leq 1$.

Concluding all the above, we have

$$\begin{aligned}
&\sum_{i=1}^{d}\lambda_i^2\sum_{j=0}^{s-1}\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2\\
&\leq \sum_{i\leq k^\ddagger}\lambda_i^2\cdot\frac{2N}{\lambda_i^2} + \sum_{k^\ddagger<i\leq k^\dagger}\lambda_i^2\cdot\frac{16N}{\lambda_i^2} + \sum_{k^\dagger<i\leq k^*}\lambda_i^2\cdot\frac{18N}{\lambda_i^2} + \sum_{i>k^*}\lambda_i^2\cdot\frac{36sN^2(q-c\delta)^2}{(1-c)^2}\\
&= 2Nk^\ddagger + 16N(k^\dagger-k^\ddagger) + 18N(k^*-k^\dagger) + \frac{36sN^2(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2\\
&\leq 18Nk^* + \frac{36sN^2(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2,
\end{aligned}$$

where the second inequality holds because all coefficients $2, 16, 18$ are bounded by 18. ■

Corollary M.7 *With $\mathbf{A}_i$ defined in (G.1), we have for all $j \geq 0$,*

$$\sum_{i=1}^{d}\lambda_i^2\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \leq 9k^* + \frac{36(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2.$$

Proof By Lemma M.5, we have

(a) For $i \leq k^\ddagger$,

$$\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \leq \frac{4}{\lambda_i^2}(c\delta/q)^{2j} \leq \frac{4}{\lambda_i^2},$$

where the second inequality holds because $c\delta/q \leq 1$.

(b) For $k^\ddagger < i \le \widehat{k}$,

$$\begin{aligned}
&\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \delta j[c(1-\delta\lambda_i)]^{(j-1)/2} \\
&\le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \delta \cdot \sum_{t=0}^{j-1}[c(1-\delta\lambda_i)]^{t/2} \\
&= \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \delta \cdot \frac{1-[c(1-\delta\lambda_i)]^{j/2}}{1-\sqrt{c(1-\delta\lambda_i)}} \\
&\le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \delta \cdot \frac{1-[c(1-\delta\lambda_i)]^{j/2}}{\delta\lambda_i/2} \\
&= \frac{2+[c(1-\delta\lambda_i)]^{j/2}}{\lambda_i} \le \frac{3}{\lambda_i},
\end{aligned} \tag{M.27}$$

where the first inequality holds because $[c(1-\delta\lambda_i)]^{(j-1)/2} \le [c(1-\delta\lambda_i)]^{t/2}$ for $t \le j-1$, the second inequality holds because $1-\sqrt{c(1-\delta\lambda_i)} \ge 1-\sqrt{1-\delta\lambda_i} \ge \delta\lambda_i/2$, and the last inequality holds because $[c(1-\delta\lambda_i)]^{j/2} \le 1$. Therefore,

$$\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \le \left(\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \delta j[c(1-\delta\lambda_i)]^{(j-1)/2}\right)^2 \le \frac{9}{\lambda_i^2},$$

where the first inequality hold due to Lemma M.5, and the second inequality holds due to (M.27).

(c) For $\widehat{k} < i \le k^\dagger$,

$$\begin{aligned}
&\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{1-c}{\lambda_i}\cdot j[c(1-\delta\lambda_i)]^{(j-1)/2} \\
&\le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{1-c}{\lambda_i} \cdot \sum_{t=0}^{j-1}[c(1-\delta\lambda_i)]^{t/2} \\
&= \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{1-c}{\lambda_i} \cdot \frac{1-[c(1-\delta\lambda_i)]^{j/2}}{1-\sqrt{c(1-\delta\lambda_i)}} \\
&\le \frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{1-c}{\lambda_i} \cdot \frac{1-[c(1-\delta\lambda_i)]^{j/2}}{(1-c)/2} \\
&= \frac{2+[c(1-\delta\lambda_i)]^{j/2}}{\lambda_i} \le \frac{3}{\lambda_i},
\end{aligned} \tag{M.28}$$

where the first inequality holds because $[c(1-\delta\lambda_i)]^{(j-1)/2} \le [c(1-\delta\lambda_i)]^{t/2}$ for $t \le j-1$, the second inequality holds because $1-\sqrt{c(1-\delta\lambda_i)} \ge 1-\sqrt{c} \ge (1-c)/2$, and the last inequality holds because $[c(1-\delta\lambda_i)]^{j/2} \le 1$. Therefore

$$\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{j+k}\begin{bmatrix}\delta\\ q\end{bmatrix}\right)_1^2 \le \left(\frac{3}{\lambda_i}[c(1-\delta\lambda_i)]^{j/2} + \frac{1-c}{\lambda_i}\cdot j[c(1-\delta\lambda_i)]^{(j-1)/2}\right)^2 \le \frac{9}{\lambda_i^2},$$

where the first inequality holds due to Lemma M.5, and the second inequality holds due to (M.28).

(d) For $i > k^*$, we have

$$\left(\sum_{k=0}^{N-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1^2 \le \frac{9}{\lambda_i^2}\left[1 - \left(1 - 2\frac{q - c\delta}{1-c}\lambda_i\right)^N\right]^2 \left(1 - \frac{q - c\delta}{1-c}\right)^{2j}$$
$$\le \frac{9}{\lambda_i^2}\left[1 - \left(1 - 2\frac{q - c\delta}{1-c}\lambda_i\right)^N\right]^2$$
$$\le \min\left\{\frac{9}{\lambda_i^2}, \frac{36N^2(q-c\delta)^2}{(1-c)^2}\right\},$$

where the second inequality holds because $1 - \frac{q-c\delta}{1-c}\lambda_i \le 1$, and the last inequality holds because $1-(1-r)^N \le 1$ and $1-(1-r)^N \le rN$ for all $r \in (0,1)$.

Combining all the above, we have

$$\sum_{i=1}^{d} \lambda_i^2 \left(\sum_{k=0}^{N-1} \mathbf{A}_i^{j+k} \begin{bmatrix} \delta \\ q \end{bmatrix}\right)_1^2$$
$$\le \sum_{i<k^\ddagger} \lambda_i^2 \cdot \frac{4}{\lambda_i^2} + \sum_{k^\ddagger < i \le k^\dagger} \lambda_i^2 \cdot \frac{9}{\lambda_i^2} + \sum_{k^\dagger < i \le k^*} \lambda_i^2 \cdot \frac{9}{\lambda_i^2} + \sum_{i>k^*} \lambda_i^2 \cdot \frac{36N^2(q-c\delta)^2}{(1-c)^2}$$
$$= 4k^\ddagger + 9(k^\dagger - k^\ddagger) + 9(k^* - k^\dagger) + \frac{36N^2(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2,$$
$$\le 9k^* + \frac{36N^2(q-c\delta)^2}{(1-c)^2}\sum_{i>k^*}\lambda_i^2,$$

where the first inequality holds because the bound $\frac{9}{\lambda_i^2}$ is applied for $k^\dagger < i \le k^*$, while the upper bound $\frac{36N^2(q-c\delta)^2}{(1-c)^2}$ is applied for $i > k^*$, and the second inequality holds because coefficient $4, 9$ can be bounded by 9. ■

Lemma M.8 *With $\mathbf{A}_i$ defined as in (G.1), let x_1 and x_2 be the eigenvalues of $\mathbf{A}_i$ defined in (G.2) and (G.3). Then*

- *For all $i \le k^\ddagger$, we have*

$$-\frac{4}{\delta\lambda_i}(c\delta/q)^j \le \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)_1 \le \frac{2}{\delta\lambda_i}(c\delta/q)^j.$$

- *For all $k^\ddagger < i \le \widehat{k}$, we have*

$$\left|\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)_1\right| \le 2j[c(1-\delta\lambda_i)]^{j/2} + \frac{4}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{j/2}.$$

- *For all $\widehat{k} < i \le k^\dagger$, we have*

$$\left| \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_1 \right| \le 2j[c(1-\delta\lambda_i)]^{j/2} + \frac{10}{1-c}[c(1-\delta\lambda_i)]^{j/2}.$$

- *For all $i > k^\dagger$, we have*

$$0 \le \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_1 \le \frac{3(1-c)}{(q-c\delta)\lambda_i} \left[1 - \left(1 - 2\frac{q-c\delta}{1-c}\lambda_i \right)^t \right] \left(1 - \frac{q-c\delta}{1-c}\lambda_i \right)^j.$$

