Trajectory of Mini-Batch Momentum:
Batch Size Saturation and Convergence in High Dimensions
Supplementary material

The appendix is organized into 4 sections as follows:

1. Appendix A derives the Volterra equation and proves the main concentration for the dynamics of SGD+M (Theorem 1).
2. We show in Appendix B that the error terms associated with concentration of measure on the high-dimensional orthogonal group disappear in the large-n limit.
3. Appendix C derives main results including Proposition 3 and speed up of convergence rate of SGD+M (Proposition 5) in the large batch regime, as well as the lower bound convergence rate in the small batch regime (Proposition 6). We also provide a proof of Proposition 7 in this section.
4. Appendix D contains details on the numerical simulations.

Potential societal impacts. The results presented in this paper concern the analysis of existing methods on a simple least squares problems. The results are theoretical and we do not anticipate any direct ethical and societal issues. We believe the results will be used by machine learning practitioners and we encourage them to use it to build a more just, prosperous world.

Notation. In this paper, we adhere whenever possible to the following notation. We denote vectors in lowercase boldface ($\boldsymbol{x}$) and matrices in upper boldface ($\boldsymbol{A}$). The entries of a vector (or matrix) are denoted by subscripts. Unless otherwise specified, the norm $\|\cdot\|_2$ is taken to be the standard Euclidean norm if it is applied to a vector and the operator 2-norm if it is applied to a matrix. For a matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$, we denote constants depending on $\boldsymbol{A}$ and $\boldsymbol{b}$, $C(\boldsymbol{A}, \boldsymbol{b})$, as those bounded by an absolute constant multiplied by $\|\boldsymbol{A}\|$ and $\|\boldsymbol{b}\|$. We say an event B holds *with overwhelming probability* (w.o.p.) if, for every fixed $D > 0$, $\Pr(B) \geq 1 - C_D d^{-D}$ for some C_D independent of d. Lastly, for $n \in \mathbb{N}$, $[n]$ denotes the set of natural numbers up to n, i.e., $[n] \stackrel{\text{def}}{=} \{1, 2, \cdots, n\}$.

A Derivation of the dynamics of SGD+M

In this section, we establish the fundamental of the proof of Theorem 1. Let us state the theorem in full detail first.

Theorem 3 (Theorem 1, detailed version)**.** *Suppose Assumptions 1.1 and 1.2 hold with the learning rate $\gamma < \frac{1+\Delta}{\zeta \sigma^2_{max}}$ and the batch size satisfies $\beta/n = \zeta$ for some $\zeta > 0$. Let the constant $T \in \mathbb{N}$. Then there exists $C > 0$ such that for any $c > 0$, there exists $D > 0$ satisfying*

$$\Pr\left[\sup_{0 \leq t \leq T, t \in \mathbb{N}} |f(\boldsymbol{x}_t) - \psi(t)| > n^{-C}\right] \leq D n^{-c}, \tag{28}$$

for sufficiently large $n \in \mathbb{N}$. $\boldsymbol{x}_t$ are the iterates of SGD+M and the function ψ is the solution to the Volterra equation

$$\psi(t+1) = \underbrace{\frac{R}{2} h_1(t+1) + \frac{\tilde{R}}{2} h_0(t+1)}_{\textit{forcing}} + \underbrace{\sum_{k=0}^{t} \gamma^2 \zeta (1-\zeta) H_2(t-k) \psi(k)}_{\textit{noise}}, \text{ and } \psi(0) = f(\boldsymbol{x}_0), \tag{29}$$

where for $k = 0, 1$,

$$h_k(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{2(\sigma_j^2)^k}{\Omega_j^2 - 4\Delta} \left(-\Delta \gamma \zeta (\sigma_j^2) \cdot \Delta^t + \frac{1}{2} (\kappa_{2,j} - \Delta)^2 \cdot (\lambda_{2,j})^t + \frac{1}{2} (\kappa_{3,j} - \Delta)^2 \cdot (\lambda_{3,j})^t \right),$$

and

$$H_2(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta} \Big(- \Delta^{t+1} + \frac{1}{2} \lambda_{2,j}^{t+1} + \frac{1}{2} \lambda_{3,j}^{t+1} \Big).$$

Here $\Omega_j, \lambda_{2,j}, \lambda_{3,j}, \kappa_{2,j}, \kappa_{3,j}, j \in [n]$ *are defined as*

$$\Omega_j = 1 - \gamma\zeta\sigma_j^2 + \Delta,\ \kappa_{2,j} = \frac{\lambda_{2,j}\Omega_j}{\lambda_{2,j}+\Delta},\ \kappa_{3,j} = \frac{\lambda_{3,j}\Omega_j}{\lambda_{3,j}+\Delta},\ \textit{and}$$
$$\lambda_{2,j} = \frac{-2\Delta + \Omega_j^2 + \sqrt{\Omega_j^2(\Omega_j^2 - 4\Delta)}}{2},\ \lambda_{3,j} = \frac{-2\Delta + \Omega_j^2 - \sqrt{\Omega_j^2(\Omega_j^2 - 4\Delta)}}{2}.$$

A.1 Change of basis

Consider the singular value decomposition of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, i.e. $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{\Sigma}$ is the $n \times d$ singular value matrix with diagonal entries $\mathbf{diag}(\sigma_j), j = 1, \ldots, n$ (in the case $n > d$, we extend the set of singular values so that $\sigma_{d+1} = \cdots = \sigma_n = 0$). We define the spectral weight vector $\boldsymbol{\nu}_k \overset{\text{def}}{=} \boldsymbol{V}^T(\boldsymbol{x}_k - \widetilde{\boldsymbol{x}})$, which therefore evolves like

$$\boldsymbol{\nu}_{k+1} = \boldsymbol{\nu}_k - \gamma\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_k(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_k - \boldsymbol{\eta}) + \Delta(\boldsymbol{\nu}_k - \boldsymbol{\nu}_{k-1}). \tag{30}$$

Moreover, we can define

$$\boldsymbol{w}_k := \boldsymbol{\Sigma}\boldsymbol{\nu}_k - \boldsymbol{U}^T\boldsymbol{\eta}, \tag{31}$$

so that

$$f(\boldsymbol{x}_t) = \frac{1}{2}\|\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta}\|_2^2 = \frac{1}{2}\sum_{j=1}^{n}\boldsymbol{w}_{t,j}^2. \tag{32}$$

Then (30) can be translated as

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \gamma\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_k\boldsymbol{U}\boldsymbol{w}_k + \Delta(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}). \tag{33}$$

From this point, we focus on the evolution of $\boldsymbol{w}$ rather than the iterates $\boldsymbol{x}$.

A.2 Evolution of f

Now we would like to demonstrate the recurrence relation of $\boldsymbol{w}_k$ and eventually that of $f(t)$, which will lead to a Volterra equation and error terms in a large scale. First, for $j \in [n]$ and $t \in \mathbb{N}$, (33) implies that

$$w_{t+1,j} = w_{t,j} - \gamma\sigma_j^2\sum_l w_{t,l}\Big(\sum_{i\in B_t} U_{ij}U_{il}\Big) + \Delta(w_{t,j} - w_{t-1,j}), \tag{34}$$

where $B_t = B$ denotes a randomly chosen mini-batch at the t-th iteration, whose size is given by $\beta \le n$. We interchangeably use the notation of B_t and B, because it is independently chosen at each iteration. By taking squares on both sides, we have

$$\begin{aligned}
w_{t+1,j}^2 &= \left(w_{t,j} - \gamma\sigma_j^2\sum_{l\in[n]} w_{t,l}\Big(\sum_{i\in B_t} U_{ij}U_{il}\Big) + \Delta(w_{t,j} - w_{t-1,j})\right)^2 \\
&= w_{t,j}^2 + \gamma^2\sigma_j^4\Big(\sum_{l\in[n]} w_{t,l}\Big(\sum_{i\in B_t} U_{ij}U_{il}\Big)\Big)^2 - 2\gamma\sigma_j^2 w_{t,j}\sum_{l\in[n]} w_{t,l}\Big(\sum_{i\in B_t} U_{ij}U_{il}\Big) \\
&\quad + \Delta^2(w_{t,j} - w_{t-1,j})^2 + 2\Delta w_{t,j}(w_{t,j} - w_{t-1,j}) \\
&\quad - 2\gamma\sigma_j^2\Delta\sum_{l\in[n]} w_{t,l}\Big(\sum_{i\in B_t} U_{ij}U_{il}\Big)(w_{t,j} - w_{t-1,j}).
\end{aligned}$$

Now let us denote the following error caused by mini-batching, i.e.,

$$\mathcal{E}_B^{(l,j)} \overset{\text{def}}{=} \sum_{i\in B} U_{il}U_{ij} - \frac{\beta}{n}\delta_{l,j}. \tag{35}$$

where $\delta_{l,j}$ is the Kronecker-delta symbol, meaning

$$\text{For } l, j \in [n],\ \delta_{l,j} = 1 \quad \text{if } l = j, \text{ and 0 otherwise.}$$

Then the iteration on w_{t+1}^2 reduces to

$$\begin{aligned}
w_{t+1,j}^2 &= w_{t,j}^2(1+\Delta^2+2\Delta) + w_{t-1,j}^2\Delta^2 + w_{t,j}w_{t-1,j}(-2\Delta^2-2\Delta) \\
&\quad - 2\gamma\sigma_j^2 w_{t,j}\sum_{l\in[n]} w_{t,l}(\mathcal{E}_B^{(l,j)} + \frac{\beta}{n}\delta_{l,j}) \\
&\quad - 2\gamma\sigma_j^2\Delta\sum_{l\in[n]} w_{t,l}(w_{t,j}-w_{t-1,j})(\mathcal{E}_B^{(l,j)} + \frac{\beta}{n}\delta_{l,j}) + \gamma^2\sigma_j^4\big(\sum_{l\in[n]} w_{t,l}(\sum_{i\in B} U_{ij}U_{il})\big)^2 \\
&= w_{t,j}^2(1+\Delta^2+2\Delta-2\gamma\sigma_j^2\frac{\beta}{n}-2\Delta\gamma\sigma_j^2\frac{\beta}{n}) + w_{t-1,j}^2\Delta^2 \\
&\quad + w_{t,j}w_{t-1,j}(-2\Delta^2-2\Delta+2\Delta\gamma\sigma_j^2\frac{\beta}{n}) \\
&\quad + \underbrace{\gamma^2\sigma_j^4\big(\sum_{l\in[n]} w_{t,l}(\sum_{i\in B} U_{ij}U_{il})\big)^2}_{\stackrel{\text{def}}{=}①} + \underbrace{\left(-2\gamma\sigma_j^2 w_{t,j}\sum_{l\in[n]}\mathcal{E}_B^{(l,j)}w_{t,l}\right)}_{\stackrel{\text{def}}{=}\mathcal{E}_{B,1}^{(j)}(t)} \\
&\quad + \underbrace{\left(-2\gamma\sigma_j^2\Delta\sum_{l\in[n]}\mathcal{E}_B^{(l,j)}w_{t,l}(w_{t,j}-w_{t-1,j})\right)}_{\stackrel{\text{def}}{=}\mathcal{E}_{B,2}^{(j)}(t)}.
\end{aligned}$$

When it comes to ①, we can decompose it into its expectation over the mini-batch B and the error generated by it. By applying the technique from [27, Lemma 8], we have

$$\begin{aligned}
\mathbb{E}[①|\mathcal{F}_t] &= \gamma^2\sigma_j^4\left[\frac{\beta(\beta-1)}{n(n-1)}w_{t,j}^2 + \left(\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{i\in[n]}U_{ij}^2\left(\sum_{l\in[n]}U_{il}w_{t,l}\right)^2\right] \\
&= \Gamma_j^2 w_{t,j}^2 + \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n}\sum_{l\in[n]}w_{t,l}^2 + \mathcal{E}_{KL}^{(j)}(t) + \mathcal{E}_{beta}^{(j)}(t),
\end{aligned}$$

where

$$\Gamma_j \stackrel{\text{def}}{=} \gamma\zeta\sigma_j^2,$$

$$\mathcal{E}_{beta}^{(j)}(t) \stackrel{\text{def}}{=} \gamma^2\sigma_j^4\left[\left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)w_{t,j}^2 + \left(-\frac{\beta(\beta-1)}{n(n-1)}+\zeta^2\right)\sum_{i\in[n]}U_{ij}^2\left(\sum_{l\in[n]}U_{il}w_{t,l}\right)^2\right],$$

$$\text{and}\quad \mathcal{E}_{KL}^{(j)}(t) \stackrel{\text{def}}{=} \gamma^2\sigma_j^4(\zeta-\zeta^2)\sum_{i\in[n]}(U_{ij}^2-\frac{1}{n})\left(\sum_l U_{il}w_{t,l}\right)^2.$$

Note that $\mathcal{E}_{beta}^{(j)}(t)$ is generated by the error between $\beta(\beta-1)/(n(n-1))$ and $\zeta^2=\beta^2/n^2$, whereas $\mathcal{E}_{KL}^{(j)}(t)$ is generated by the replacement of U_{ij}^2 by $1/n$; In Appendix B, we establish that this error can be bounded by the *Key Lemma* (this is where the acronym "KL" comes from). Let $\mathcal{E}_{B^2}^{(j)}(t) \stackrel{\text{def}}{=} ① - \mathbb{E}[①|\mathcal{F}_t]$. Then observe

$$① = \Gamma_j^2 w_{t,j}^2 + \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n}\sum_{l\in[n]}w_{t,l}^2 + \mathcal{E}_{B^2}^{(j)}(t) + \mathcal{E}_{beta}^{(j)}(t) + \mathcal{E}_{KL}^{(j)}(t).$$

Therefore, we obtain

$$\begin{aligned}
w_{t+1,j}^2 &= \Omega_j^2 w_{t,j}^2 + \Delta^2 w_{t-1,j}^2 - 2\Delta\Omega_j w_{t,j}w_{t-1,j} + \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n}\sum_{l\in[n]}w_{t,l}^2 \\
&\quad + \mathcal{E}_{beta}^{(j)}(t) + \mathcal{E}_{KL}^{(j)}(t) + \mathcal{E}_B^{(j)}(t),
\end{aligned} \tag{36}$$

where

$$\mathcal{E}_B^{(j)}(t) \stackrel{\text{def}}{=} \mathcal{E}_{B^2}^{(j)}(t) + \mathcal{E}_{B,1}^{(j)}(t) + \mathcal{E}_{B,2}^{(j)}(t).$$

Similarly, we have

$$\begin{aligned} w_{t+1,j}w_{t,j} &= w_{t,j}(w_{t,j} - \gamma\sigma_j^2 \sum_{l\in[n]} w_{t,l}(\sum_{i\in B_t} U_{ij}U_{il}) + \Delta(w_{t,j} - w_{t-1,j})) \\ &= w_{t,j}^2 - \gamma\sigma_j^2 w_{t,j} \sum_{l\in[n]} w_{t,l}(\mathcal{E}_B^{(l,j)} + \frac{\beta}{n}\delta_{l,j}) + \Delta w_{t,j}(w_{t,j} - w_{t-1,j}) \\ &= \Omega_j w_{t,j}^2 - \Delta w_{t,j}w_{t-1,j} - \underbrace{\gamma\sigma_j^2 w_{t,j} \sum_l \mathcal{E}_B^{(l,j)} w_{t,l}}_{=\frac{1}{2}\mathcal{E}_{B,1}^{(j)}(t)}, \end{aligned} \tag{37}$$

where $\Omega_j \stackrel{\text{def}}{=} 1 - \Gamma_j + \Delta$.

Therefore, (36) and (37) imply

$$\begin{pmatrix} w_{t+1,j}^2 \\ w_{t,j}^2 \\ w_{t+1,j}w_{t,j} \end{pmatrix} = \underbrace{\begin{pmatrix} \Omega_j^2 & \Delta^2 & -2\Delta\Omega_j \\ 1 & 0 & 0 \\ \Omega_j & 0 & -\Delta \end{pmatrix}}_{\stackrel{\text{def}}{=}\boldsymbol{M}_j} \underbrace{\begin{pmatrix} w_{t,j}^2 \\ w_{t-1,j}^2 \\ w_{t,j}w_{t-1,j} \end{pmatrix}}_{\stackrel{\text{def}}{=}\tilde{\mathcal{X}}_{t,j}} + \underbrace{\begin{pmatrix} \tilde{N}_{t,j} + \mathcal{E}_1^{(j)}(t) \\ 0 \\ \mathcal{E}_2^{(j)}(t) \end{pmatrix}}_{\stackrel{\text{def}}{=}\tilde{\mathcal{Y}}_{t,j}}, \tag{38}$$

where

$$\tilde{N}_{t,j} \stackrel{\text{def}}{=} \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n} \sum_l w_{t,l}^2 = \varphi_j^{(n)} \sum_l w_{t,l}^2, \text{ with } \varphi_j^{(n)} \stackrel{\text{def}}{=} \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n},$$

$$\mathcal{E}_1^{(j)}(t) \stackrel{\text{def}}{=} \mathcal{E}_{beta}^{(j)}(t) + \mathcal{E}_{KL}^{(j)}(t) + \mathcal{E}_B^{(j)}(t), \text{ and}$$

$$\mathcal{E}_2^{(j)}(t) \stackrel{\text{def}}{=} -\frac{1}{2}\mathcal{E}_{B,1}^{(j)}(t).$$

Let us rewrite (38) as

$$\begin{aligned} \tilde{\mathcal{X}}_{t+1,j} &= \boldsymbol{M}_j\tilde{\mathcal{X}}_{t,j} + \tilde{\mathcal{Y}}_{t,j} \\ &= \boldsymbol{M}_j^2\tilde{\mathcal{X}}_{t-1,j} + \boldsymbol{M}_j\tilde{\mathcal{Y}}_{t-1,j} + \tilde{\mathcal{Y}}_{t,j} \\ &= \boldsymbol{M}_j^t\tilde{\mathcal{X}}_{1,j} + \sum_{k=1}^{t} \boldsymbol{M}_j^{t-k}\tilde{\mathcal{Y}}_{k,j}. \end{aligned}$$

The eigendecomposition of $\boldsymbol{M}_j$ is given by $\boldsymbol{M}_j = \boldsymbol{X}_j\Lambda_j\boldsymbol{X}_j^{-1}$, $\Lambda_j = \mathbf{diag}(\lambda_{1,j}, \lambda_{2,j}, \lambda_{3,j})$, $\boldsymbol{X}_j = (x_{1,j}, x_{2,j}, x_{3,j})^T$ where

$$\lambda_{1,j} = \Delta, \lambda_{2,j} = \frac{-2\Delta + \Omega_j^2 + \sqrt{(\Omega_j^2)(\Omega_j^2 - 4\Delta)}}{2}, \lambda_{3,j} = \frac{-2\Delta + \Omega_j^2 - \sqrt{(\Omega_j^2)(\Omega_j^2 - 4\Delta)}}{2},$$
$$\text{and } \boldsymbol{X}_j = \begin{pmatrix} \Delta & \lambda_{2,j} & \lambda_{3,j} \\ 1 & 1 & 1 \\ \frac{\Omega_j}{2} & \kappa_{2,j} & \kappa_{3,j} \end{pmatrix} \text{ with } \kappa_{i,j} = \frac{\lambda_i\Omega_j}{\lambda_i + \Delta}. \tag{39}$$

Also, its inverse, assuming $\det \boldsymbol{X}_j \neq 0$, is given by

$$\boldsymbol{X}_j^{-1} = \frac{2}{\Omega_j^2 - 4\Delta} \begin{pmatrix} -1 & -\Delta & \Omega_j \\ \frac{1}{2} & \frac{\lambda_{3,j}}{2} & -\kappa_{3,j} \\ \frac{1}{2} & \frac{\lambda_{2,j}}{2} & -\kappa_{2,j} \end{pmatrix}.$$

Note that for each $j \in [n]$ and $i \in \{2, 3\}$, $\kappa_{i,j}$ satisfies

- $\kappa_{i,j}^2 = \lambda_{i,j}$ and $\kappa_{i,j} = \sqrt{\lambda_{i,j}}$ when $\Omega_j \geq 0$,
- $\kappa_{2,j} + \kappa_{3,j} = \Omega_j$, and
- $\kappa_{2,j}\kappa_{3,j} = \Delta$.

This implies that

$$\begin{aligned}\boldsymbol{M}_j^{t-k}\tilde{\mathcal{Y}}_{k,j} &= \boldsymbol{X}_j \Lambda_j^{t-k} \boldsymbol{X}_j^{-1} \begin{pmatrix} \tilde{N}_{k,j} + \mathcal{E}_1^{(j)}(k) \\ 0 \\ \mathcal{E}_2^{(j)}(k) \end{pmatrix} \\ &= \boldsymbol{X}_j \cdot \frac{2}{\Omega_j^2 - 4\Delta} \begin{pmatrix} -\lambda_{1,j}^{t-k}\left(\tilde{\mathcal{N}}_{k,j} + \mathcal{E}_1^{(j)}(k)\right) + \Omega_j \lambda_{1,j}^{t-k}\mathcal{E}_2^{(j)}(k) \\ \frac{1}{2}\lambda_{2,j}^{t-k}\left(\tilde{\mathcal{N}}_{k,j} + \mathcal{E}_1^{(j)}(k)\right) - \kappa_{3,j}\lambda_{2,j}^{t-k}\mathcal{E}_2^{(j)}(k) \\ \frac{1}{2}\lambda_{3,j}^{t-k}\left(\tilde{\mathcal{N}}_{k,j} + \mathcal{E}_1^{(j)}(k)\right) - \kappa_{2,j}\lambda_{3,j}^{t-k}\mathcal{E}_2^{(j)}(k) \end{pmatrix}.\end{aligned}$$

In particular, if we just focus on the (first coordinate of $\tilde{\mathcal{X}}_{t+1,j}) = w_{t+1,j}^2$, we have

$$\begin{aligned} w_{t+1,j}^2 = (\boldsymbol{M}_j^t \tilde{\mathcal{X}}_{1,j})_1 &+ \frac{2}{\Omega_j^2 - 4\Delta}\sum_{k=1}^{t}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\varphi_j^{(n)}\sum_{l\in[n]} w_{k,l}^2 \\ &+ \frac{2}{\Omega_j^2 - 4\Delta}\sum_{k=1}^{t}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_1^{(j)}(k) \\ &+ \frac{2}{\Omega_j^2 - 4\Delta}\sum_{k=1}^{t}(\Omega_j\lambda_{1,j}\cdot\lambda_{1,j}^{t-k} - \kappa_{3,j}\lambda_{2,j}\cdot\lambda_{2,j}^{t-k} - \kappa_{2,j}\lambda_{3,j}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_2^{(j)}(k)\end{aligned}$$

(Here $(\cdot)_1$ denotes the first coordinate of a vector). Summing over $j \in [n]$ and dividing both sides by 2 gives

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{n} w_{t+1,j}^2 &= \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{M}_j^t\tilde{\mathcal{X}}_{1,j})_1 \\ &+ \sum_{k=1}^{t}\left(\sum_{j=1}^{n}\frac{2\varphi_j^{(n)}}{\Omega_j^2 - 4\Delta}(-\Delta\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\right) f(k) \\ &+ \sum_{k=1}^{t}\left(\sum_{j=1}^{n}\frac{1}{\Omega_j^2 - 4\Delta}(-\Delta\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_1^{(j)}(k)\right) \\ &+ \sum_{k=1}^{t}\left(\sum_{j=1}^{n}\frac{1}{\Omega_j^2 - 4\Delta}(\Omega_j\lambda_{1,j}\cdot\lambda_{1,j}^{t-k} - \kappa_{3,j}\lambda_{2,j}\cdot\lambda_{2,j}^{t-k} - \kappa_{2,j}\lambda_{3,j}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_2^{(j)}(k)\right).\end{aligned} \tag{40}$$

Note that $\sum_{j=1}^{n}(\boldsymbol{M}_j^t\tilde{\mathcal{X}}_{1,j})_1$ describes the *forcing* term (see Section 2). In order to analyze this term, observe

$$\begin{aligned}\tilde{\mathcal{X}}_{1,j} &= \begin{pmatrix} w_{1,j}^2 \\ w_{0,j}^2 \\ w_{1,j}w_{0,j}\end{pmatrix} = \begin{pmatrix} (1-\Gamma_j)^2 w_{0,j}^2 + \varphi_j^{(n)}\sum_l w_{0,l}^2 + \mathcal{E}_{beta}^{(j)}(0) + \mathcal{E}_{KL}^{(j)}(0) + \mathcal{E}_B^{(j)}(0) \\ w_{0,j}^2 \\ (1-\Gamma_j)w_{0,j}^2 - \frac{1}{2}\mathcal{E}_{B,1}^{(j)}(0)\end{pmatrix} \\ &= \begin{pmatrix} (1-\Gamma_j)^2\left(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\right) + 2\varphi_j^{(n)}f(0) + \mathcal{E}_{beta}^{(j)}(0) + \mathcal{E}_{KL}^{(j)}(0) + \mathcal{E}_B^{(j)}(0) + (1-\Gamma_j)^2\mathcal{E}_{w_0}^{(j)} \\ \sigma_j^2\frac{R}{n} + \frac{\tilde{R}}{n} + \mathcal{E}_{w_0}^{(j)} \\ (1-\Gamma_j)\left(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\right) + (1-\Gamma_j)\mathcal{E}_{w_0}^{(j)} - \frac{1}{2}\mathcal{E}_{B,1}^{(j)}(0)\end{pmatrix} \\ &= \left(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\right)\begin{pmatrix}(1-\Gamma_j)^2 \\ 1 \\ (1-\Gamma_j)\end{pmatrix} + \begin{pmatrix}2\varphi_j^{(n)}f(0) + \mathcal{E}_1^{(j)}(0) \\ 0 \\ 0\end{pmatrix} + \mathcal{E}_{w_0}^{(j)}\begin{pmatrix}(1-\Gamma_j)^2 \\ 1 \\ (1-\Gamma_j)\end{pmatrix} + \mathcal{E}_2^{(j)}(0)\begin{pmatrix}0\\0\\1\end{pmatrix},\end{aligned}$$

where

$$\begin{aligned}\mathcal{E}_{w_0}^{(j)} &\stackrel{\text{def}}{=} w_{0,j}^2 - \mathbb{E}[w_{0,j}^2] = w_{0,j}^2 - \Big(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\Big),\\ \mathcal{E}_1^{(j)}(0) &= \mathcal{E}_{beta}^{(j)}(0) + \mathcal{E}_{KL}^{(j)}(0) + \mathcal{E}_B^{(j)}(0), \quad \text{and } \mathcal{E}_2^{(j)}(0) = -\frac{1}{2}\mathcal{E}_{B,1}^{(j)}(0).\end{aligned} \tag{41}$$