Proof Note that

$$\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (\mathbf{A}_i^j - \mathbf{A}_i^{j+t})(\mathbf{I} - \mathbf{A}_i)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (\mathbf{A}_i^j - \mathbf{A}_i^{j+t}) \cdot \frac{1}{(q-c\delta)\lambda_i} \begin{bmatrix} 1-c+(q-\delta)\lambda_i \\ 1-c \end{bmatrix}$$
$$= \frac{1}{(q-c\delta)\lambda_i} \begin{bmatrix} (1-c+(q-\delta)\lambda_i)((\mathbf{A}_i^j)_{11} - (\mathbf{A}_i^{j+t})_{11}) + (1-c)(\mathbf{A}_i^j)_{12} - (\mathbf{A}_i^{j+t})_{12}) \\ (1-c+(q-\delta)\lambda_i)((\mathbf{A}_i^j)_{21} - (\mathbf{A}_i^{j+t})_{21}) + (1-c)((\mathbf{A}_i^j)_{22} - (\mathbf{A}_i^{j+t})_{22}). \end{bmatrix} \tag{M.29}$$

Combine (M.29) with Lemma G.3, and we have

$$\begin{aligned} \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_1 &= \frac{1}{(q-c\delta)\lambda_i} \Bigg[-(1-c+(q-\delta)\lambda_i) \cdot \frac{x_1 x_2^j(1-x_2^t) - x_2 x_1^j (1-x_1^t)}{x_2 - x_1} \\ &\quad + (1-c)(1-\delta\lambda_i) \frac{x_2^j(1-x_2^t) - x_1^j(1-x_1^t)}{x_2-x_1} \Bigg] \\ &= \frac{1}{(q-c\delta)\lambda_i} \Bigg\{ \frac{[(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_1] x_2^j (1-x_2^t)}{x_1 - x_2} \\ &\quad - \frac{[(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2] x_1^j (1-x_1^t)}{x_2 - x_1} \Bigg\}. \end{aligned} \tag{M.30}$$

For $i \le k^\ddagger$, note that

$$(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_1 = -(1-c)(x_1 + \delta\lambda_i - 1) - (q-\delta)\lambda_i x_1 \le 0$$

due to (M.8) and $q - \delta \ge 0$. The upper bound of (M.30) is thus given by

$$\begin{aligned} \left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_1 &\le \frac{1}{(q-c\delta)\lambda_i} \Bigg\{ \frac{[(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_1] x_1^j (1-x_2^t)}{x_1 - x_2} \\ &\quad - \frac{[(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2] x_1^j (1-x_1^t)}{x_2 - x_1} \Bigg\} \\ &= \frac{x_1^j}{(q-c\delta)\lambda_i} \Big\{ (1-c+(q-\delta)\lambda_i)(1-x_1^t) \end{aligned}$$

$$
\begin{aligned}
&\quad + [(1 - c + (q - \delta)\lambda_i)x_1 - (1-c)(1-\delta\lambda_i)] \cdot \frac{x_2^t - x_1^t}{x_2 - x_1}\bigg\} \\
&\leq \frac{x_1^j}{(q - c\delta)\lambda_i}\bigg\{(1 - c + (q-\delta)\lambda_i)(1 - x_1^t) \\
&\quad + [(1 - c + (q - \delta)\lambda_i)x_1 - (1-c)(1-\delta\lambda_i)] \cdot \frac{1 - x_1^t}{1 - x_1}\bigg\} \\
&= \frac{x_1^j(1 - x_1^t)}{(q - c\delta)\lambda_i} \cdot \frac{(q - c\delta)\lambda_i}{1 - x_1} = \frac{x_1^j(1 - x_1^t)}{1 - x_1} \\
&\leq \frac{2}{\delta\lambda_i}(c\delta/q)^j,
\end{aligned}
$$

where the first inequality holds because $x_2 \geq x_1$, the second inequality holds due to Lemma M.12, and the last inequality holds because $x_1 \leq x_2 \leq c\delta/q$, $1 - x_1^t \leq 1$ and (M.10). The lower bound of (M.30) is given by

$$
\begin{aligned}
\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1 &\geq \frac{1}{(q - c\delta)\lambda_i}\left\{\frac{-[(1 - c + (q-\delta)\lambda_i)x_1 - (1-c)(1-\delta\lambda_i)]x_2^j(1 - x_2^t)}{x_1 - x_2}\right. \\
&\quad \left. - \frac{[(1 - c + (q-\delta)\lambda_i)x_2 - (1-c)(1-\delta\lambda_i)]x_1^j(1 - x_2^t)}{x_2 - x_1}\right\} \\
&= \frac{1 - x_2^t}{(q - c\delta)\lambda_i}\bigg\{(1-c)(1-\delta\lambda_i)x_1^{j-1} \\
&\quad - [(1 - c + (q-\delta)\lambda_i)x_1 - (1-c)(1-\delta\lambda_i)]x_2 \cdot \frac{x_2^{j-1} - x_1^{j-1}}{x_2 - x_1}, \qquad \text{(M.31)}
\end{aligned}
$$

where the inequality holds because $x_1 \leq x_2$. If $j \geq 1$, then

$$
\begin{aligned}
&(1-c)(1-\delta\lambda_i)x_1^{j-1} - [(1 - c + (q-\delta)\lambda_i)x_1 - (1-c)(1-\delta\lambda_i)]x_2 \cdot \frac{x_2^{j-1} - x_1^{j-1}}{x_2 - x_1} \\
&\geq \frac{(1-c)(1-\delta\lambda_i)x_2 - (1 - c + (q-\delta)\lambda_i)x_1x_2}{c\delta/q - x_1} \cdot (c\delta/q)^{j-1} \\
&= \frac{(1-c)(1-\delta\lambda_i)x_2 - (1 - c + (q-\delta)\lambda_i) \cdot c(1-\delta\lambda_i)}{c\delta/q - x_1} \cdot (c\delta/q)^{j-1} \\
&= -(1-\delta\lambda_i)\frac{(1-c)(c - x_2) + c(q-\delta)\lambda_i}{c\delta/q - x_1} \cdot (c\delta/q)^{j-1}, \qquad \text{(M.32)}
\end{aligned}
$$

where the ineuqality holds because $(1-c)(1-\delta\lambda_i)x_1^{j-1} \geq 0$ and due to Lemma M.12. Note that

$$
\begin{aligned}
&(1-\delta\lambda_i)\frac{(1-c)(c - x_2) + c(q-\delta)\lambda_i}{c\delta/q - x_1} \\
&= (1-\delta\lambda_i)\frac{(1-c)\frac{-1+c+q\lambda_i-\sqrt{(1+c-q\lambda_i)^2-4c(1-\delta\lambda_i)}}{2} + c(q-\delta)\lambda_i}{c\delta/q - \frac{1+c-q\lambda_i-\sqrt{(1+c-q\lambda_i)^2-4c(1-\delta\lambda_i)}}{2}}
\end{aligned}
$$

$$
\begin{aligned}
&\leq (1-\delta\lambda_i)\frac{(1-c)\frac{-1+c+q\lambda_i}{2}+c(q-\delta)\lambda_i}{c\delta/q-\frac{1+c-q\lambda_i}{2}} = q(1-\delta\lambda_i)\frac{[(1+c)q-2c\delta]\lambda_i-(1-c)^2}{q^2\lambda_i-[(1+c)q-2c\delta]} \\
&\leq q\left[1-\delta\cdot\frac{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}{q^2}\right]\cdot\frac{[(1+c)q-2c\delta]\frac{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}{q^2}-(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2-[(1+c)q-2c\delta]} \\
&= \frac{(c\delta-\sqrt{c(q-\delta)(q-c\delta)})^2(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}{cq^3} \\
&\leq \frac{(c\delta)^2\cdot 4(q-c\delta)}{cq^3}\leq\frac{4c(q-c\delta)}{q},
\end{aligned}
\tag{M.33}
$$

where the first inequality holds because $\sqrt{(1+c-q\lambda_i)^2-4c(1-\delta\lambda_i)}\geq 0$, the second inequality holds due to (G.6), the third inequality holds because $c\delta-\sqrt{c(q-\delta)(q-c\delta)}\leq c\delta$ and $c(q-\delta)\leq q-c\delta$, and the last inequality holds because $\delta\leq q$. Combining (M.33) with (M.32) and (M.31), we have

$$
\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1 \geq -\frac{1-x_2^t}{(q-c\delta)\lambda_i}\cdot\frac{4c(q-c\delta)}{q}(c\delta/q)^{j-1}\geq -\frac{4}{\delta\lambda_i}(c\delta/q)^j,
$$

where the second inequality holds because $1-x_2^t\leq 1$.