Therefore, by using the eigendecomposition of $\boldsymbol{M}_j$ again, the first coordinate of $\boldsymbol{M}_j^t\tilde{\mathcal{X}}_{1,j}$ is given by

$$\begin{aligned}(\boldsymbol{M}_j^t\tilde{\mathcal{X}}_{1,j})_1 &= \Bigg[\boldsymbol{X}_j\Lambda_j^t\begin{pmatrix}-(1-\Gamma_j)^2-\Delta+\Omega_j(1-\Gamma_j)\\ \frac{1}{2}(1-\Gamma_j)^2+\frac{\lambda_{3,j}}{2}-\kappa_{3,j}(1-\Gamma_j)\\ \frac{1}{2}(1-\Gamma_j)^2+\frac{\lambda_{2,j}}{2}-\kappa_{2,j}(1-\Gamma_j)\end{pmatrix}\cdot\frac{2}{\Omega_j^2-4\Delta}\Big(\frac{R}{n}\sigma_j^2+\frac{\tilde{R}}{n}+\mathcal{E}_{w_0}^{(j)}\Big)\\ &\quad+\boldsymbol{X}_j\Lambda_j^t\begin{pmatrix}-1\\1/2\\1/2\end{pmatrix}\cdot\frac{2}{\Omega_j^2-4\Delta}\Big(2\varphi_j^{(n)}f(0)+\mathcal{E}_1^{(j)}(0)\Big)+\boldsymbol{X}_j\Lambda_j^t\begin{pmatrix}\Omega_j\\-\kappa_{3,j}\\-\kappa_{2,j}\end{pmatrix}\cdot\frac{2}{\Omega_j^2-4\Delta}\mathcal{E}_2^{(j)}(0)\Bigg]_1\\ &=\frac{2(\frac{R}{n}\sigma_j^2+\frac{\tilde{R}}{n})}{\Omega_j^2-4\Delta}\Big(-\Delta\Gamma_j\cdot\lambda_{1,j}^{t+1}+\frac{1}{2}(1-\Gamma_j-\kappa_{3,j})^2\cdot\lambda_{2,j}^{t+1}+\frac{1}{2}(1-\Gamma_j-\kappa_{2,j})^2\cdot\lambda_{3,j}^{t+1}\Big)\\ &\quad+\frac{2\mathcal{E}_{w_0}^{(j)}}{\Omega_j^2-4\Delta}\Big(-\Delta\Gamma_j\cdot\lambda_{1,j}^{t+1}+\frac{1}{2}(1-\Gamma_j-\kappa_{3,j})^2\cdot\lambda_{2,j}^{t+1}+\frac{1}{2}(1-\Gamma_j-\kappa_{2,j})^2\cdot\lambda_{3,j}^{t+1}\Big)\\ &\quad+\left(\frac{2\varphi_j^{(n)}}{\Omega_j^2-4\Delta}(-\lambda_{1,j}^{t+1}+\frac{1}{2}\cdot\lambda_{2,j}^{t+1}+\frac{1}{2}\cdot\lambda_{3,j}^{t+1})\right)2f(0)\\ &\quad+\left(\frac{2}{\Omega_j^2-4\Delta}(-\lambda_{1,j}^{t+1}+\frac{1}{2}\cdot\lambda_{2,j}^{t+1}+\frac{1}{2}\cdot\lambda_{3,j}^{t+1})\right)\mathcal{E}_1^{(j)}(0)\\ &\quad+\left(\frac{2}{\Omega_j^2-4\Delta}(\Omega_j\Delta\cdot\Delta^t-\kappa_{3,j}\lambda_{2,j}\cdot\lambda_{2,j}^t-\kappa_{2,j}\lambda_{3,j}\cdot\lambda_{3,j}^t)\right)\mathcal{E}_2^{(j)}(0).\end{aligned}$$

Simple algebra shows $1-\Gamma_j-\kappa_{3,j} = (\Omega_j-\Delta)-(\Omega_j-\kappa_{2,j}) = \Delta-\kappa_{2,j}$, and similarly, $1-\Gamma_j-\kappa_{2,j}=\Delta-\kappa_{3,j}$. Hence, we conclude that

$$f(t+1) = \frac{R}{2}h_1(t+1)+\frac{\tilde{R}}{2}h_0(t+1)+\sum_{k=0}^{t}\gamma^2\zeta(1-\zeta)H_2(t-k)f(k)+\mathcal{E}(t).$$

Here for $k=0,1$,

$$h_k(t) = \frac{1}{n}\sum_{j=1}^{n}\frac{2(\sigma_j^2)^k}{\Omega_j^2-4\Delta}\left(-\Delta\Gamma_j\cdot\Delta^t+\frac{1}{2}(\Delta-\kappa_{2,j})^2\cdot\lambda_{2,j}^t+\frac{1}{2}(\Delta-\kappa_{3,j})^2\cdot\lambda_{3,j}^t\right),$$

and

$$H_2(t) = \frac{1}{n}\sum_{j=1}^{n}\frac{2\sigma_j^4}{\Omega_j^2-4\Delta}\Big(-\lambda_{1,j}^{t+1}+\frac{1}{2}\lambda_{2,j}^{t+1}+\frac{1}{2}\lambda_{3,j}^{t+1}\Big).$$

Also, the error term $\mathcal{E}(t)$ is defined as

$$\mathcal{E}(t) \stackrel{\text{def}}{=} \mathcal{E}_{IC}(t)+\mathcal{E}_{beta}(t)+\mathcal{E}_{KL}(t)+\mathcal{E}_M(t), \tag{42}$$

where

$$\mathcal{E}_{IC}(t) \stackrel{\text{def}}{=} \sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}\left(-\Delta\Gamma_j \cdot \Delta^{t+1} + \frac{1}{2}(\Delta - \kappa_{2,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{3,j})^2 \cdot \lambda_{3,j}^{t+1}\right)\mathcal{E}_{w_0}^{(j)},$$

$$\mathcal{E}_{beta}(t) \stackrel{\text{def}}{=} \sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta \cdot \Delta^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k})\mathcal{E}_{beta}^{(j)}(k)\right),$$

$$\mathcal{E}_{KL}(t) \stackrel{\text{def}}{=} \sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta \cdot \Delta^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k})\mathcal{E}_{KL}^{(j)}(k)\right), \text{ and}$$

$$\mathcal{E}_M(t) \stackrel{\text{def}}{=} \sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta \cdot \Delta^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k})\mathcal{E}_B^{(j)}(k)\right)$$
$$+ \sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(\Omega_j\Delta \cdot \Delta^{t-k} - \kappa_{3,j}\lambda_{2,j} \cdot \lambda_{2,j}^{t-k} - \kappa_{2,j}\lambda_{3,j} \cdot \lambda_{3,j}^{t-k})\mathcal{E}_2^{(j)}(k)\right).$$

A few comments on the naming of errors: IC in $\mathcal{E}_{IC}(t)$ stands for *initial condition*. This error is generated from the initial bias on $w_{0,j}^2$. On the other hand, M in $\mathcal{E}_M(t)$ stands for *Martingale*; the error is an accumulation of martingales over each time iteration. We deal with these errors in detail in following sections. And note that Theorem 3 can be proved once we control the error $\mathcal{E}(t)$ *with overwhelming probability*.

B Estimates based on concentration of measure on the high–dimensional orthogonal group

In this section, we give a high-level overview of the errors and how to bound them with overwhelming probability. Recall that we have the following error pieces:

$$\mathcal{E}(t) \stackrel{\text{def}}{=} \mathcal{E}_{IC}(t) + \mathcal{E}_{beta}(t) + \mathcal{E}_{KL}(t) + \mathcal{E}_M(t). \tag{43}$$

In order to bound the errors, we follow the methods that are used in [27]: we would like to make an *a priori* estimate that shows the function values remain bounded. Thus, we define the stopping time, for any fixed $\theta > 0$ and *large enough* $n \in \mathbb{N}$, by

$$\vartheta \stackrel{\text{def}}{=} \inf\left\{t \geq 0 : \|\boldsymbol{w}_t\|(= \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}\|) > n^\theta\right\}.$$

We then need to show:

Lemma 1. *For any $\theta > 0$, and for any $T > 0$, $\vartheta > T$ with overwhelming probability.*

Proof. From (33), we have

$$\boldsymbol{w}_{k+1} = \left((1+\Delta)\boldsymbol{I}_n - \gamma\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_k\boldsymbol{U}\right)\boldsymbol{w}_k - \Delta\boldsymbol{w}_{k-1},$$

where $\boldsymbol{I}_n$ denotes an identity matrix of dimension $n \times n$. Therefore, by taking norm on both sides and applying triangle inequality, we have

$$\|\boldsymbol{w}_{k+1}\| \leq \left(1 + \Delta + \gamma\|\boldsymbol{\Sigma}\|_2^2\right)\|\boldsymbol{w}_k\| + \Delta\|\boldsymbol{w}_{k-1}\|.$$

Let $C := 1 + 2\Delta + \gamma\|\boldsymbol{\Sigma}\|_2^2$ and $\epsilon > 0$ is small enough so that $C^T \cdot n^\epsilon \leq n^\theta$. By induction hypothesis, if we are given $\|\boldsymbol{w}_l\| \leq C^l n^\epsilon$ for $l = 0, \cdots, k < T$, we have

$$\|\boldsymbol{w}_{k+1}\| \leq (1 + 2\Delta + \gamma\sigma_{\max}^2)C^k n^\epsilon \leq C^{k+1}n^\epsilon,$$

and this finishes the proof once we check the initial conditions, i.e., $\|\boldsymbol{w}_0\|, \|\boldsymbol{w}_1\|$ are small enough with overwhelming probability. Observe, for any $\epsilon > 0$ and sufficiently large n,

$$\|\boldsymbol{w}_0\|^2 = \sum_{j\in[n]}\left(\sigma_j\nu_{0,j} - (\boldsymbol{U}^T\boldsymbol{\eta})_j\right)^2 \leq 2(\sigma_{\max}^2\|\boldsymbol{\nu}_0\|_2^2 + \|\boldsymbol{\eta}\|_2^2) = \mathcal{O}(1) \leq n^\epsilon,$$

w.o.p. by assumption 1.1. Similarly, $\boldsymbol{w}_1$ is generated by the following formula

$$\boldsymbol{w}_1 = \left(\boldsymbol{I}_n - \gamma \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_k \boldsymbol{U}\right) \boldsymbol{w}_0,$$

and applying norm on both sides gives

$$\|\boldsymbol{w}_1\| \leq (1 + \gamma \sigma_{\max}^2)\|\boldsymbol{w}_0\| \leq (1 + \gamma \sigma_{\max}^2) n^{\epsilon} \leq C n^{\epsilon}.$$

□

We will need the result in what follows. Also, as an input, we work with the stopped process defined for any $t \geq 0$ by $\boldsymbol{w}_t^{\vartheta} \stackrel{\text{def}}{=} \boldsymbol{w}_{t \wedge \vartheta}$. Moreover, we condition on $\boldsymbol{\Sigma}$ going forward.

B.1 Control of the errors from the Initial Conditions

In this section, we focus on controlling the errors generated by the initial conditions:

$$\mathcal{E}_{IC}(t) = \sum_{j=1}^{n} \frac{1}{\Omega_j^2 - 4\Delta} \left(-\Delta \Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{2,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{3,j})^2 \cdot \lambda_{3,j}^{t+1} \right) \mathcal{E}_{w_0}^{(j)},$$

where

$$\mathcal{E}_{w_0}^{(j)} = w_{0,j}^2 - \mathbb{E}[w_{0,j}^2] = w_{0,j}^2 - \left(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\right).$$

The next Proposition shows that the error $\mathcal{E}_{IC}(t)$ can be bounded w.o.p.