For $k^\ddagger < i\leq k^\dagger$, i.e., $\mathbf{A}_i$ has complex eigenvalues x_1, x_2, we have

$$
\begin{aligned}
&\left|\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1\right| = \frac{1}{(q-c\delta)\lambda_i}\Bigg|(1-c)(1-\delta\lambda_i)x_2^{j-1}(1-x_2^t) \\
&\quad + [(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2]\cdot x_1\cdot\frac{(x_2^{j-1}-x_1^{j-1})(1-x_2^t)-x_1^{j-1}(x_2^t-x_1^t)}{x_2-x_1}\Bigg| \\
&\leq \frac{1}{(q-c\delta)\lambda_i}|(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2|\cdot|x_1| \\
&\quad\cdot\left[\left|\frac{x_2^{j-1}-x_1^{j-1}}{x_2-x_1}\right|\cdot|1-x_2^t|+|x_1^{j-1}|\cdot\left|\frac{x_2^t-x_1^t}{x_2-x_1}\right|\right]+\frac{(1-c)(1-\delta\lambda_i)}{(q-c\delta)\lambda_i}\cdot|x_2^{j-1}|\cdot|1-x_2^t|,
\end{aligned}
\tag{M.34}
$$

where the inequality holds due to triangle inequality. Note that

$$
\begin{aligned}
&|(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2| \\
&= \sqrt{[(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2][(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_1]},
\end{aligned}
$$

where

$$
\begin{aligned}
&[(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2][(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_1] \\
&= (1-c)^2(1-\delta\lambda_i)^2-(1-c)(1-\delta\lambda_i)(1-c+(q-\delta)\lambda_i)(x_1+x_2) \\
&\quad + (1-c+(q-\delta)\lambda_i)^2\cdot x_1x_2 \\
&= (1-c)^2(1-\delta\lambda_i)^2-(1-c)(1-\delta\lambda_i)(1-c+(q-\delta)\lambda_i)(1+c-q\lambda_i) \\
&\quad + (1-c+(q-\delta)\lambda_i)^2\cdot c(1-\delta\lambda_i)
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\delta\lambda_i)(q-\delta)(q-c\delta)\lambda_i^2 \\
&\le c(1-\delta\lambda_i)(q-c\delta)^2\lambda_i^2,
\end{aligned}
$$

where the inequality holds because $q-\delta \le c(q-c\delta)$. Therefore,

$$
|(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2| \le (q-c\delta)\lambda_i\sqrt{c(1-\delta\lambda_i)}. \tag{M.35}
$$

(M.34) can thus be further bounded by

$$
\begin{aligned}
&\left|\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1\right| \\
&\le \sqrt{c(1-\delta\lambda_i)}\cdot|x_1|\cdot\left[2\left|\frac{x_2^{j-1}-x_1^{j-1}}{x_2-x_1}\right| + |x_1|^{j-1}\cdot\left|\frac{x_2^t-x_1^t}{x_2-x_1}\right|\right] + \frac{2(1-c)(1-\delta\lambda_i)}{(q-c\delta)\lambda_i}\cdot|x_2^{j-1}| \\
&\le \left\{t[c(1-\delta\lambda_i)]^{t/2} + \frac{2(1-c)}{(q-c\delta)\lambda_i}\sqrt{\frac{1-\delta\lambda_i}{c}} + 2(j-1)\right\}\cdot[c(1-\delta\lambda_i)]^{j/2},
\end{aligned} \tag{M.36}
$$

where the first inequality holds because $|1-x_2^t|\le 2$ and due to (M.33), and the second inequality holds due to Lemma G.2 and Lemma M.13.

For $k^\ddagger < i \le \widehat{k}$, we can further bound (M.36) as

$$
\begin{aligned}
\left|\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1\right| &\le \left[\frac{2\sqrt{c(1-\delta\lambda_i)}}{\delta\lambda_i} + \frac{2(1-c)}{(q-c\delta)\lambda_i}\cdot\sqrt{\frac{1-\delta\lambda_i}{c}} + 2(j-1)\right][c(1-\delta\lambda_i)]^{j/2} \\
&\le \left[\frac{2}{\delta\lambda_i} + \frac{2}{\delta\lambda_i}\cdot 1 + 2j\right][c(1-\delta\lambda_i)]^{j/2} \\
&= 2j[c(1-\delta\lambda_i)]^{j/2} + \frac{4}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{j/2},
\end{aligned}
$$

where the first inequality holds due to Lemma M.14, and the second inequality holds because $\sqrt{c(1-\delta\lambda_i)}\le 1$, $(1-c)/(q-c\delta)\le 1/\delta$, $1-\delta\lambda_i\le c$ and $j-1<j$.

For $\widehat{k} < i \le k^\dagger$, note that

$$
\begin{aligned}
&\frac{1-c}{(q-c\delta)\lambda_i}\sqrt{\frac{1-\delta\lambda_i}{c}} \\
&\le \frac{1-c}{\sqrt{c}(q-c\delta)}\cdot\frac{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}{(1-c)^2}\cdot\sqrt{1-\frac{\delta(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}} \\
&= \frac{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})(\sqrt{q-\delta}+\sqrt{c(q-c\delta)})}{\sqrt{c}(1-c)(q-c\delta)} \\
&\le \frac{(1+c)\sqrt{q-c\delta}\cdot 2\sqrt{c(q-c\delta)}}{\sqrt{c}(1-c)(q-c\delta)} \\
&\le \frac{4}{1-c},
\end{aligned} \tag{M.37}
$$

where the first inequality holds due to (G.5), the second inequality holds because $q-\delta \le c(q-c\delta)$, and the third inequality holds because $c \le 1$. Therefore, (M.36) can be further bounded by

$$\begin{aligned}
\left|\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1\right| &\le \left[\frac{2\sqrt{c(1-\delta\lambda_i)}}{1-c}+\frac{2(1-c)}{(q-c\delta)\lambda_i}\sqrt{\frac{1-\delta\lambda_i}{c}}+2(j-1)\right][c(1-\delta\lambda_i)]^{j/2}\\
&\le \left[\frac{2}{1-c}+\frac{8}{1-c}+2j\right][c(1-\delta\lambda_i)]^{j/2}\\
&= 2j[c(1-\delta\lambda_i)]^{j/2}+\frac{10}{1-c}[c(1-\delta\lambda_i)]^{j/2},
\end{aligned}$$

where the first inequality holds due to Lemma M.12, and the second inequality holds because $\sqrt{c(1-\delta\lambda_i)} \le 1$, $j-1<j$ and due to (M.37).

For $j=0$ and $t \ge 1$, we have

$$\begin{aligned}
&\left|\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1\right|\\
&= \frac{1}{(q-c\delta)\lambda_i}\left|\frac{[(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_1](1-x_2^t)}{x_2-x_1}\right.\\
&\quad\left.-\frac{[(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2](1-x_1^t)}{x_2-x_1}\right|\\
&= \frac{1}{(q-c\delta)\lambda_i}\left|(1-c+(q-\delta)\lambda_i)-(1-c)(1-\delta\lambda_i)x_1^{t-1}\right.\\
&\quad\left.-[(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_1]x_2\cdot\frac{x_2^{t-1}-x_1^{t-1}}{x_2-x_1}\right|\\
&\le \frac{1}{(q-c\delta)\lambda_i}\left[(1-c)+(q-\delta)\lambda_i+(1-c)(1-\delta\lambda_i)|x_1^{t-1}|\right.\\
&\quad\left.+|(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_1|\cdot|x_2|\cdot\left|\frac{x_2^{t-1}-x_1^{t-1}}{x_2-x_1}\right|\right]\\
&\le \frac{1-c}{(q-c\delta)\lambda_i}+\frac{q-\delta}{q-c\delta}+\frac{(1-c)(1-\delta\lambda_i)}{(q-c\delta)\lambda_i}[c(1-\delta\lambda_i)]^{(t-1)/2}\\
&\quad+c(1-\delta\lambda_i)\cdot(t-1)[c(1-\delta\lambda_i)]^{(t-2)/2},
\end{aligned} \tag{M.38}$$

where the first inequality holds due to triangle inequality, and the second inequality holds due to Lemma M.13. When $k^\ddagger < i \le \widehat{k}$, (M.38) can be further bounded by

$$\begin{aligned}
&\left|\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1\right|\\
&\le \frac{1-c}{(q-c\delta)\lambda_i}+\frac{q-\delta}{q-c\delta}+\frac{(1-c)(1-\delta\lambda_i)}{(q-c\delta)\lambda_i}+\frac{c(1-\delta\lambda_i)}{\delta\lambda_i}
\end{aligned}$$

$$\le \frac{1-c}{(q-c\delta)\lambda_i} + 1 + \frac{1-c}{(q-c\delta)\lambda_i} + \frac{1}{\delta\lambda_i}$$
$$\le \frac{1}{\delta\lambda_i} + \frac{1}{\delta\lambda_i} + \frac{1}{\delta\lambda_i} + \frac{1}{\delta\lambda_i} = \frac{4}{\delta\lambda_i},$$

where the first inequality holds due to Lemma M.14, the second inequality holds because $(q - c\delta)/(1-c) \ge \delta$, $1-\delta\lambda_i \le 1$ and $c \le 1$, and the last inequality holds because $\frac{q-c\delta}{1-c} \ge \delta$ and $\delta\lambda_i \le 1$. When $\widehat{k} < i \le k^\dagger$, (M.38) can be further bounded by

$$\begin{aligned}
&\left|\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1\right| \\
&\le \frac{1-c}{(q-c\delta)\lambda_i} + \frac{q-\delta}{q-c\delta} + \frac{(1-c)(1-\delta\lambda_i)}{(q-c\delta)\lambda_i} + \frac{c(1-\delta\lambda_i)}{1-c} \\
&\le \frac{1-c}{(q-c\delta)\lambda_i} + 1 + \frac{1-c}{(q-c\delta)\lambda_i} + \frac{1}{1-c} \\
&\le \frac{2(1-c)}{(q-c\delta)} \cdot \frac{(\sqrt{q-c\delta} + \sqrt{c(q-\delta)})^2}{(1-c)^2} + \frac{1}{1-c} + \frac{1}{1-c} \\
&= \frac{2}{1-c} \cdot \left(1 + \sqrt{\frac{c(q-\delta)}{q-c\delta}}\right)^2 + \frac{2}{1-c} \\
&\le \frac{2}{1-c} \cdot (1+1)^2 + \frac{2}{1-c} = \frac{10}{1-c},
\end{aligned}$$

where the first inequality holds due to Lemma M.12, the second inequality holds because $q-\delta \le q - c\delta$, $1-\delta\lambda_i \le 1$ and $c \le 1$, the third inequality holds due to (G.5), and the last inequality holds because $c(q-\delta) \le q - c\delta$. Therefore, the upper bounds hold for $j = 0$.