Proposition 8. *For any $T > 0$ and for any $\epsilon > 0$, with overwhelming probability,*

$$\max_{0 \leq t \leq T} |\mathcal{E}_{IC}(t)| \leq n^{\epsilon - 1/2}.$$

Proof. The proof is similar to that of [27, Lemma 10]. We rely on Chebyshev's inequality and the law of total probability to control the error. Fix $t \in [T]$ and let

$$C^{(j)}(t) \stackrel{\text{def}}{=} \frac{1}{\Omega_j^2 - 4\Delta} \left(-\Delta \Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{2,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{3,j})^2 \cdot \lambda_{3,j}^{t+1} \right),$$

and

$$W(t) \stackrel{\text{def}}{=} \sum_{j=1}^{n} C^{(j)}(t) w_{0,j}^2,$$

so that $\mathcal{E}_{IC}(t) = W(t) - \mathbb{E}[W(t)]$. From [27, Lemma 10], we know that the vector $\boldsymbol{\nu}_0^2$ follows the Dirichlet distribution (recall $\boldsymbol{\nu}_k = \boldsymbol{V}^T(\boldsymbol{x}_k - \widetilde{\boldsymbol{x}})$), and in particular, $\mathbb{E}(\nu_{0,j}^4) \leq \mathcal{O}(n^{-2})$ leads to $\mathbb{E}(w_{0,j}^4) \leq \mathcal{O}(n^{-2})$ (also recall $\boldsymbol{w}_k = \boldsymbol{\Sigma}\boldsymbol{\nu}_k - \boldsymbol{U}^T\boldsymbol{\eta}$, (31)). Therefore, the (conditional) variance of $W(t)$ is bounded by

$$\begin{aligned}
\text{Var}\Big[W(t)\Big] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} C^{(j)}(t) w_{0,j}^2 - \sum_{j=1}^{n} C^{(j)}(t)\Big(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\Big)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n} C^{(j)}(t)\Big(n w_{0,j}^2 - \big(R\sigma_j^2 + \tilde{R}\big)\Big)\right)^2\right] \\
&\leq \frac{1}{n^2}\mathbb{E}\left[\sum_{j=1}^{n}\big(C^{(j)}(t)\big)^2\Big(n w_{0,j}^2 - \big(R\sigma_j^2 + \tilde{R}\big)\Big)^2\right] \\
&= \frac{1}{n}\left[\frac{1}{n}\sum_{j=1}^{n}\big(C^{(j)}(t)\big)^2\Big(n^2\mathbb{E}[w_{0,j}^4] - \big(R\sigma_j^2 + \tilde{R}\big)^2\Big)\right] = \mathcal{O}\Big(\frac{1}{n}\Big),
\end{aligned}$$

where the Cauchy-Schwarz inequality was used in the second last line. Therefore, for $\epsilon > 0$, Chebyshev inequality gives

$$\Pr\left[\left|\sum_{j=1}^{n} C^{(j)}(t) w_{0,j}^2 - \sum_{j=1}^{n} C^{(j)}(t)\Big(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\Big)\right| \geq n^{\epsilon-1/2}\right] \leq \frac{1}{n^{2\epsilon-1}}\mathrm{Var}\Big[W(t)\Big] \xrightarrow[n\to\infty]{} 0.$$

Now applying the law of total probability (over $t = 1, \cdots, T$) to this gives the claim. □

B.2 Control of the beta errors

In this section, we control the errors generated by the difference of $\frac{\beta(\beta-1)}{n(n-1)}$ and $\zeta^2 = (\frac{\beta}{n})^2$. For $t \in [T \wedge \vartheta]$, recall

$$\mathcal{E}_{beta}(t) = \sum_{k=0}^{t}\left(\sum_{j=1}^{n} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k})\mathcal{E}_{beta}^{(j)}(k)\right),$$

with

$$\mathcal{E}_{beta}^{(j)}(t) = \gamma^2\sigma_j^4\left[\left(\frac{\beta(\beta-1)}{n(n-1)} - \zeta^2\right)w_{t,j}^2 + \left(-\frac{\beta(\beta-1)}{n(n-1)} + \zeta^2\right)\sum_i U_{ij}^2\left(\sum_l U_{il}w_{t,l}\right)^2\right].$$

First of all, note that

$$\delta \stackrel{\text{def}}{=} \frac{\beta(\beta-1)}{n(n-1)} - \zeta^2 = \frac{\beta}{n} \cdot \frac{(\beta-1)n - \beta(n-1)}{n(n-1)} = \frac{\zeta(\zeta-1)}{n-1} = \mathcal{O}(n^{-1}).$$

Then we can show the following:

Proposition 9. *For any $T > 0$ and for any $\epsilon > 0$, with overwhelming probability,*

$$\max_{0\leq t\leq T\wedge\vartheta} |\mathcal{E}_{beta}(t)| \leq n^{\alpha-1/2},$$

for some $1/4 > \alpha > \epsilon$.

Proof. Let

$$C^{(j)}(t,k) \stackrel{\text{def}}{=} \frac{\gamma^2\sigma_j^4}{\Omega_j^2 - 4\Delta}(-\Delta \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}).$$

Then $C^{(j)}(t,k), j \in [n]$ are uniformly bounded by our assumptions, and we have

$$\mathcal{E}_{beta}(t) = \sum_{k=0}^{t}\sum_{j=1}^{n} C^{(j)}(t,k)\left[\delta w_{t,j}^2 - \delta\sum_i U_{ij}^2\left(\sum_l U_{il}w_{t,l}\right)^2\right].$$

Now Lemma 1 (boundedness on the norm of $\boldsymbol{w}_t$) and Lemma 3 (uniform boundedness on the coordinates of $\boldsymbol{U}\boldsymbol{w}_t$) gives

$$\mathcal{E}_{beta}(t \wedge \vartheta) \leq C\delta\big(\|\boldsymbol{w}_t^{\vartheta}\|^2 + n \cdot \max_i(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i^2\big) = \mathcal{O}(n^{2\alpha-1}),$$

for some $C > 0$, which shows our claim. □

B.3 Control of the Key lemma errors

In this section, we show that $\mathcal{E}_{KL}(t)$ can be bounded with overwhelming probability. The following Key Lemma from [27, Lemma 14] will be useful in the following:

Lemma 2 (Key Lemma)**.** *For any $T > 0$ and for any $\epsilon > 0$, for some $\{C^{(j)}(t)\}, j \in [n], 0 \leq t \leq T$ that are uniformly bounded, with overwhelming probability*

$$\max_{1\leq i\leq n}\max_{0\leq t\leq T}\left|\sum_{j=1}^{n} C^{(j)}(t)\left(\left(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\right)^2 - \frac{1}{n}\right)\right| \leq n^{\epsilon-1/2}.$$

Given this lemma, combined with the Key Lemma, we can bound the error $\mathcal{E}_{KL}(t)$ with overwhelming probability.

Proposition 10. *For any $T > 0$ and for any $\epsilon > 0$, with overwhelming probability,*

$$\max_{0\le t\le T\wedge\vartheta} |\mathcal{E}_{KL}(t)| \le n^{\epsilon-1/2}.$$

Proof. By definition, we have

$$\mathcal{E}_{KL}(t) = \sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_{KL}^{(j)}(k)\right),$$

with

$$\mathcal{E}_{KL}^{(j)}(t) = \gamma^2\sigma_j^4(\zeta-\zeta^2)\sum_{i\in[n]}(U_{ij}^2 - \frac{1}{n})\left(\sum_{l\in[n]} U_{il}w_{t,l}\right)^2.$$

Thus for a sufficiently small $\tilde{\epsilon} > 0$ and some $C > 0$, and by applying Lemma 2 and Lemma 1,

$$|\mathcal{E}_{KL}^{(n)}(t\wedge\vartheta)| \le C\sum_{k=0}^{t}\sum_{i=1}^{n}\left(e_i^T\boldsymbol{U}\boldsymbol{w}_t^{\vartheta}\right)^2\cdot n^{\tilde{\epsilon}-1/2} \le CTn^{\tilde{\epsilon}-1/2}\cdot\|\boldsymbol{w}_t^{\vartheta}\|^2 \le n^{\epsilon-1/2}.$$

□

B.4 Control of the Martingale error

In this section, we bound the error caused by Martingale terms. Recall that

$$\begin{aligned}\mathcal{E}_M(t) = &\sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta\cdot\Delta^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_B^{(j)}(k)\right)\\ &+\sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(\Omega_j\Delta\cdot\Delta^{t-k} - \kappa_{3,j}\lambda_{2,j}\cdot\lambda_{2,j}^{t-k} - \kappa_{2,j}\lambda_{3,j}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_2^{(j)}(k)\right),\end{aligned}$$

where

$$\mathcal{E}_B^{(j)}(t) = \mathcal{E}_{B^2}^{(j)}(t) + \mathcal{E}_{B,1}^{(j)}(t) + \mathcal{E}_{B,2}^{(j)}(t),$$

with

$$\begin{aligned}&\mathcal{E}_B^{(j)}(t) = \mathcal{E}_{B^2}^{j}(t) + \mathcal{E}_{B,1}^{(j)}(t) + \mathcal{E}_{B,2}^{(j)}(t), \text{ and } \mathcal{E}_2^{(j)}(t) = -\frac{1}{2}\mathcal{E}_{B,1}^{(j)}(t), \text{ with}\\ &\mathcal{E}_{B,1}^{(j)}(t) = -2\gamma\sigma_j^2 w_{t,j}\sum_{l\in[n]}\mathcal{E}_B^{(l,j)}w_{t,l},\ \mathcal{E}_B^{(l,j)} = \sum_{i\in B}U_{il}U_{ij} - \zeta\delta_{l,j},\\ &\mathcal{E}_{B,2}^{(j)}(t) = -2\gamma\sigma_j^2\Delta\sum_{l\in[n]}\mathcal{E}_B^{(l,j)}w_{t,l}(w_{t,j} - w_{t-1,j}), \text{ and}\\ &\mathcal{E}_{B^2}^{(j)}(t) = \underbrace{\gamma^2\sigma_j^4\Big(\sum_{l\in[n]} w_{t,l}\big(\sum_{i\in B}U_{ij}U_{il}\big)\Big)^2}_{\stackrel{\text{def}}{=}①} - \mathbb{E}[①].\end{aligned}$$

In view of the expression of $\mathcal{E}_M(t)$, we define

$$\mathcal{E}_{B,1}(t) \stackrel{\text{def}}{=} \sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_{B,1}^{(j)}(k)\right),$$

and

$$\mathcal{E}_{B^2}(t) \stackrel{\text{def}}{=} \sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_{B^2}^{(j)}(k)\right).$$

Then it is easy to see that controlling these two terms will lead to the control of the entire Martingale error. Control of $\mathcal{E}_{B,2}(t)$, which can be defined similarly to $\mathcal{E}_{B,1}(t)$, can be done with exactly the same as that of $\mathcal{E}_{B,1}(t)$. As for the second term of $\mathcal{E}_M(t)$ which includes $\mathcal{E}_2^{(j)}(t)$, our analysis will show that the coefficients won't play an important rule in the control of the error; so that term can be controlled for the same reason as $\mathcal{E}_{B,1}(t)$.

We organize the proof as follows. First, we introduce a proposition from [5] that gives an overwhelming probability concentration for sampling with replacement. Also, we claim that $\{\boldsymbol{U}\boldsymbol{w}_t\}, t \in [T \wedge \vartheta]$ is uniformly distributed with overwhelming probability over different coordinates. This lemma will lead to bounding the "first-order" error $\mathcal{E}_{B,1}(t)$ (similarly for $\mathcal{E}_{B,2}(t)$). As for bounding the "second-order" error $\mathcal{E}_{B^2}(t)$, we will use the Hanson-Wright inequality for sampling without replacement [1].

B.4.1 Control of $\mathcal{E}_{B,1}(t)$

The Martinagle error originates from randomly sampling a mini-batch at every iteration. We begin by presenting the following Bernstein-type concentration result for sampling without replacement so that we see that randomness does not deviate too much from the "expectation".

Proposition 11 (Proposition 1.4, [5]). *Let $\mathcal{X} = (x_1, \cdots, x_n)$ be a finite population of n points and $X_1, \cdots, X_\beta$ be a random sample drawn without replacement from $\mathcal{X}$. Let*

$$a = \min_{1 \le i \le n} x_i \text{ and } b = \max_{1 \le i \le n} x_i.$$

Also let

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

be the mean and variance of $\mathcal{X}$, respectively. Then for all $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_i - \mu \ge \epsilon\right) \le \exp\left(-\frac{\beta\epsilon^2}{2\sigma^2 + (2/3)(b-a)\epsilon}\right).$$

Now we can show that $\boldsymbol{U}\boldsymbol{w}_t$ is more or less uniformly distributed over coordinates.

Lemma 3. $\max_k |(\boldsymbol{U}\boldsymbol{w}_t^\vartheta)_k| = \mathcal{O}(n^{\alpha - 1/2})$ *with overwhelming probability for some* $1/4 > \alpha > \epsilon$.

Proof. We show a more general result, which is

$$\begin{aligned} &MB^{(t)} \overset{\text{def}}{=} \max_{1 \le k \le n} \max_{1 \le m \le n} |B_{k,m}^{(t)}| = \mathcal{O}(n^{\alpha(t) - 1/2}) \text{ w.o.p.,} \\ &\text{where } B_{k,m}^{(t)} \overset{\text{def}}{=} \sum_{j=1}^{m} U_{kj} w_{t,j}^\vartheta \text{ and } 1/4 > \alpha(T \wedge \vartheta) > \alpha((T \wedge \vartheta) - 1) > \cdots > \alpha(0) > \epsilon. \end{aligned} \tag{44}$$

Note that $B_{k,n}^{(t)} = (\boldsymbol{U}\boldsymbol{w}_t^\vartheta)_k$, so $\max_{1 \le k \le n} |(\boldsymbol{U}\boldsymbol{w}_t^\vartheta)_k| \le MB^{(t)}$. One approach is to apply the Proposition 11 and the induction hypothesis. Note that the initial condition for the induction hypothesis will be treated later. From (34), we have

$$w_{t+1,j}^\vartheta = w_{t,j}^\vartheta - \gamma\sigma_j^2 \sum_{l \in [n]} w_{t,l}^\vartheta \Big(\sum_{i \in B_{t+1}} U_{ij} U_{il} \Big) + \Delta(w_{t,j}^\vartheta - w_{t-1,j}^\vartheta).$$

By multiplying U_{kj} and summing over $j = 1, \cdots, m$, on both sides, we have

$$B_{k,m}^{(t+1)} = B_{k,m}^{(t)} - \underbrace{\gamma \sum_{j=1}^{m} \sigma_j^2 U_{kj} \sum_{l \in [n]} w_{t,l}^\vartheta \Big(\sum_{i \in B_{t+1}} U_{ij} U_{il} \Big)}_{\overset{\text{def}}{=} ①} + \Delta(B_{k,m}^{(t)} - B_{k,m}^{(t-1)}). \tag{45}$$

Let

$$X_{i,k,m}^{(t)} \overset{\text{def}}{=} \sum_{j=1}^{m} \sigma_j^2 U_{kj} U_{ij} \sum_{l \in [n]} U_{il} w_{t,l}^\vartheta = (\boldsymbol{U}\boldsymbol{\Sigma}_m^2 \boldsymbol{U}^T)_{ik} (\boldsymbol{U}\boldsymbol{w}_t^\vartheta)_i,$$

where $\mathbf{\Sigma}_m = diag(\sigma_1^2, \cdots, \sigma_m^2, 0, \cdots, 0)$, so that $① = \sum_{i \in B_{t+1}} X_{i,k,m}^{(t)}$. Note that we can assume that $k \notin B_{t+1}$ so that $k \neq i$, because we can deal with the term $X_{k,k,m}^{(t)} = (\boldsymbol{U}\mathbf{\Sigma}_m^2\boldsymbol{U}^T)_{kk}(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_k$ separately. In order to use Proposition 11, we evaluate

$$\mu = \frac{1}{n}\sum_{i\in[n]} X_{i,k,m}^{(t)} = \frac{1}{n}\sum_{j=1}^{m}\sigma_j^2 U_{kj}\sum_{l\in[n]} w_{t,l}^{\vartheta}\delta_{j,l} = \frac{1}{n}\sum_{j=1}^{m}\sigma_j^2 U_{kj} w_{t,j}^{\vartheta},$$

and

$$\sigma^2 = \frac{1}{n}\sum_i (X_{i,k,m}^{(t)})^2 - \mu^2 = \frac{1}{n}\sum_i (\boldsymbol{U}\mathbf{\Sigma}_m^2\boldsymbol{U}^T)_{ik}^2 (\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i^2 - \mu^2.$$

Now observe,

1. As for μ, by applying Abel's inequality,
$$|\mu| \leq \frac{1}{n}\sigma_{\max}^2 \max_m |B_{k,m}^{(t)}| \leq \frac{1}{n}\sigma_{\max}^2 M B^{(t)}.$$

2. When it comes to controlling σ^2, by using Lemma 1,
$$\sigma^2 \leq \frac{1}{n}\left(\max_{i\neq k} |(\boldsymbol{U}\mathbf{\Sigma}_m^2\boldsymbol{U}^T)_{ik}|^2\right)\|\boldsymbol{U}\boldsymbol{w}_t^{\vartheta}\|_2^2 + \mu^2 \leq n^{-1+2\theta}\left(\max_{i\neq k} |(\boldsymbol{U}\mathbf{\Sigma}_m^2\boldsymbol{U}^T)_{ik}|^2\right).$$
When $i \neq k$, by referring to [27, Lemma 25],
$$|(\boldsymbol{U}\mathbf{\Sigma}_m^2\boldsymbol{U}^T)_{ik}| = |\sum_{j=1}^{m}\sigma_j^2 U_{ij}U_{kj}| = \mathcal{O}(n^{-1/2+\epsilon}) \text{ w.o.p.}$$
Therefore, we have, with overwhelming probability,
$$\sigma^2 = \mathcal{O}(n^{-2+2\theta+2\epsilon}).$$

3. Observe,
$$b = \max_i X_{i,k,m}^{(t)} = \max_i (\boldsymbol{U}\mathbf{\Sigma}_m^2\boldsymbol{U}^T)_{ik}(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i \leq \mathcal{O}(n^{\epsilon})\max_i |(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i| \leq \mathcal{O}(n^{\epsilon})\cdot MB^{(t)}$$
$$= \mathcal{O}(n^{\epsilon+\alpha(t)-1/2}) \text{ w.o.p., and similar for } a = \min_i |X_{i,k,m}^{(t)}|.$$

Now applying Proposition 11 gives

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_{i,k,m} - \mu \geq t\right) \leq \exp\left(-\frac{\beta t^2}{2\sigma^2 + (2/3)(b-a)t}\right),$$

where the concentration with overwhelming probability is attained for $t = n^{-3/2+\alpha'(t)}, \alpha'(t) > \alpha(t) > \theta + \epsilon$, and therefore

$$\mathbb{P}\left(\sum_{i=1}^{\beta} X_{i,k,m}^{(t)} - \beta\mu \geq \tilde{\epsilon}\right) \to 0 \text{ as } n \to \infty \text{ when } \tilde{\epsilon} = n^{-1/2+\alpha'(t)}.$$

So applying this to (45) gives

$$B_{k,m}^{(t+1)} = B_{k,m}^{(t)} - \mathbb{1}_{i=k}\cdot X_{k,k,m}^{(t)} - \left(\beta\mu + \mathcal{O}(n^{-1/2+\alpha'(t)})\right) + \Delta(B_{k,m}^{(t)} - B_{k,m}^{(t-1)}),$$

or

$$\begin{aligned}|B_{k,m}^{(t+1)}| &\leq MB^{(t)} + \sigma_{\max}^2 MB^{(t)} + \left(\frac{\beta}{n}\sigma_{\max}^2 MB^{(t)} + \mathcal{O}(n^{-1/2+\alpha'(t)})\right) \\ &\quad + \Delta(MB^{(t)} + MB^{(t-1)}) \\ &\leq C^{(k,m)}\mathcal{O}(n^{-1/2+\alpha'(t)}),\end{aligned}$$

for some $C^{(k,m)} > 0$. Now taking maximum on k and m gives

$$MB^{(t+1)} \leq \left(\max_{k,m} C^{(k,m)}\right)\mathcal{O}(n^{-1/2+\alpha'(t)}) = \mathcal{O}(n^{-1/2+\alpha(t+1)}) \text{ w.o.p.},$$

for some $\alpha(t+1) > \alpha'(t)$. Now once we show that the initial value $MB^{(0)}$ is small enough, by the induction hypothesis, we prove the theorem. Note that as $n \to \infty$, we can always make the increment $\alpha(t+1) - \alpha(t), t \in [T \wedge \vartheta - 1]$ small enough so that $\alpha(T \wedge \vartheta) < 1/4$.

Now it suffices to check the initial condition, i.e., $MB^{(0)}$ is small enough:

Claim. $MB^{(0)} = \max_k \max_m |B^{(0)}_{k,m}| = \mathcal{O}(n^{\alpha(0)-1/2})$ w.o.p., $\quad \alpha(0) > \theta + \epsilon$.

First note that $w_{0,j} = \sigma_j \nu_{0,j} - (\boldsymbol{U}^T\boldsymbol{\eta})_j, \boldsymbol{\nu}_t = \boldsymbol{V}^T(\boldsymbol{x}_t - \tilde{\boldsymbol{x}})$. Therefore

$$B^{(0)}_{k,m} = \sum_{j=1}^m U_{kj}(\sigma_j \nu_{0,j} - \sum_{l\in[n]} U_{lj}\eta_l) = \underbrace{\sum_{j=1}^m \sigma_j U_{kj}\nu_{0,j}}_{\stackrel{\text{def}}{=}①} - \underbrace{\sum_{j=1}^m U_{kj}(\sum_{l\in[n]} U_{lj}\eta_l)}_{\stackrel{\text{def}}{=}②}.$$

We first show that $B_{k,m} = B^{(0)}_{k,m}$ for a fixed k and m attains the desired error order. As for ①, we show that $f_m(\boldsymbol{U}_k) \stackrel{\text{def}}{=} ①$ is a Lipschitz function on S^{n-1}: observe, for $\boldsymbol{U}_k, \boldsymbol{U}'_k \in S^{n-1}$,

$$\begin{aligned} f_m(\boldsymbol{U}_k) - f_m(\boldsymbol{U}'_k) &= \sum_{j=1}^m \sigma_j(U_{kj} - U'_{kj})\nu_{0,j} \\ &\leq \sqrt{\sum_{j=1}^m \sigma_j^2\nu_{0,j}^2}\sqrt{\sum_{j=1}^m (U_{kj} - U'_{kj})^2} \leq C\|\boldsymbol{U}_k - \boldsymbol{U}'_k\|_2, \end{aligned}$$

for some $C > 0$. Therefore, the concentration result for Lipschitz function ([34, Ex 5.1.12]) gives

$$\Pr\{|f_m(\boldsymbol{U}_k) - \mathbb{E}f_m(\boldsymbol{U}_k)| \geq t\} \leq 2\exp(-cnt^2),$$

and the overwhelming probability concentration is attained for $t = n^{-1/2+\epsilon}$, $\epsilon > 0$.

As for ②, observe that

$$② = \sum_{j=1}^n g_j(t)(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j,$$

where $g_j(t) = 1$ for $1 \leq j \leq m$ and 0 otherwise, $\boldsymbol{a} = e_k$, and $\boldsymbol{b} = \boldsymbol{\eta}$. Given $\boldsymbol{\eta}$ fixed, we have $\mathbb{E}_\eta[②|\boldsymbol{\eta}] = \frac{m}{n}\eta_k$. Therefore, by [27, Lemma 25], $② = \frac{m}{n}\eta_k + \mathcal{O}(n^{\epsilon-1/2})$ w.o.p. As $\max_k |\eta_k| \leq n^{\epsilon-1/2}$ w.o.p. ($f(x) = \max_i |x_i|$, $x \in S^{n-1}$ is a Lipschitz function on S^{n-1} with Lipschitz constant 1), we conclude that $② = \mathcal{O}(n^{\epsilon-1/2})$ w.o.p. Therefore $B^{(0)}_{k,m} = \mathcal{O}(n^{\alpha(0)-1/2})$ w.o.p. for arbitrarily small enough $\epsilon + \theta < \alpha(0) < 1/4$ and taking maximum over k and m shows our claim. □

Above lemma leads to the control of $\mathcal{E}_{B,1}(t)$. Note that control of $\mathcal{E}_{B,2}(t)$ can be done very similarly to $\mathcal{E}_{B,1}(t)$.

Proposition 12 (Error bound for $\mathcal{E}_{B,1}(t)$)**.**

$$\max_{0\leq t\leq T\wedge\vartheta} |\mathcal{E}_{B,1}(t)| = \mathcal{O}(n^{\alpha'-1/2}) \textit{ w.o.p.},$$

where $1/2 > \alpha' > \alpha$, with α from Lemma 3.

Proof. Our strategy is to apply Proposition 11 as well as Lemma 3. Recall that

$$\begin{aligned} \mathcal{E}_{B,1}(t) &= \sum_{k=0}^t \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathcal{E}^{(j)}_{B,1}(k)\right), \\ &= \sum_{k=0}^t \left(\sum_{j\in[n]} C^{(j)}(t,k)w_{t,j}\sum_{l\in[n]} \mathcal{E}_B^{(l,j)}w_{t,l}\right), \end{aligned}$$

where $C^{(j)}(t,k) \stackrel{\text{def}}{=} \frac{1}{\Omega_j^2 - 4\Delta}(-\Delta \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \cdot (-2\gamma\sigma_j^2)$. Let us define

$$X_i^{(t,k)} \stackrel{\text{def}}{=} \sum_{j\in[n]} C^{(j)}(t,k) U_{ij} w_{t,j}^{\vartheta} \sum_{l\in[n]} U_{il} w_{t,l}^{\vartheta}, \text{ and } \mu_{(t,k)} = \frac{1}{n}\sum_{i\in[n]} X_i^{(t,k)} = \frac{1}{n}\sum_{j\in[n]} C^{(j)}(t,k)(w_{t,j}^{\vartheta})^2,$$

so that $\mathcal{E}_{B,1}(t\wedge\vartheta) = \sum_{k=0}^{t}\left(\sum_{i\in B} X_i^{(t,k)} - \beta\mu_{(t,k)}\right)$. Let $\sigma_{(t,k)}^2$ be the variance of $X_i^{(t,k)}$:

$$\sigma_{(t,k)}^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i\in[n]}\left(\sum_{j\in[n]} C^{(j)}(t,k) U_{ij} w_{t,j}^{\vartheta}\sum_{l} U_{il} w_{t,l}^{\vartheta}\right)^2 - \left(\frac{1}{n}\sum_{j\in[n]} C^{(j)}(t,k)(w_{t,j}^{\vartheta})^2\right)^2.$$