For $i > k^\dagger$, note that

$$\begin{aligned}
&(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2 \\
&\ge (1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)\left(1 - \frac{q-c\delta}{1-c}\cdot\lambda_i\right) \\
&= \frac{(q-\delta)(q-c\delta)}{1-c}\lambda_i^2 \ge 0, \qquad \text{(M.39)}
\end{aligned}$$

where the first inequality holds due to Lemma G.2, and the second inequality holds because $q - c\delta \ge q - \delta \ge 0$. We thus have

$$\begin{aligned}
\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1 &\le \frac{1}{(q-c\delta)\lambda_i}\left\{\frac{[(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_1]x_2^j(1-x_2^t)}{x_1 - x_2}\right. \\
&\qquad \left. - \frac{[(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2]x_1^j(1-x_2^t)}{x_2 - x_1}\right\} \\
&= \frac{1-x_2^t}{(q-c\delta)\lambda_i}\left\{(1-c)(1-\delta\lambda_i)x_2^{j-1}\right.
\end{aligned}$$

$$+ [(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2]x_1 \cdot \frac{x_2^{j-1} - x_1^{j-1}}{x_2 - x_1}\Bigg\}, \quad \text{(M.40)}$$

where the inequality holds due to (M.39) and $x_1 \le x_2$. If $j \ge 1$, (M.40) is further bounded by

$$\begin{aligned}
&(1-c)(1-\delta\lambda_i)x_2^{j-1} + [(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2]x_1 \cdot \frac{x_2^{j-1} - x_1^{j-1}}{x_2 - x_1} \\
&\le \left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^{j-1}\left[(1-c)(1-\delta\lambda_i) + \frac{(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2}{1 - \frac{q-c\delta}{1-c}\lambda_i - x_1} \cdot x_1\right] \\
&= \left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^{j-1} \\
&\quad \cdot \left[(1-c)(1-\delta\lambda_i) + \frac{(1-c)(1-\delta\lambda_i)x_1 - (1-c+(q-\delta)\lambda_i)\cdot c(1-\delta\lambda_i)}{1 - \frac{q-c\delta}{1-c}\lambda_i - x_1}\right] \\
&= \left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^{j} \cdot \frac{1-\delta\lambda_i}{1 - \frac{q-c\delta}{1-c}\lambda_i} \cdot \frac{(1-c)^2 - [(1+c)q - 2c\delta]\lambda_i}{1 - \frac{q-c\delta}{1-c}\lambda_i - x_1}, \qquad \text{(M.41)}
\end{aligned}$$

where the inequality holds due to Lemma M.12 and Lemma G.2. We already have

$$\frac{1-\delta\lambda_i}{1 - \frac{q-c\delta}{1-c}\lambda_i} \le \frac{4}{3+c}$$

by (M.19). We also have

$$\begin{aligned}
&\frac{(1-c)^2 - [(1+c)q - 2c\delta]\lambda_i}{1 - \frac{q-c\delta}{1-c}\lambda_i - x_1} = \frac{(1-c)^2 - [(1+c)q - 2c\delta]\lambda_i}{1 - \frac{q-c\delta}{1-c}\lambda_i - \frac{(1+c-q\lambda_i) - \sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}}{2}} \\
&= 2(1-c) \cdot \frac{(1-c)^2 - [(1+c)q - 2c\delta]\lambda_i}{(1-c)^2 - [(1+c)q - 2c\delta]\lambda_i + (1-c)\sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)}} \\
&\le 2(1-c),
\end{aligned}$$

where the inequality holds because $(1-c)\sqrt{(1+c-q\lambda_i)^2 - 4c(1-\delta\lambda_i)} \ge 0$. We thus have

$$\frac{1-\delta\lambda_i}{1 - \frac{q-c\delta}{1-c}\lambda_i} \cdot \frac{(1-c)^2 - [(1+c)q - 2c\delta]\lambda_i}{1 - \frac{q-c\delta}{1-c}\lambda_i - x_1} \le \frac{4}{3+c} \cdot 2(1-c) \le \frac{8}{3}(1-c) \le 3(1-c), \quad \text{(M.42)}$$

where the second inequality holds because $c \ge 0$, and the last inequality holds because $8/3 < 3$. Combining (M.42) with (M.40) and (M.41), we have

$$\begin{aligned}
\left(\sum_{k=0}^{t-1} \mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1 &\le \frac{3(1-c)}{(q-c\delta)\lambda_i}\left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^j (1 - x_2^t) \\
&\le \frac{3(1-c)}{(q-c\delta)\lambda_i}\left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^j \left[1 - \left(1 - 2\frac{q-c\delta}{1-c}\lambda_i\right)^t\right],
\end{aligned}$$

where the second inequality holds due to Lemma G.2. For $j = 0$, we have

$$
\begin{aligned}
&[(1-c)(1-\delta\lambda_i) - (1-c+(q-\delta)\lambda_i)x_2]\frac{x_2^0 - x_1^0}{x_2 - x_1} + (1-c+(q-\delta)\lambda_i)x_2^0 \\
&= 1-c+(q-\delta)\lambda_i \le 1-c+(q-\delta)\cdot\frac{(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2} \\
&= (1-c)\frac{(q-\delta)+(q-c\delta)+2\sqrt{c(q-\delta)(q-c\delta)}}{c(q-\delta)+(q-c\delta)+2\sqrt{c(q-\delta)(q-c\delta)}} \\
&\le (1-c)\frac{(q-\delta)+(q-\delta)/c+2(q-\delta)}{c(q-\delta)+(q-\delta)/c+2(q-\delta)} \\
&= (1-c)\left[\frac{1}{1+c}+\frac{2}{(1+c)^2}\right] \le 3(1-c),
\end{aligned}
$$

where the first inequality holds due to (G.6), the second inequality holds because $q - c\delta \ge (q-\delta)/c$, and the last inequality holds because $c \ge 0$. Therefore, the upper bound also holds for $j = 0$.

The lower bound of (M.30) is given by

$$
\begin{aligned}
\left(\sum_{k=0}^{t-1}\mathbf{A}_i^{j+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1 &\ge \frac{1}{(q-c\delta)\lambda_i}\left\{\frac{[(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_1]x_2^j(1-x_2^t)}{x_1-x_2}\right. \\
&\qquad\left. - \frac{[(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2]x_2^j(1-x_1^t)}{x_2-x_1}\right\} \\
&= \frac{x_2^j}{(q-c\delta)\lambda_i}\left\{(1-c+(q-\delta)\lambda_i)(1-x_2^t)\right. \\
&\qquad\left. - [(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2]\frac{x_2^t-x_1^t}{x_2-x_1}\right\} \\
&\ge \frac{x_2^j}{(q-c\delta)\lambda_i}\left\{(1-c+(q-\delta)\lambda_i)(1-x_2^t)\right. \\
&\qquad\left. - [(1-c)(1-\delta\lambda_i)-(1-c+(q-\delta)\lambda_i)x_2]\frac{1-x_2^t}{1-x_2}\right\} \\
&= \frac{x_2^j(1-x_2^t)}{1-x_2} \ge 0,
\end{aligned}
$$

where the first inequality holds due to (M.39) and $x_1 < x_2$, the second inequality holds due to Lemma M.12, and the third inequality holds because $0 < x_2 < 1$. ■

The following corollary follows from Lemma M.8.

Corollary M.9 *With $\mathbf{A}_i$ defined in (G.1), we have*

$$
\begin{aligned}
&\sum_{i=1}^d \lambda_i w_i^2\left(\sum_{k=0}^{N-1}\mathbf{A}_i^{s+k}\begin{bmatrix}1\\1\end{bmatrix}\right)_1^2 \\
&\le \frac{16}{\delta^2}(c\delta/q)^{2s}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{0:k^\ddagger}^{-1}}^2 + 8s^2c^s\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|_{\mathbf{H}_{k^\ddagger:k^\dagger}}^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{32}{\delta^2} \cdot c^s \|(\mathbf{I} - \delta \mathbf{H})^{s/2}(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{k^\ddagger:\widehat{k}}} + \frac{200}{(1-c)^2} \cdot c^s \|(\mathbf{I} - \delta \mathbf{H})^{s/2}(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}} \\
&+ \frac{9(1-c)^2}{(q-c\delta)^2} \left\| \left(\mathbf{I} - \frac{q-c\delta}{1-c}\mathbf{H}\right)^s (\mathbf{w}_0 - \mathbf{w}^*) \right\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}} \\
&+ 36N^2 \left\| \left(\mathbf{I} - \frac{q-c\delta}{1-c}\mathbf{H}\right)^s (\mathbf{w}_0 - \mathbf{w}^*) \right\|^2_{\mathbf{H}_{k^*\infty}}
\end{aligned}
$$

Proof By Lemma M.8, we have

(a) For $i \le k^\ddagger$, we have

$$
\left(\sum_{k=0}^{N-1} \mathbf{A}_i^{s+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)_1^2 \le \frac{16}{\delta^2 \lambda_i^2} (c\delta/q)^{2s}.
$$

(b) For $k^\ddagger < i \le \widehat{k}$, we have

$$
\begin{aligned}
\left(\sum_{k=0}^{N-1} \mathbf{A}_i^{s+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)_1^2 &\le \left(2s[c(1-\delta\lambda_i)]^{s/2} + \frac{4}{\delta\lambda_i}[c(1-\delta\lambda_i)]^{s/2}\right)^2 \\
&\le 8s^2[c(1-\delta\lambda_i)]^s + \frac{32}{\delta^2\lambda_i^2}[c(1-\delta\lambda_i)]^s,
\end{aligned}
$$

where the inequality holds due to Cauchy-Schwarz inequality.

(c) For $\widehat{k} < i \le k^\dagger$, we have

$$
\begin{aligned}
\left(\sum_{k=0}^{N-1} \mathbf{A}_i^{s+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)_1^2 &\le \left(2s[c(1-\delta\lambda_i)]^{s/2} + \frac{10}{1-c}[c(1-\delta\lambda_i)]^{s/2}\right)^2 \\
&\le 8s^2[c(1-\delta\lambda_i)]^s + \frac{200}{(1-c)^2}[c(1-\delta\lambda_i)]^s,
\end{aligned}
$$

where the inequality holds due to Cauchy-Schwarz inequality.

(d) For $i > k^\dagger$, we have

$$
\begin{aligned}
\left(\sum_{k=0}^{N-1} \mathbf{A}_i^{s+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)_1^2 &\le \frac{9(1-c)^2}{(q-c\delta)^2\lambda_i^2} \left[1 - \left(1 - 2\frac{q-c\delta}{1-c}\lambda_i\right)^N\right]^2 \left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^{2s} \\
&\le \min\left\{\frac{9(1-c)^2}{(q-c\delta)^2\lambda_i^2}, 36N^2\right\} \left(1 - \frac{q-c\delta}{1-c}\lambda_i\right)^{2s},
\end{aligned}
$$

where the second inequality holds because $1-(1-r)^N \le 1$ and $1-(1-r)^N \le rN$ hold for all $r \in (0,1)$.

Concluding all the above, we have

$$
\sum_{i=1}^d \lambda_i w_i^2 \left(\sum_{k=0}^{N-1} \mathbf{A}_i^{s+k} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)_1^2
$$

$$
\begin{aligned}
&\leq \sum_{i\leq k^\ddagger} \lambda_i w_i^2 \cdot \frac{16}{\delta^2\lambda_i^2}(c\delta/q)^{2s} + \sum_{k^\ddagger<i\leq \widehat{k}} \lambda_i w_i^2 \cdot \left(8s^2[c(1-\delta\lambda_i)]^s + \frac{32}{\delta^2\lambda_i^2}[c(1-\delta\lambda_i)]^s\right) \\
&\quad + \sum_{\widehat{k}<i\leq k^\dagger} \lambda_i w_i^2 \cdot \left(8s^2[c(1-\delta\lambda_i)]^s + \frac{200}{(1-c)^2}[c(1-\delta\lambda_i)]^s\right) \\
&\quad + \sum_{k^\dagger<i\leq k^*} \lambda_i w_i^2 \cdot \frac{9(1-c)^2}{(q-c\delta)^2\lambda_i^2}\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{2s} + \sum_{i>k^*} \lambda_i w_i^2 \cdot 36N^2\left(1-\frac{q-c\delta}{1-c}\lambda_i\right)^{2s} \\
&= \frac{16}{\delta^2}(c\delta/q)^{2s}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^\ddagger}} + 8s^2c^s\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^\ddagger:k^\dagger}} \\
&\quad + \frac{32}{\delta^2}\cdot c^s\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{k^\ddagger:\widehat{k}}} + \frac{200}{(1-c)^2}\cdot c^s\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}} \\
&\quad + \frac{9(1-c)^2}{(q-c\delta)^2}\left\|\left(\mathbf{I}-\frac{q-c\delta}{1-c}\mathbf{H}\right)^s(\mathbf{w}_0-\mathbf{w}^*)\right\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}} \\
&\quad + 36N^2\left\|\left(\mathbf{I}-\frac{q-c\delta}{1-c}\mathbf{H}\right)^s(\mathbf{w}_0-\mathbf{w}^*)\right\|^2_{\mathbf{H}_{k^*\infty}},
\end{aligned}
$$

where the first inuequality holds because the upper bound $\frac{9(1-c)^2}{(q-c\delta)^2\lambda_i^2}$ is applied for $k^\dagger < i \leq k^*$ and $36N^2$ is applied for $i > k^*$. ■

Lemma M.10 *With $\mathbf{A}_i$ defined in (G.1), let x_1 and x_2 be the eigenvalues of $\mathbf{A}_i$ as defined in (G.2) and (G.3). Then*

- *For all $i \leq k^\ddagger$, we have*
$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \leq \frac{7}{2\delta\lambda_i};$$
- *For all $k^\ddagger < i \leq \widehat{k}$, we have*
$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \leq \frac{14}{\delta\lambda_i};$$
- *For all $\widehat{k} < i \leq k^\dagger$, we have*
$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \leq \frac{10}{1-c};$$
- *For all $i > k^\dagger$, we have*
$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \leq \frac{1-c}{(q-c\delta)\lambda_i}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^{2t}\right].$$

Proof Note that

$$\left(\mathbf{A}_i^k \begin{bmatrix}1\\1\end{bmatrix}\right)_2 = (\mathbf{A}_i^k)_{21} + (\mathbf{A}_i^k)_{22} = -c\frac{x_2^k - x_1^k}{x_2 - x_1} + \frac{x_2^{k+1} - x_1^{k+1}}{x_2 - x_1}$$
$$= \frac{(x_2 - c)x_2^k - (x_1 - c)x_1^k}{x_2 - x_1}, \tag{M.43}$$

where the second equality holds due to Lemma G.3. Summing up the square of (M.43) yields

$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k \begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 = \sum_{k=0}^{t-1}\left[\frac{(x_2 - c)x_2^k - (x_1 - c)x_1^k}{x_2 - x_1}\right]^2$$
$$= \sum_{k=0}^{t-1}\frac{(x_2 - c)^2 x_2^{2k}}{(x_2 - x_1)^2} - 2\sum_{k=0}^{t-1}\frac{(x_2 - c)(x_1 - c)(x_1 x_2)^k}{(x_2 - x_1)^2} + \sum_{k=0}^{t-1}\frac{(x_1 - c)^2 x_1^{2k}}{(x_2 - x_1)^2}$$
$$= \frac{(x_2 - c)^2(1 - x_2^{2t})}{(1 - x_2^2)(x_2 - x_1)^2} - 2\frac{(x_2 - c)(x_1 - c)[1 - (x_1 x_2)^t]}{(1 - x_1 x_2)(x_2 - x_1)^2} + \frac{(x_1 - c)^2(1 - x_1^{2t})}{(1 - x_1^2)(x_2 - x_1)^2}. \tag{M.44}$$

Denote

$$A := \frac{(x_2 - c)^2}{1 - x_2^2}, \quad B := \frac{(x_1 - c)(x_2 - c)}{1 - x_1 x_2}, \quad C := \frac{(x_1 - c)^2}{1 - x_1^2},$$

then we have

$$\frac{A - B}{x_2 - x_1} = \frac{(x_2 - c)(1 - c x_2)}{(1 - x_2^2)(1 - x_1 x_2)}, \quad \frac{B - C}{x_2 - x_1} = \frac{(x_1 - c)(1 - c x_1)}{(1 - x_1^2)(1 - x_1 x_2)},$$
$$\frac{A - 2B + C}{(x_2 - x_1)^2} = \frac{(1 + c^2)(1 + x_1 x_2) - 2c(x_1 + x_2)}{(1 - x_1^2)(1 - x_2^2)(1 - x_1 x_2)}.$$