In order to determine its order, note that

$$\frac{1}{n}\sum_{i\in[n]}\left(\sum_{j\in[n]} C^{(j)}(t,k) U_{ij} w_{t,j}^{\vartheta}\sum_{l} U_{il} w_{t,l}^{\vartheta}\right)^2 = \frac{1}{n}\sum_{i\in[n]} (\boldsymbol{U}\boldsymbol{\Sigma}_C \boldsymbol{w}_t^{\vartheta})_i^2 (\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i^2,$$

where $\boldsymbol{\Sigma}_C \stackrel{\text{def}}{=} \mathbf{diag}\{C^{(j)}(t,k)\}_{j\in[n]}$. By applying Lemma 3, we have with overwhelming probability

$$\sigma_{(t,k)}^2 \le \max_i \|(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i\|^2 \frac{1}{n}\|\boldsymbol{U}\boldsymbol{\Sigma}_C\boldsymbol{w}_t^{\vartheta}\|_2^2 \le \max_i \|(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i\|^2 \frac{1}{n}\|\boldsymbol{\Sigma}_C\|_2^2 \|\boldsymbol{w}_t^{\vartheta}\|_2^2 \le \mathcal{O}(n^{2\alpha(t)+2\theta-1}).$$

Now Proposition 11 gives

$$\Pr\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_i^{(t,k)} - \mu_{(t,k)} \ge \tilde{\epsilon}\right) \le \exp\left(-\frac{\beta\tilde{\epsilon}^2}{2\sigma_{(t,k)}^2 + (2/3)(b-a)\tilde{\epsilon}}\right),$$

where, by using Cauchy-Schwarz's inequality and applying Lemma 3 again,

$$\begin{aligned} b &= \max_{1\le i\le n}\left(\sum_{j\in[n]} C^{(j)}(t,k) U_{ij} w_{t,j}^{\vartheta}\sum_{l} U_{il} w_{t,l}^{\vartheta}\right) \\ &\le \max_i |(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i| \sqrt{\sum_{j\in[n]} (C^{(j)}(t,k))^2 (w_{t,j}^{\vartheta})^2}\sqrt{\sum_{j\in[n]} U_{ij}^2} = \mathcal{O}(n^{\alpha(t)+\theta-1/2}) \text{ w.o.p.} \end{aligned}$$

So by applying the same argument used in Proposition 11, and applying the union bound, we have

$$\mathbb{P}\left(|\mathcal{E}_{B,1}(t)| \ge \tilde{\epsilon}\right) \le T\cdot\mathbb{P}\left(\left|\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu_{(t,k)}\right| \ge c\tilde{\epsilon}\right) \searrow 0 \text{ as } n\to\infty \text{ when } \tilde{\epsilon} = n^{-1/2+\alpha'(t)},$$

for $c = 1/t$ and any $1/2 > \alpha'(t) > \alpha(t) + \theta$. Note that θ can be taken as small as possible. Now taking maximum over $t, 0 \le t \le T\wedge\vartheta, t\in\mathbb{N}$, gives the claim, with $\alpha' \stackrel{\text{def}}{=} \alpha'(T\wedge\vartheta)$. □

B.4.2 Control of $\mathcal{E}_{B^2}^{(j)}(t)$

This section deals with controlling the error $\mathcal{E}_{B^2}^{(j)}(t)$. Recall that

$$\begin{aligned} \mathcal{E}_{B^2}(t) &= \sum_{k=0}^{t}\left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathcal{E}_{B^2}^{(j)}(k)\right) \\ &= \sum_{k=0}^{t}\left(\sum_{j\in[n]} C^{(j)}(t,k)\left(\big(\sum_{l\in[n]} w_{t,l}\big(\sum_{i\in B_k} U_{ij}U_{il}\big)\big)^2 - \mathbb{E}\Big[\big(\sum_{l\in[n]} w_{t,l}\big(\sum_{i\in B_k} U_{ij}U_{il}\big)\big)^2\Big|\mathcal{F}_k\Big]\right)\right), \end{aligned}$$

where $C^{(j)}(t,k) \stackrel{\text{def}}{=} \frac{\gamma^2\sigma_j^4}{\Omega_j^2 - 4\Delta}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})$. Observe that the expression in the summand of k can be translated as a *quadratic form*:

$$\begin{aligned} \sum_{j\in[n]} C^{(j)}(t,k)\big(\sum_{l\in[n]} w_{k,l}\big(\sum_{i\in B_k} U_{ij}U_{il}\big)\big)^2 &= \sum_{j\in[n]} C^{(j)}(t,k)\left(e_j^T\boldsymbol{U}^T\boldsymbol{P}_k\boldsymbol{U}\boldsymbol{w}_k\right)^2 \\ &= (\boldsymbol{U}\boldsymbol{w}_k)^T\boldsymbol{P}_k\boldsymbol{U}\boldsymbol{\Sigma}_C\boldsymbol{U}^T\boldsymbol{P}_k(\boldsymbol{U}\boldsymbol{w}_k), \end{aligned}$$

where $\mathbf{\Sigma}_C \overset{\text{def}}{=} \mathbf{diag}\{C^{(j)}(t,k)\}_{j\in[n]}$. Let $\boldsymbol{X}_k \overset{\text{def}}{=} \boldsymbol{P}_k(\boldsymbol{U}\boldsymbol{w}_k)$ and $\boldsymbol{D} \overset{\text{def}}{=} \boldsymbol{U}\mathbf{\Sigma}_C\boldsymbol{U}^T$. Note that, for a fixed time t and k, and conditioned on $\boldsymbol{U}$, $\boldsymbol{D}$ is a fixed symmetric matrix and $\boldsymbol{X}_k$ has a randomness only depending on $\boldsymbol{P}_k$. Therefore, our error $\mathcal{E}_{B^2}(t)$ can be expressed as

$$\mathcal{E}_{B^2}(t) = \sum_{k=0}^{t} \left(\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k|\mathcal{F}_k]\right). \tag{46}$$

As we did in the previous section, in view of union bounds, it suffices to impose bounds on each summand of (46) at $k = 0, \cdots, t$. In order to have the *Hanson-Wright* type concentration for our expression, we introduce the concept of *Convex concentration property*.

Definition 1 (Convex concentration property, [1])**.** *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$. We will say that $\boldsymbol{X}$ has the convex concentration property with constant K if for every $1-$Lipschitz convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$. we have $\mathbb{E}[\varphi(\boldsymbol{X})] < \infty$ and for every $t > 0$,*

$$Pr(|\varphi(\boldsymbol{X}) - \mathbb{E}\varphi(\boldsymbol{X})| \geq t) \leq 2\exp(-t^2/K^2).$$

Remark 2. *By a simple scaling, the previous remark can extend to $x_1, \cdots, x_n \in [a,b]$, in which case K in the definition above will be replaced by $K(b-a)$.*

What is interesting for us is that vectors obtained via sampling without replacement follow the convex concentration property ([1, Remark 2.3]). More precisely, if $x_1, \cdots, x_n \in [0,1]$ and for $m \leq n$ the random vector $\boldsymbol{X} = (X_1, \cdots, X_m)$ is obtained by sampling without replacement m numbers from the set $\{x_1, \cdots, x_n\}$, then $\boldsymbol{X}$ satisfies the convex concentration property with an absolute constant K. In this sense, the following lemma ([1, Theorem 2.5]) will be useful to us.

Lemma 4 (Hanson-Wright concentration for sampling without replacement)**.** *Let $\boldsymbol{X}$ be a mean zero random vector in $\mathbb{R}^n$. If $\boldsymbol{X}$ has the convex concentration property with constant K, then for any $n \times n$ matrix $\boldsymbol{A}$ and every $t > 0$,*

$$\mathbb{P}(|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}| \geq t) \leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{2K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{K^2\|\boldsymbol{A}\|}\right)\right),$$

for some universal constant C.

Remark 3. *The assumption that $\boldsymbol{X}$ is centered is introduced just to simplify the statement of the theorem. Note that if $\boldsymbol{X}$ has the convex concentration property with constant K, then so does $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbb{E}\boldsymbol{X}$. Moreover, observe,*

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = (\tilde{\boldsymbol{X}} + \mathbb{E}\boldsymbol{X})^T\boldsymbol{A}(\tilde{\boldsymbol{X}} + \mathbb{E}\boldsymbol{X}) = \tilde{\boldsymbol{X}}^T\boldsymbol{A}\boldsymbol{X} + \tilde{\boldsymbol{X}}^T\boldsymbol{A}(\mathbb{E}\boldsymbol{X}) + (\mathbb{E}\boldsymbol{X})^T\boldsymbol{A}\boldsymbol{X} + (\mathbb{E}\boldsymbol{X})^T\boldsymbol{A}(\mathbb{E}\boldsymbol{X}),$$

and this implies

$$\begin{aligned}
&\boldsymbol{P}(|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}| \geq t) \\
&\quad\leq \boldsymbol{P}(|\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}} - \mathbb{E}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}| \geq t/3) + \boldsymbol{P}(|\tilde{\boldsymbol{X}}^T A(\mathbb{E}\boldsymbol{X}) - \mathbb{E}\tilde{\boldsymbol{X}}^T\boldsymbol{A}(\mathbb{E}\boldsymbol{X})| \geq t/3) \\
&\quad+ \boldsymbol{P}(|(\mathbb{E}\boldsymbol{X})^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}(\mathbb{E}\boldsymbol{X})^T\boldsymbol{A}\boldsymbol{X}| \geq t/3) \\
&\quad\leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{2\cdot 9K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{3K^2\|\boldsymbol{A}\|}\right)\right) + 2\cdot 2\exp\left(-\frac{t^2}{9K^2\|\boldsymbol{A}(\mathbb{E}\boldsymbol{X})\|_2^2}\right).
\end{aligned}$$

Finally, we can bound the error $\mathcal{E}_{B^2}(t)$ using Lemma 4.

Proposition 13. *For any $\epsilon > 0$, we have*

$$\max_{0\leq t\leq T\wedge\vartheta} |\mathcal{E}_{B^2}(t)| = \mathcal{O}(n^{-1/2+2\tilde{\alpha}}) \text{ w.o.p.},$$

where $1/4 > \tilde{\alpha} > \alpha$, with α from Lemma 3.

Proof. Recall that

$$\mathcal{E}_{B^2}(t) = \sum_{k=0}^{t} \left(\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k|\mathcal{F}_k]\right),$$

and we apply Lemma 4 to each summand of $\mathcal{E}_{B^2}(t \wedge \vartheta)$. More precisely,

- K is replaced by $K \cdot M_k$, where $M_k \overset{\text{def}}{=} \max_{l\in[n]} |(\boldsymbol{U}\boldsymbol{w}_k^{\vartheta})_l| = \mathcal{O}(n^{\alpha(k)-1/2})$, by Lemma 3.
- Observe that
$$\|\boldsymbol{D}\|_{HS}^2 \le \|\boldsymbol{\Sigma}_C\|_{HS}^2 = \mathcal{O}(n),$$
and
$$\|\boldsymbol{D}\| = \|\boldsymbol{\Sigma}_C\| = \mathcal{O}(1).$$
- $\mathbb{E}\boldsymbol{X}_k = (\mu_1, \cdots, \mu_n)$ where $\mu_l = \frac{\beta}{n}(\boldsymbol{U}\boldsymbol{w}_k^{\vartheta})_l, l \in [n]$, so that $\|\boldsymbol{D}\mathbb{E}\boldsymbol{X}\|_2 \le \|\boldsymbol{D}\|_2\|\mathbb{E}\boldsymbol{X}\|_2 \le \mathcal{O}(n^{\theta})$.

Therefore, by using Lemma 4, we have

$$\begin{aligned}
&\boldsymbol{P}(|\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k - \mathbb{E}\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k| \ge \tilde{\epsilon}|\mathcal{F}_k) \\
&\quad \le 2\exp\left(-\frac{1}{C}\min\left(\frac{\tilde{\epsilon}^2}{2\cdot 9M_k^4K^4\|\boldsymbol{D}\|_{HS}^2}, \frac{\tilde{\epsilon}}{3M_k^2K^2\|\boldsymbol{D}\|}\right)\right) \\
&\qquad + 2\cdot 2\exp\left(-\frac{\tilde{\epsilon}^2}{M_k^2K^2\|\boldsymbol{D}(\mathbb{E}\boldsymbol{X}_k)\|_2^2}\right),
\end{aligned}$$

and for $\tilde{\epsilon} = n^{2\tilde{\alpha}(k)-1/2}, 1/4 > \tilde{\alpha}(k) > \alpha(k)$, we obtain the desired concentration result. Now taking union bound over $k = 0, \cdots, T \wedge \vartheta$ gives the desired result, with $\tilde{\alpha} \overset{\text{def}}{=} \tilde{\alpha}(T \wedge \vartheta)$. □

B.5 Proof of Theorem 3

Proof of Theorem 3. We have observed that Proposition 8, Proposition 9, Proposition 10, Proposition 12 and Proposition 13 imply that there exists $C > 0$ such that for any $c > 0$, there exists $D > 0$ such that

$$\Pr\left[\sup_{0\le t\le T\wedge\vartheta, t\in\mathbb{N}} |\mathcal{E}(t)| > n^{-C}\right] < Dn^{-c}.$$

Now combining this result with Lemma 1 proves the Theorem. □

C Proof of Main Results

In this section, we prove various statements from Section 3. First, we analyze assumptions on the learning rate γ so that the kernel K is convergent (Proposition 2). Second, we define the Malthusian exponent and show under which conditions the convergence rate of our algorithm is determined by $\lambda_{2,\max}$ (Proposition 3). Third, We find an optimal set of learning rate and momentum parameter so that the SGD+M outperforms SGD in the large batch regime (Proposition 5). Lastly, we show the lower bound of the convergence rate of SGD+M in the small batch regime (Proposition 6).