For all $i \le k^{\ddagger}$, (M.44) is bounded by

$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k \begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 = (1 - x_2^{2t}) \cdot \frac{A - 2B + C}{(x_2 - x_1)^2} + 2\frac{C - B}{x_2 - x_1} \cdot x_2^t \cdot \frac{x_2^t - x_1^t}{x_2 - x_1} - C \cdot \left(\frac{x_2^t - x_1^t}{x_2 - x_1}\right)^2$$
$$\le (1 - x_2^{2t}) \cdot \frac{A - 2B + C}{(x_2 - x_1)^2} + 2x_2 \cdot \frac{C - B}{x_2 - x_1} \cdot \frac{x_2^{2t} - (x_1 x_2)^t}{x_2^2 - x_1 x_2}$$
$$\le \frac{A - 2B + C}{(x_2 - x_1)^2} + \frac{2x_2(C - B)}{x_2 - x_1} \cdot \frac{1}{1 - x_1 x_2}$$
$$= \frac{(1 + c^2)(1 + x_1 x_2) - 2c(x_1 + x_2)}{(1 - x_1^2)(1 - x_2^2)(1 - x_1 x_2)} + \frac{2x_2(c - x_1)(1 - c x_1)}{(1 - x_1^2)(1 - x_1 x_2)^2}, \tag{M.45}$$

where the first inequality holds because $C\left(\frac{x_2^t - x_1^t}{x_2 - x_1}\right)^2 \ge 0$, and the second inequality holds because due to Lemma M.12. Note that

$$\frac{(1 + c^2)(1 + x_2 x_2) - 2c(x_1 + x_2)}{(1 - x_1^2)(1 - x_2^2)}$$
$$= \frac{(1 + c)^2}{2(1 + x_1)(1 + x_2)} + \frac{(1 - c)^2}{2(1 - x_1)(1 - x_2)} \le \frac{(1 + c)^2}{2} + \frac{(1 - c)^2}{2(q - c\delta)\lambda_i}$$

$$\leq \frac{(1+c)^2}{2} + \frac{(\sqrt{q-c\delta} - \sqrt{c(q-\delta)})^2}{2(q-c\delta)} \leq \frac{(1+1)^2}{2} + \frac{q-c\delta}{2(q-c\delta)} = \frac{5}{2}, \tag{M.46}$$

where the first inequality holds because $1+x_2 \geq 1+x_1 \geq 1$, the second inequality holds due to (G.6), and the last inequality holds because $\sqrt{q-c\delta} - \sqrt{c(q-\delta)} \leq \sqrt{q-c\delta}$. We also have

$$\begin{aligned}
\frac{(c-x_1)(1-cx_1)x_2}{1-x_1^2} &\leq (c-x_1)x_2 = \frac{(c-x_1)(c-x_2)x_2}{c-x_2} = \frac{c(q-\delta)\lambda_i \cdot x_2}{c-x_2} \\
&\leq \frac{c(q-\delta)\lambda_i \cdot \frac{c\delta - \sqrt{c(q-\delta)(q-c\delta)}}{q}}{c - \frac{c\delta - \sqrt{c(q-\delta)(q-c\delta)}}{q}} = \frac{\sqrt{c(q-\delta)}(c\delta - \sqrt{c(q-\delta)(q-c\delta)})}{\sqrt{q-c\delta} + \sqrt{c(q-\delta)}}\lambda_i \\
&\leq \frac{\sqrt{c(q-\delta)} \cdot \delta}{\sqrt{c(q-\delta)} + \sqrt{c(q-\delta)}}\lambda_i = \frac{\delta\lambda_i}{2},
\end{aligned} \tag{M.47}$$

where the first inequality holds because $1-cx_1 \leq 1-x_1^2$ (due to the fact that $x_1 \leq x_2 \leq c\delta/q \leq c$), the second inequality holds due to Lemma G.2, and the last inequality holds because $\sqrt{q-c\delta} \geq \sqrt{c(q-\delta)}$ and $c\delta - \sqrt{c(q-\delta)(q-c\delta)} \leq c\delta \leq \delta$. We finally have

$$1 - x_1 x_2 = 1 - c(1-\delta\lambda_i) \geq \delta\lambda_i. \tag{M.48}$$

Substituting (M.46), (M.47) and (M.48) into (M.45), we have

$$\sum_{k=0}^{t-1} \left(\mathbf{A}_i^k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_2^2 \leq \frac{5}{2} \cdot \frac{1}{\delta\lambda_i} + \delta\lambda_i \cdot \frac{1}{(\delta\lambda_i)^2} = \frac{7}{2\delta\lambda_i}.$$

For $k^\ddagger < i \leq k^\dagger$, (M.44) can be bounded as

$$\begin{aligned}
&\sum_{k=0}^{t-1} \left(\mathbf{A}_i^k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_2^2 \\
&= (1-(x_1x_2)^t) \cdot \frac{A-2B+C}{(x_2-x_1)^2} - \frac{A+C}{2} \cdot \left(\frac{x_2^t - x_1^t}{x_2 - x_1} \right)^2 + \frac{C-A}{2(x_2-x_1)} \cdot \frac{x_2^{2t} - x_1^{2t}}{x_2 - x_1} \\
&\leq |1-(x_1x_2)^t| \cdot \left| \frac{A-2B+C}{(x_2-x_1)^2} \right| + \frac{|A+C|}{2} \cdot \left| \frac{x_2^t - x_1^t}{x_2 - x_1} \right|^2 + \frac{1}{2} \cdot \left| \frac{C-A}{x_2-x_1} \right| \cdot \left| \frac{x_2^{2t} - x_1^{2t}}{x_2 - x_1} \right| \\
&\leq \left| \frac{A-2B+C}{(x_2-x_1)^2} \right| + \frac{|A+C|}{2} \cdot (t[(1-\delta\lambda_i)]^{(t-1)/2})^2 + \frac{1}{2} \cdot \left| \frac{C-A}{x_2-x_1} \right| \cdot 2t[c(1-\delta\lambda_i)]^{(2t-1)/2} \\
&= \frac{1}{1-x_1x_2} \cdot \left| \frac{(1+c^2)(1+x_1x_2) - 2c(x_1+x_2)}{(1-x_1^2)(1-x_2^2)} \right| + \frac{|A+C|}{2} \cdot (t[(1-\delta\lambda_i)]^{(t-1)/2})^2 \\
&\quad + \left| \frac{2c(1+x_1x_2) - (1+c^2)(x_1+x_2)}{2(1-x_1^2)(1-x_2^2)} \right| \cdot 2t[c(1-\delta\lambda_i)]^{(2t-1)/2},
\end{aligned} \tag{M.49}$$

where the first inequality holds due to triangle inequality, and the second inequality holds because $0 \leq 1-(x_1x_2)^t \leq 1$ and due to Lemma M.13. We now bound the coefficients. Note that

$$\frac{(1-c)^2}{(1-x_1)(1-x_2)} = \frac{(1-c)^2}{(q-c\delta)\lambda_i} \leq \frac{(\sqrt{q-c\delta} + \sqrt{c(q-\delta)})^2}{q-c\delta} \leq \frac{(\sqrt{q-c\delta} + \sqrt{q-c\delta})^2}{q-c\delta} = 4$$

where the first inequality holds due to (G.5), and the second inequality holds because $c(q-\delta) \le q - c\delta$. We also note that

$$\frac{(1+c)^2}{(1+x_1)(1+x_2)} = \frac{(1+c)^2}{2(1+c)-(q+c\delta)\lambda_i} \le \frac{(1+c)^2}{2(1+c)-(1+2c)\delta\lambda_i} \le \frac{(1+c)^2}{2(1+c)-(1+2c)} \le \frac{(1+1)^2}{1} = 4,$$

where the first equality holds due to Lemma M.2(c), the first inequality holds because $q \le (1+c)\delta$, the second inequality holds because $\delta\lambda_i \le 1$, and the last inequality holds because $c \le 1$. We thus have

$$\left|\frac{(1+c^2)(1+x_1x_2)-2c(x_1+x_2)}{(1-x_1^2)(1-x_2^2)}\right| = \frac{1}{2}\left|\frac{(1+c)^2}{(1+x_1)(1+x_2)} + \frac{(1-c)^2}{(1-x_1)(1-x_2)}\right| \le 4, \quad \text{(M.50)}$$

$$\left|\frac{2c(1+x_1x_2)-(1+c^2)(x_1+x_2)}{(1-x_1^2)(1-x_2^2)}\right| = \frac{1}{2}\left|\frac{(1+c)^2}{(1+x_1)(1+x_2)} - \frac{(1-c)^2}{(1-x_1)(1+x_2)}\right| \le 2. \quad \text{(M.51)}$$

We then bound $|A+C|/2$. Note that

$$\begin{aligned}
\frac{A+C}{2} &= \frac{1}{2}\left[\frac{(x_1-c)^2}{1-x_1^2} + \frac{(x_2-c)^2}{1-x_2^2}\right] = \frac{(x_1-c)^2(1-x_2^2)+(x_2-c)^2(1-x_1^2)}{2(1-x_1^2)(1-x_2^2)} \\
&= \frac{2c^2-2c(x_1+x_2)+(1-c^2)(x_1^2+x_2^2)+2cx_1x_2(x_1+x_2)-2x_1^2x_2^2}{2(q-c\delta)\lambda_i\cdot[2(1+c)-(q+c\delta)\lambda_i]} \\
&= \frac{(1+c)(1-c)^3-2(1-c)[(1+c+c^2)q-c(1+2c)\delta]\lambda_i+[(1-c^2)q^2+2c^2\delta q-2c^2\delta^2]\lambda_i^2}{2(q-c\delta)\lambda_i\cdot[2(1+c)-(q+c\delta)\lambda_i]}.
\end{aligned} \quad \text{(M.52)}$$

For $k^\ddagger < i \le \widehat{k}$, we aim to bound the denominator of (M.52) by λ_i^2 multiplied by a constant. Denote the denominator divided by λ_i^2 as

$$\phi(\lambda_i) := \frac{(1+c)(1-c)^3}{\lambda_i^2} - \frac{2(1-c)[(1+c+c^2)q-c(1+2c)\delta]}{\lambda_i} + [(1-c^2)q^2+2c^2\delta q-2c^2\delta^2],$$

then

$$\begin{aligned}
\frac{\partial\phi}{\partial\frac{1}{\lambda_i}} &= 2(1-c)\left[\frac{(1+c)(1-c)^2}{\lambda_i} - [(1+c+c^2)q-c(1+2c)\delta]\right] \\
&\le 2(1-c)\left[\frac{(1+c)(1-c)^2}{(1-c)/\delta} - [(1+c+c^2)q-c(1+2c)\delta]\right] \\
&= -2(1-c)(1+c+c^2)(q-\delta) \le 0,
\end{aligned}$$

where the second inequality holds due to (G.7), and the last inequality holds because $q \ge \delta$, so ϕ is a decreasing function in $1/\lambda_i$. We thus have

$$\begin{aligned}
&\phi(\lambda_i) \ge \phi((1-c)/\delta) \\
&= \frac{(1+c)(1-c)^3}{(1-c)^2/\delta^2} - \frac{2(1-c)[(1+c+c^2)q-c(1+2c)\delta]}{(1-c)/\delta} + [(1-c^2)q^2+2c^2\delta q-2c^2\delta^2]
\end{aligned}$$

$$
\begin{aligned}
&= (1+c)(q-\delta)[(1-c)q-(1+c)\delta] \\
&\geq (1+c)(q-\delta)[(1-c)\delta-(1+c)\delta] = -2c\delta(1+c)(q-\delta) \\
&\geq -4\delta(q-c\delta),
\end{aligned}
$$

where the first inequality holds because $\lambda_i \geq (1-c)/\delta$, the second inequality holds because $q \geq \delta$, and the last inequality holds because $c \leq 1$ and $q-\delta \leq q-c\delta$. We also note that $2(1+c)-(q+c\delta)\lambda_i \geq 1$, so $(A+C)/2 \leq 2\delta\lambda_i$. We also have

$$
\begin{aligned}
\phi(\lambda_i) &\leq \phi\left(\frac{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}{q^2}\right), \\
2(1+c)-(q+c\delta)\lambda_i &\geq 2(1+c)-(q+c\delta)\cdot\frac{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}{q^2},
\end{aligned}
$$

so the upper bound of $\frac{A+C}{2}$ is

$$
\begin{aligned}
\frac{A+C}{2} &\leq \frac{(c(q-\delta)+\sqrt{c(q-\delta)(q-c\delta)})^2}{1-(c\delta-\sqrt{c(q-\delta)(q-c\delta)})^2/q^2}\cdot\frac{\lambda_i}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2} \\
&\leq \frac{(c(q-\delta)+\sqrt{c(q-\delta)(q-c\delta)})^2}{1-(c\delta-\sqrt{c(q-\delta)(q-c\delta)})/q}\cdot\frac{\lambda_i}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2} \\
&= \frac{q\cdot c(q-\delta)\lambda_i}{(q-c\delta)+\sqrt{c(q-\delta)(q-c\delta)}} \\
&\leq \frac{2\delta(q-c\delta)\lambda_i}{q-c\delta} = 2\delta\lambda_i,
\end{aligned}
$$

where the second inequality holds because $(c\delta-\sqrt{c(q-\delta)(q-c\delta)})/q \leq 1$, and the last inequality holds because $c(q-\delta) \leq q-c\delta$, $q \leq (1+c)\delta \leq 2\delta$ and $\sqrt{c(q-\delta)(q-c\delta)} \geq 0$. Therefore,

$$
\frac{|A+C|}{2} \leq 2\delta\lambda_i, \tag{M.53}
$$

where the second inequality holds because $2(1+c)-(q+c\delta)\lambda_i \geq 1$. For $\widehat{k} < i \leq k^\dagger$, we aim to bound the denominator of (M.52) as λ_i multiplied by a constant. Denote the denominator devided by λ_i as

$$
\varphi(\lambda_i) := \frac{(1+c)(1-c)^3}{\lambda_i} - 2(1-c)[(1+c+c^2)q-c(1+2c)\delta] + [(1-c^2)q^2+2c^2\delta q-2c^2\delta^2]\lambda_i,
$$

then the lower bound of $\varphi(\lambda_i)$ is given by

$$
\begin{aligned}
\varphi(\lambda_i) &\geq -2(1-c)[(1+c+c^2)q-c(1+2c)\delta] \\
&= -2(1-c)[(1+c)(q-c\delta)+c^2(q-\delta)] \\
&\geq -2(1-c)(1+c+c^2)(q-c\delta),
\end{aligned}
$$

where the first inequality holds because $\frac{(1+c)(1-c)^3}{\lambda_i} \geq 0$ and $[(1-c^2)q^2+2c^2\delta q-2c^2\delta^2]\lambda_i \geq 0$, and the second inequality holds because and $q-\delta \leq q-c\delta$. Note that the maximum of $\varphi(\lambda_i)$ is attained at either $\frac{1-c}{\delta}$ or $\frac{(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2}$. For the former, we have

$$
\begin{aligned}
\varphi((1-c)/\delta) &= \frac{(1+c)(1-c)^3}{(1-c)/\delta} - 2(1-c)[(1+c+c^2)q - c(1+2c)\delta] \\
&\quad + [(1-c^2)q^2 + 2c^2\delta q - 2c^2\delta^2] \cdot \frac{1-c}{\delta} \\
&= (1-c)(1+c)(q-\delta)[(1-c)q/\delta - (1+c)] \\
&\le (1-c)(1+c)(q-\delta)[(1-c)(1+c) - (1+c)] \\
&= -c(1-c)(1+c)(q-\delta) \le 0,
\end{aligned}
$$

where the first inequality holds because $q \le (1+c)\delta$, and the second inequality holds because $q \le \delta$. For the latter,

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

where the inequality holds because $q-\delta \le c(q-c\delta)$ and $c(q-\delta) \le q-c\delta$. We finally have

$$
2(1+c) - (q+c\delta)\lambda_i \ge 2(1+c) - (1+2c)\delta\lambda_i \ge 2(1+c) - (1+2c)(1-c) = 1+c+2c^2,
$$

where the first inequality holds because $q \le (1+c)\delta$, and the second inequality holds because $\delta\lambda_i \le 1-c$ (due to definition of $\widehat{k}$). Therefore,

$$
\frac{|A+C|}{2} \le \max\left\{\frac{1+c+c^2}{1+c+2c^2}, \frac{c^2}{1+c+2c^2}\right\} \cdot (1-c) \le (1-c), \tag{M.54}
$$

where the second inequality holds because $1+c+c^2 \le 1+c+2c^2$ and $c^2 \le 1+c+2c^2$.

Therefore, when $k^\ddagger < i \le \widehat{k}$, $1-x_1x_2 \ge \delta\lambda_i$, so (M.49) can be further bounded by

$$
\begin{aligned}
\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k \begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 &\le \frac{4}{\delta\lambda_i} + 2\delta\lambda_i \cdot (t[c(1-\delta\lambda_i)]^{(t-1)/2})^2 + \frac{1}{2}\cdot 2\cdot 2t[c(1-\delta\lambda_i)]^{(2t-1)/2} \\
&\le \frac{4}{\delta\lambda_i} + 2\delta\lambda_i \cdot \frac{4}{\delta^2\lambda_i^2} + \frac{2}{\delta\lambda_i} = \frac{14}{\delta\lambda_i},
\end{aligned}
$$

where the first inequality holds due to (M.50), (M.51) and (M.53), and the second inequality holds due to Lemma M.14. When $\widehat{k} < i \le k^\dagger$, $1-x_1x_2 \ge 1-c$, so (M.49) is further bounded by

$$
\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k \begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \le \frac{4}{1-c} + (1-c)\cdot([t(1-\delta\lambda_i)]^{(t-1)/2})^2 + \frac{1}{2}\cdot 2\cdot 2t[c(1-\delta\lambda_i)]^{(2t-1)/2}
$$

$$\leq \frac{4}{1-c} + (1-c) \cdot \frac{4}{(1-c)^2} + \frac{2}{1-c} = \frac{10}{1-c},$$

where the first inequality holds due to (M.50), (M.51), and the second inequality holds due to Lemma M.14.