C.1 Learning rate assumption and kernel bound

First, we show that the kernel K is always a nonnegative function, regardless of whether the eigenvalues $\{\lambda_{2,j}, \lambda_{3,j}\}, j \in [n]$ are real or complex values.

Lemma 5 (Positivity of the kernel)**.** *The kernel function satisfies $K(t) \ge 0$ for any $t \ge 0$.*

Proof. Fix $j \in [n]$ and let

$$H_{2,j}(t) \overset{\text{def}}{=} \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta}\Big(-\Delta\cdot\Delta^t + \frac{1}{2}\lambda_{2,j}\cdot\lambda_{2,j}^t + \frac{1}{2}\lambda_{3,j}\cdot\lambda_{3,j}^t\Big)$$

be the j-th summand of $H_2(t)$. We address two cases. In the first case, assume $\Omega_j^2 - 4\Delta \ge 0$. Then $\lambda_{2,j}$ and $\lambda_{3,j}$ are positive real numbers and one can easily verify that $\lambda_{2,j} \ge \Delta \ge \lambda_{3,j}$ and $\lambda_{2,j}\lambda_{3,j} = \Delta^2$. By the arithmetic-geometric inequality, we have

$$H_{2,j}(t) \ge \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta}\Big(-\Delta^{t+1} + \sqrt{\lambda_{2,j}^{t+1}\lambda_{3,j}^{t+1}}\Big) = \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta}\Big(-\Delta^{t+1} + \Delta^{t+1}\Big) = 0.$$

In the second case, we assume $\Omega_j^2 - 4\Delta < 0$. In this case, $\lambda_{2,j}$ and $\lambda_{3,j}$ are complex conjugates with magnitude Δ, and therefore we have the relation

$$\lambda_{2,j}^t = \Delta^t e^{i\theta_j t}, \quad \text{and } \lambda_{3,j}^t = \Delta^t e^{-i\theta_j t},$$

for some $\theta_j \in \mathbb{R}$. By Euler's formula, we obtain

$$-\Delta^{t+1} + \frac{1}{2}\left(\lambda_{2,j}^{t+1} + \lambda_{3,j}^{t+1}\right) = -\Delta^{t+1} + \Delta^{t+1}\cos(\theta_j t) \leq 0.$$

and combined with the condition $\Omega_j^2 - 4\Delta < 0$ gives $H_{2,j}(t) \geq 0$. Hence these two cases give the claim. □

The next proposition establishes that, under an upper bound on the learning rate, the maximum of the eigenvalues $\{\lambda_{2,j}\}$ for $j \in [n]$ has its magnitude less than one. Let $\lambda_{2,\max} \stackrel{\text{def}}{=} \max_j |\lambda_{2,j}|$. A simple computation shows that when $\lambda_{2,j}$ is complex then $|\lambda_{2,j}| = \Delta$. In particular, when all the eigenvalues $\lambda_{2,j}$ are complex numbers, $\lambda_{2,\max} = \Delta$. Otherwise, $\lambda_{2,\max} > \Delta$. Recall again that $\sigma_{\max}^2$ and $\sigma_{\min}^2$ be the largest and smallest (nonzero) eigenvalue of $\boldsymbol{A}\boldsymbol{A}^T$, respectively.

Proposition 14. *If* $\gamma < \frac{2(1+\Delta)}{\zeta\sigma_{max}^2}$ *and* $0 \leq \Delta < 1$, *then* $\lambda_{2,\max} < 1$.

Proof. First observe that

$$\gamma < \frac{2(1+\Delta)}{\zeta\sigma_{\max}^2} \iff \Omega_{\min} \stackrel{\text{def}}{=} 1 - \gamma\zeta\sigma_{\max}^2 + \Delta > -1 - \Delta,$$

so we conclude $\Omega_j > -1 - \Delta$ for all $j \in [n]$. Note that Ω_j increases as σ_j decreases. Fix $j \in [n]$. First, when Ω_j is non-positive, i.e.

$$0 \geq \Omega_j > -1 - \Delta,$$

this implies $0 \leq \Omega_j < (1+\Delta)^2$. Second, let $\Omega_j \geq 0$. Then by the definition of $\Omega_j = 1 - \gamma\zeta\sigma_j^2 + \Delta$, and as $\sigma_j^2 > 0$, we have $\Omega_j \leq 1 + \Delta$, or $\Omega_j^2 < (1+\Delta)^2$. So in both cases, we have

$$\Omega_j^2 < (1+\Delta)^2. \tag{47}$$

Then plugging in (47) into the expression of $\lambda_{2,j}$ gives

$$\begin{aligned}
|\lambda_{2,j}| = \left|\frac{-2\Delta + \Omega_j^2 + \sqrt{\Omega_j^2(\Omega_j^2 - 4\Delta)}}{2}\right| &< \left|\frac{\Delta^2 + 1\sqrt{(1+\Delta)^2(\Delta^2 - 2\Delta + 1)}}{2}\right| \\
&= \frac{\Delta^2 + 1\sqrt{(1+\Delta)^2(\Delta-1)^2}}{2} \\
&= \frac{\Delta^2 + 1 + (1+\Delta)(1-\Delta)}{2} \\
&= 1,
\end{aligned}$$

where the second last inequality comes from the constraint $0 \leq \Delta < 1$. □

Now we are ready to prove Proposition 2.

Proof of Proposition 2

Proof. Note that $\gamma < \frac{1+\Delta}{\zeta\sigma_{\max}^2}$ implies not only $\lambda_{2,\max} < 1$ from Proposition 14, but also $\Omega_j > 0$ for all $j \in [n]$. Let $\tilde{C}_j \stackrel{\text{def}}{=} \gamma^2\zeta(1-\zeta)\sigma_j^4/(\Omega_j^2 - 4\Delta)$ for the following. Using the the fact that

$\lambda_{2,j}\lambda_{3,j} = \Delta^2$ and $\lambda_{2,j} + \lambda_{3,j} = -2\Delta + \Omega_j^2$, we have

$$\begin{aligned}
\sum_{t=0}^{\infty} K(t) &= \sum_{t=0}^{\infty} \frac{1}{n}\bigg(\sum_{j=1}^{n} \tilde{C}_j(-2\Delta \cdot \Delta^t + \lambda_{2,j} \cdot \lambda_{2,j}^t + \lambda_{3,j} \cdot \lambda_{3,j}^t)\bigg) \\
&= \frac{1}{n}\sum_{j=1}^{n} \tilde{C}_j \bigg(-2\frac{\Delta}{1-\Delta} + \frac{\lambda_{2,j}}{1-\lambda_{2,j}} + \frac{\lambda_{3,j}}{1-\lambda_{3,j}} \bigg) \\
&= \frac{1}{n}\sum_{j=1}^{n} \tilde{C}_j \bigg(\frac{-2\Delta}{1-\Delta} + \frac{-2\Delta + \Omega_j^2 - 2\Delta^2}{1 + 2\Delta - \Omega_j^2 + \Delta^2} \bigg) \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{(1-\zeta)\zeta\gamma^2\sigma_j^4}{\Omega_j^2 - 4\Delta} \cdot \frac{(1+\Delta)(\Omega_j^2 - 4\Delta)}{(1-\Delta)(1+\Delta+\Omega_j)(1+\Delta-\Omega_j)} \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{(1-\zeta)\gamma\sigma_j^2}{\Omega_j^2 - 4\Delta} \cdot \frac{(1+\Delta)(\Omega_j^2 - 4\Delta)}{(1-\Delta)(1+\Delta+\Omega_j)} \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{(1-\zeta)\gamma\sigma_j^2(1+\Delta)}{(1-\Delta)(1+\Delta+\Omega_j)} \\
&\leq \frac{1}{n}\sum_{j=1}^{n} \frac{(1-\zeta)\gamma\sigma_j^2}{1-\Delta} = \frac{(1-\zeta)\gamma}{1-\Delta} \cdot \frac{1}{n}\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A}) < 1,
\end{aligned}$$

where $\Omega_j > 0$ was used in the last inequality. □

When the norm of the kernel is less than 1, we can specify the limit of the solution $\psi(t)$ to the Volterra equation when $t \to \infty$, as Proposition 1 states.

Proof of Proposition 1

Proof. This is immediate from [4, Proposition 7.4]. In particular, from our expression of the renewal equation (13), we have

$$\psi(t) \to \frac{F(\infty)}{1 - \|K\|} \quad \text{as} \quad t \to \infty.$$

Now the proof is done once we evaluate the limit of $F(t) = \frac{R}{2}h_1(t) + \frac{\tilde{R}}{2}h_0(t)$. Note that $\lim_{t\to\infty} h_1(t) = 0$. On the other hand, as for $h_0(t)$, if $n > d$, $\sigma_j = 0$ for $j = d+1, \cdots, n$. And for such j's satisfying $\sigma_j = 0$, we can easily verify that $\lambda_{2,j} = 1, \lambda_{3,j} = \Delta^2, \Omega_j = 1 + \Delta, \kappa_{2,j} = 1, \kappa_{3,j} = \Delta$. Therefore,

$$\lim_{t\to\infty} h_0(t) = \lim_{t\to\infty} \left\{ \frac{1}{n}\sum_{j=d+1}^{n} \frac{2}{\Omega_j^2 - 4\Delta}\big(0 + \frac{1}{2}(1-\Delta)^2 \cdot 1 + 0\big) \right\} = \frac{n-d}{n} = 1 - r,$$

and this proves the claim. □

C.2 Malthusian exponent and convergence rate

In this section, we show that the Malthusian exponent Ξ is always smaller than $\lambda_{2,\max}^{-1}$ for a finite dimension n. Also, in the problem constrained regime we show that SGD+M shares the same convergence rate with full batch gradient descent with momentum with adjusted learning rate.

Proposition 15. *The Malthusian exponent defined in* (15) *satisfies*

$$\Xi < (\lambda_{2,\max})^{-1}$$

when the dimension n is finite.

Proof. It suffices to observe that the convergence rate of $H_2(t)$ is determined by $\lambda_{2,\max}$; if all $\lambda_{2,j}, j \in [n]$, are real numbers, then we can easily show that $\lambda_{2,j} > \Delta > \lambda_{3,j}$. Therefore $\lambda_{2,\max}$

takes over the convergence rate of $H_2(t)$. If, for some $j \in [n]$, $\lambda_{2,j}$ and $\lambda_{3,j}$ are both complex numbers, observe that $|\lambda_{2,j}| = |\lambda_{3,j}| = \Delta$. In that case, if we let $\lambda_{2,j} = \Delta \exp(i\theta_j)$ for some $\theta_j \in \mathbb{R}$, $\lambda_{3,j} = \Delta \exp(-i\theta_j)$ then

$$-\Delta^{t+1} + \frac{1}{2}\lambda_{2,j}^{t+1} + \frac{1}{2}\lambda_{3,j}^{t+1} = -\Delta^{t+1} + \frac{1}{2}\Delta^{t+1} \cdot 2\cos(i(t+1)\theta_j) = \Delta^{t+1}(-1 + \cos(i(t+1)\theta_j)).$$

Therefore, Δ is the governing convergence rate of such j-th summand of $H_2(t)$ and the overall convergence rate of $H_2(t)$ is still determined by $\lambda_{2,\max}$. If all $\lambda_{2,j}, j \in [n]$, are complex numbers then the observation above shows that the governing convergence rate of $H_2(t)$ should be $\Delta = \lambda_{2,\max}$ and this proves our claim. □

When $\lambda_{2,\max}$ takes over the convergence behavior of SGD+M, we can easily see that its convergence dynamics is nothing but its analogue with full batch size but with adjusted learning rate. This can be easily obtained by $\zeta = 1$ in Theorem 1, but we provide a statement for full batch SGD+M and its proof for completeness.