For all $i > k^\dagger$, (M.44) can by bounded as

$$\begin{aligned}\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 &= (1-x_2^{2t})\cdot\frac{A-2B+C}{(x_2-x_1)^2} - 2\frac{B-C}{x_2-x_1}\cdot\frac{x_2^{2t}-(x_1x_2)^t}{x_2-x_1} - C\left(\frac{x_2^t-x_1^t}{x_2-x_1}\right)^2\\ &\leq (1-x_2^{2t})\cdot\frac{A-2B+C}{(x_2-x_1)^2} = (1-x_2^{2t})\frac{(1+c^2)(1+x_1x_2)-2c(x_1+x_2)}{(1-x_1^2)(1-x_2^2)(1-x_1x_2)},\end{aligned} \tag{M.55}$$

where the inequality holds because negative terms are dropped. Note that

$$\frac{(1-c)^2}{(1-x_1)(1-x_2)} = \frac{(1-c)^2}{(q-c\delta)\lambda_i},$$

and

$$\frac{(1+c)^2}{(1+x_1)(1+x_2)} \leq \frac{(1+c)^2}{(1+c)^2} = 1 \leq \frac{(1-c)^2}{(\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2\lambda_i} \leq \frac{(1-c)^2}{(q-c\delta)\lambda_i},$$

where the first inequality holds because $c \leq x_1 \leq x_2$, the second inequality holds due to (G.5), and the last inequality holds because $\sqrt{c(q-\delta)} \geq 0$. We thus have

$$\begin{aligned}&\frac{(1+c^2)(1+x_1x_2)-2c(x_1+x_2)}{(1-x_1^2)(1-x_2^2)} = \frac{(1+c)^2}{2(1+x_1)(1+x_2)} + \frac{(1-c)^2}{2(1-x_1)(1-x_2)}\\ &\leq \frac{(1-c)^2}{2(q-c\delta)\lambda_i} + \frac{(1-c)^2}{2(q-c\delta)\lambda_i} = \frac{(1-c)^2}{(q-c\delta)\lambda_i}.\end{aligned} \tag{M.56}$$

We also have

$$1 - x_1x_2 = 1 - c + c\delta\lambda_i \geq 1 - c, \tag{M.57}$$

where the inequality holds because $c\delta\lambda_i \geq 0$. Substituting (M.56) and (M.57) into (M.55), we have

$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \leq (1-x_2^{2t})\cdot\frac{1-c}{(q-c\delta)\lambda_i} \leq \frac{1-c}{(q-c\delta)\lambda_i}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^{2t}\right],$$

where the second inequality holds due to Lemma G.2. ■

The following lemma follows from Lemma M.8.

Corollary M.11 *With $\mathbf{A}_i$ defined in (G.1), we have*

$$\begin{aligned}\sum_{i=1}^{d}\lambda_i w_i^2\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 &\leq \frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:\hat{k}}}^2 + \frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{\hat{k}:k^\dagger}}^2\\ &+ \frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{k^\dagger:k^*}}^2 + 4t\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2.\end{aligned}$$

Proof By Lemma M.8, specifically for $k^\dagger < i \le k^*$, we have

$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \le \frac{1-c}{(q-c\delta)\lambda_i}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^{2t}\right] \le \frac{1-c}{(q-c\delta)\lambda_i},$$

where the inequality holds because $1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^{2t}\le 1$; For $i > k^*$,

$$\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2 \le \frac{1-c}{(q-c\delta)\lambda_i}\left[1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^{2t}\right] \le \frac{1-c}{(q-c\delta)\lambda_i}\cdot 4t\frac{(q-c\delta)\lambda_i}{1-c} = 4t,$$

where the inequality holds because $1-\left(1-2\frac{q-c\delta}{1-c}\lambda_i\right)^{2t}\le 4t\frac{(q-c\delta)\lambda_i}{1-c}$. Therefore,

$$\begin{aligned}
&\sum_{i=1}^{d}\lambda_i w_i^2\sum_{k=0}^{t-1}\left(\mathbf{A}_i^k\begin{bmatrix}1\\1\end{bmatrix}\right)_2^2\\
&\le \sum_{i\le k^\ddagger}\lambda_i w_i^2\cdot\frac{7}{2\delta\lambda_i}+\sum_{k^\ddagger<i\le\widehat{k}}\lambda_i w_i^2\cdot\frac{14}{\delta\lambda_i}+\sum_{\widehat{k}<i\le k^\dagger}\lambda_i w_i^2\cdot\frac{10}{1-c}\\
&\quad+\sum_{k^\dagger<i\le k^*}\lambda_i w_i^2\cdot\frac{1-c}{(q-c\delta)\lambda_i}+\sum_{i>k^*}\lambda_i w_i^2\cdot 4t\\
&=\frac{7}{2\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\ddagger}}+\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{k^\ddagger:\widehat{k}}}+\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}}\\
&\quad+\frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{k^\dagger:k^*}}+4t\sum_{i>k^*}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\\
&\le\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:\widehat{k}}}+\frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}}\\
&\quad+\frac{1-c}{q-c\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{k^\dagger:k^*}}+4t\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}},
\end{aligned}$$

where the second inequality holds because $7/2 < 14$. ■

Lemma M.12 *For any* $0 < x_1, x_2 \le \theta < 1$ *(*$x_1 \ne x_2$*) and integer* $t \ge 0$*, we have*

$$\frac{x_2^t - x_1^t}{x_2 - x_1} \le \frac{\theta^t - x_1^t}{\theta - x_1}.$$

Proof The lemma holds trivially for $t = 0$. For $t \ge 1$, we have

$$\frac{x_2^t - x_1^t}{x_2 - x_1} = \sum_{k=0}^{t-1}x_1^k x_2^{t-1-k} \le \sum_{k=0}^{t-1}x_1^k\cdot\theta^{t-1-k} = \frac{\theta^t - x_1^t}{\theta - x_1},$$

where the inequality holds because $x_2 \le \theta$. ■

Lemma M.13 *Suppose x_1, x_2 are complex eigenvalues of $\mathbf{A}_i$ for $k^\ddagger < i \le k^\dagger$. Then for any $t \ge 0$,*

$$\left|\frac{x_2^t - x_1^t}{x_2 - x_1}\right| \le t[c(1-\delta\lambda_i)]^{(t-1)/2}.$$

Proof We have

$$\left|\frac{x_2^t - x_1^t}{x_2 - x_1}\right| = \left|\sum_{k=0}^{t-1} x_2^k x_1^{t-1-k}\right| \le \sum_{k=0}^{t-1} |x_2^k| \cdot |x_1^{t-1-k}| = t[c(1-\delta\lambda_i)]^{(t-1)/2},$$

where the inequality holds due to triangle inequality, and the second equality holds due to Lemma G.2. ■

Lemma M.14 *For any $t \ge 0$, we have*

$$t[c(1-\delta\lambda_i)]^{(t-1)/2} \le \min\left\{\frac{2}{\delta\lambda_i}, \frac{2}{1-c}\right\}.$$

Proof Note that

$$\begin{aligned} t[c(1-\delta\lambda_i)]^{(t-1)/2} &= \sum_{k=0}^{t-1}[c(1-\delta\lambda_i)]^{(t-1)/2} \le \sum_{k=0}^{t-1}[c(1-\delta\lambda_i)]^{k/2} = \frac{1-[c(1-\delta\lambda_i)]^{t/2}}{1-\sqrt{c(1-\delta\lambda_i)}} \\ &\le \frac{1}{1-\sqrt{c(1-\delta\lambda_i)}} \le \frac{1}{1-\sqrt{1-\delta\lambda_i}} \le \frac{2}{\delta\lambda_i}, \end{aligned}$$

where the first inequality holds because $c(1-\delta\lambda_i) \le 1$, the second inequality holds because $1 - [c(1-\delta\lambda_i)]^{t/2} \le 1$, the third inequality holds because $c \le 1$, and the last inequality holds because $1 - \sqrt{1-\delta\lambda_i} \ge \delta\lambda_i/2$. Similarly we have

$$t[c(1-\delta\lambda_i)]^{(t-1)/2} \le \frac{1}{1-\sqrt{c(1-\delta\lambda_i)}} \le \frac{1}{1-\sqrt{c}} \le \frac{2}{1-c},$$

where the second inequality holds because $1-\delta\lambda_i \le c$, and the last inequality holds because $1-\sqrt{c} \ge (1-c)/2$. ■