Proof of Proposition 4

Proof. Basically, we follow the same arguments introduced in A.2, but with $\zeta = 1$; so we would not have any errors generated by selecting mini-batches. In other words, $\mathcal{E}_B^{(l,j)} = 0$. This implies the following, which is an analogue of (38),

$$\begin{pmatrix} w_{t+1,j}^2 \\ w_{t,j}^2 \\ w_{t+1,j}w_{t,j} \end{pmatrix} = \underbrace{\begin{pmatrix} \Omega_j^2 & \Delta^2 & -2\Delta\Omega_j \\ 1 & 0 & 0 \\ \Omega_j & 0 & -\Delta \end{pmatrix}}_{=\boldsymbol{M}_j} \begin{pmatrix} w_{t,j}^2 \\ w_{t-1,j}^2 \\ w_{t,j}w_{t-1,j} \end{pmatrix}. \tag{48}$$

This implies $w_{t+1,j}^2 = (\boldsymbol{M}_j^t \tilde{\mathcal{X}}_{1,j})_1$ and following the same arguments in A.2 gives

$$\begin{aligned}(\boldsymbol{M}_j^t \tilde{\mathcal{X}}_{1,j})_1 = {} & \frac{2(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n})}{\Omega_j^2 - 4\Delta}\left(-\Delta\Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{3,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{2,j})^2 \cdot \lambda_{3,j}^{t+1}\right) \\ & + \frac{2\mathcal{E}_{w_0}^{(j)}}{\Omega_j^2 - 4\Delta}\left(-\Delta\Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{3,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{2,j})^2 \cdot \lambda_{3,j}^{t+1}\right).\end{aligned}$$

Therefore, this leads to

$$f(t+1) = \frac{R}{2}h_1(t+1) + \frac{\tilde{R}}{2}h_0(t+1) + \mathcal{E}(t),$$

with the error term $\mathcal{E}(t) = \mathcal{E}_{IC}(t)$. Now taking $n \to \infty$ combined with Proposition 8 gives (20). Note that the convergence rate of $\psi_{\text{full}}(t)$ is determined by $\lambda_{2,\max}^{(\text{full})} := \max_j |\lambda_{2,j}^{(\text{full})}|$, where

$$\lambda_{2,j}^{(\text{full})} = \frac{-2\Delta + (\Omega_j^{(\text{full})})^2 + \sqrt{(\Omega_j^{(\text{full})})^2((\Omega_j^{(\text{full})})^2 - 4\Delta)}}{2},\ \Omega_j^{(\text{full})} \overset{\text{def}}{=} 1 - \gamma_{\text{full}}\sigma_j^2 + \Delta.$$

And observing that $\lambda_{2,j}^{(\text{full})} = \lambda_{2,j}$ if $\gamma_{\text{full}} = \gamma\zeta$ gives our conclusion. □

C.3 Choice of optimal learning rate and momentum

In this section, we prove Proposition 3 which states a sufficient condition for a set of learning rate and momentum parameters to be in the problem constrained regime. We also offer the proof of Proposition 5, which gives an optimal learning rate and momentum so that SGD+M outperforms SGD in terms of convergence rate. Finally, the proof of Proposition 6 will be given as well.

Proof of Proposition 3

Remark on the assumption. The first assumption on the learning rate, i.e., $\gamma \le \frac{1+\Delta}{\zeta\sigma_{\max}^2}$ implies that $\Omega_j \ge 0$ for all $j \in [n]$. On the other hand, the second condition, i.e., $\gamma \le \frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2}$, implies that $\Omega_{\max} \ge 2\sqrt{\Delta}$. Note that when $\Omega_{\max} = 2\sqrt{\Delta}$, $\lambda_{2,\max} = \frac{1}{2}\big(-2\Delta + \Omega_{\max}^2 + \sqrt{\Omega_{\max}^2(\Omega_{\max}^2 - 4\Delta)}\big) = \Delta$.

Proof. First recall that $\varphi_j^{(n)} = \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n}$ and observe that, for $1 < \Upsilon < \lambda_{2,\max}^{-1}$,

$$\begin{aligned}
\tilde{K}(\Upsilon) \stackrel{\text{def}}{=} \sum_{t=0}^{\infty} \Upsilon^t K(t) &= \sum_{t=0}^{\infty}\Bigg(\sum_{j=1}^{n}\frac{\varphi_j^{(n)}}{\Omega_j^2 - 4\Delta}(-2\Delta\cdot(\Upsilon\lambda_{1,j})^t + \lambda_{2,j}\cdot(\Upsilon\lambda_{2,j})^t + \lambda_{3,j}\cdot(\Upsilon\lambda_{3,j})^t)\Bigg) \\
&= \sum_{j=1}^{n}\frac{\varphi_j^{(n)}}{\Omega_j^2 - 4\Delta}\left(\frac{-2\Delta}{1-\Upsilon\Delta} + \frac{\lambda_{2,j}}{1-\Upsilon\lambda_{2,j}} + \frac{\lambda_{3,j}}{1-\Upsilon\lambda_{3,j}}\right) \\
&= \sum_{j=1}^{n}\frac{\varphi_j^{(n)}}{\Omega_j^2 - 4\Delta}\left(\frac{-2\Delta}{1-\Upsilon\Delta} + \frac{-2\Delta + \Omega_j^2 - 2\Upsilon\Delta^2}{1 + \Upsilon(2\Delta - \Omega_j^2) + \Upsilon^2\Delta^2}\right) \\
&= \sum_{j=1}^{n}\frac{(1-\zeta)\zeta\gamma^2\sigma_j^4}{n}\left(\frac{(1+\Upsilon\Delta)}{(1-\Upsilon\Delta)(1-\Upsilon(-2\Delta+\Omega_j^2)+\Upsilon^2\Delta^2)}\right) \\
&= \sum_{j=1}^{n}\frac{C\zeta\gamma\sigma_j^4}{n}\left(\frac{(1-\Delta)(1+\Upsilon\Delta)}{(1-\Upsilon\Delta)(1+\Upsilon\Delta+\sqrt{\Upsilon}\Omega_j)(1+\Upsilon\Delta-\sqrt{\Upsilon}\Omega_j)}\right),
\end{aligned}$$

where $C = (1-\zeta)\gamma/(1-\Delta)$. Observe, as $\Omega_j \ge 0$,

$$\tilde{K}(\Upsilon) \le \frac{C\zeta\gamma}{n}\cdot\frac{(1-\Delta)}{(1-\Upsilon\Delta)}\sum_{j=1}^{n}\frac{\sigma_j^4}{1+\Upsilon\Delta-\sqrt{\Upsilon}\Omega_j}. \tag{49}$$

Let us analyze the denominator of the summand first. Let $f_j(x) := 1 + x^2\Delta - x\Omega_j, 1 < x < \sqrt{\Delta^{-1}}$. Then the denominator in the summand is $f_j(\sqrt{\Upsilon})$. Especially, $f_{\min}(x) := \min_j f_j(x) = 1 + x^2\Delta - x\Omega_{\max}, \Omega_{\max} = 1 - \gamma\zeta{\sigma_{\min}}^2 + \Delta$. Note that $f_{\min}(x)$ is a quadratic function of x and the solution to $f_{\min}(x) = 0$ is $x = \sqrt{\lambda_{2,\max}^{-1}}$ (the other root $\sqrt{\lambda_{3,\max}^{-1}}$ exceeds the valid domain of x). Also, observe that this is where the assumption $\Omega_{\max} \ge 2\sqrt{\Delta}$ is used.

Note that $f_j(1) = \gamma\zeta\sigma_j^2$. Simple algebra shows that for $1 < x < \alpha < \beta$, $c_1(x-\alpha)^2 \le c_2(x-\alpha)(x-\beta)$ where $c_1, c_2 > 0$ satisfies $c_1(1-\alpha)^2 = c_2(1-\alpha)(1-\beta)$, i.e. two functions coincide at $x = 1$ and $x = \alpha$. If $\lambda_{2,j} \ge 0$, or $\Omega_j^2 - 4\Delta \ge 0$, then the argument above gives

$$f_{\min}\Big(\frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}\Big) \ge \frac{\gamma\zeta\sigma_{\min}^2}{4}.$$

Now for any $j \in [n]$, note that $f_j(x) - f_{\min}(x) = x\gamma\zeta(\sigma_j^2 - \sigma_{\min}^2)$ is an increasing function of $x \in \mathbb{R}$. So observe,

$$\begin{aligned}
f_j\Big(\frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}\Big) &\ge f_{\min}\Big(\frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}\Big) + \gamma\zeta(\sigma_j^2 - \sigma_{\min}^2) \\
&\ge \frac{\gamma\zeta\sigma_{\min}^2}{4} + \frac{1}{4}\gamma\zeta(\sigma_j^2 - \sigma_{\min}^2) = \frac{1}{4}\gamma\zeta\sigma_j^2.
\end{aligned}$$

Therefore, when $\sqrt{\Upsilon} = \frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}$, (49) gives

$$\tilde{K}(\Upsilon) \le \frac{4C}{n}\cdot\frac{(1-\Delta)}{(1-\Upsilon\Delta)}\sum_{j=1}^{n}\sigma_j^2.$$

Moreover, in order to bound the denominator $(1-\Upsilon\Delta)$ on the right-hand side, if we define $g(x) \stackrel{\text{def}}{=} 1-\Delta x^2$, g is a decreasing function on $[1, \sqrt{\Delta^{-1}}]$ and

$$g(\frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}) \geq g(\frac{1+\sqrt{\Delta^{-1}}}{2}) \geq \frac{1-\Delta}{2},$$

by considering a linear line passing through $(1, 1-\Delta)$ and $(\sqrt{\Delta^{-1}}, 0)$ that lies below g. Therefore,

$$\tilde{K}(\Upsilon) \leq \frac{4C}{n} \cdot \frac{(1-\Delta)}{(1-\Upsilon\Delta)} \cdot \sum_{j=1}^{n} \sigma_j^2 \leq \frac{8(1-\zeta)\gamma}{(1-\Delta)} \frac{1}{n} \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A}).$$

□

Proof of Proposition 5

Proof. First, when the assumption $\frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2} \leq \frac{(1+\sqrt{\Delta})^2}{2\zeta\sigma_{\max}^2}$ is met, we have

$$\frac{1-\sqrt{\Delta}}{1+\sqrt{\Delta}} \leq \frac{1}{\sqrt{2\kappa}}.$$

Solving this inequality with respect to Δ gives

$$\Delta \geq \left(\frac{1-\frac{1}{\sqrt{2\kappa}}}{1+\frac{1}{\sqrt{2\kappa}}}\right)^2.$$

Furthermore, from Proposition 3, when $\gamma = \frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2}$, observe that $\lambda_{2,\max} = \Delta$ and

$$\frac{8(1-\zeta)\gamma}{(1-\Delta)} \frac{1}{n} \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A}) = \frac{8(1-\zeta)(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2(1-\Delta)} \frac{1}{n} \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A}) = \frac{8(1-\zeta)}{\zeta} \cdot \frac{1-\sqrt{\Delta}}{1+\sqrt{\Delta}}\bar{\kappa} < 1.$$

Therefore, this condition implies

$$\frac{1-\sqrt{\Delta}}{1+\sqrt{\Delta}} < \frac{\mathcal{C}}{\bar{\kappa}},$$

where $\mathcal{C} = \mathcal{C}(\zeta) \stackrel{\text{def}}{=} \zeta/(8(1-\zeta))$ and solving this inequality gives

$$\sqrt{\Delta} > \frac{1-\frac{\mathcal{C}}{\bar{\kappa}}}{1+\frac{\mathcal{C}}{\bar{\kappa}}}.$$

□

Next, we present the proof of Proposition 6.

Proof of Proposition 6

Proof. For brevity and clarity, we define the following quantities:

$$\gamma_1 \stackrel{\text{def}}{=} \frac{1+\Delta}{\zeta\sigma_{\max}^2}, \quad \gamma_2 \stackrel{\text{def}}{=} \frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2}, \text{ and } \gamma_3 \stackrel{\text{def}}{=} \frac{1}{\bar{\kappa}\sigma_{\min}^2} \cdot \frac{1-\Delta}{1-\zeta}.$$

Note that the assumptions on the learning rate γ in Proposition 2 imply that $\gamma \leq \min(\gamma_1, \gamma_3)$.

First, let us assume that $\gamma \geq \gamma_2$. Recall that this condition implies that $\Omega_{\max}^2 - 4\Delta \leq 0$ and therefore $\lambda_{2,\max} = \Delta$. In this case, $\gamma_2 \leq \gamma \leq \gamma_3$ implies that

$$\frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2} \leq \frac{1}{\bar{\kappa}\sigma_{\min}^2} \cdot \frac{1-\Delta}{1-\zeta} \Rightarrow \frac{1-\sqrt{\Delta}}{1+\sqrt{\Delta}} \leq \frac{\zeta}{(1-\zeta)\bar{\kappa}}, \text{ or}$$

$$\sqrt{\Delta} \geq \frac{1-\frac{\zeta}{(1-\zeta)\bar{\kappa}}}{1+\frac{\zeta}{(1-\zeta)\bar{\kappa}}}.$$

So, combining the condition $\zeta \leq 1/2$ with the above inequality gives the claim. Therefore, for the following arguments, we assume that $\gamma \leq \gamma_2$. It is worthwhile to note that by the definition of $\lambda_{2,\max}$ and $\Omega_{\max} = 1 - \gamma\zeta\sigma_{\min}^2 + \Delta$, we know that $\lambda_{2,\max}$ is an increasing function of $\Omega_{\max}$ when $\Omega_{\max}^2 - 4\Delta \geq 0$ and $\Omega_{\max} \geq 0$ and $\Omega_{\max}$ is a decreasing function of γ. Therefore, $\lambda_{2,\max}$ attains its minimum at the maximum feasible learning rate γ.

First, let as assume that $\gamma \leq \gamma_3 \leq \gamma_1$. Then $\lambda_{2,\max}$ attains its minimum at $\gamma = \gamma_3$ and

$$\Omega_{\max} \geq 1 + \Delta - \zeta\sigma_{\min}^2 \cdot \frac{1}{\bar{\kappa}\sigma_{\min}^2} \cdot \frac{1-\Delta}{1-\zeta} = 1 + \Delta - \frac{\zeta}{(1-\zeta)\bar{\kappa}}(1-\Delta).$$

By observing that

$$\sqrt{\lambda_{2,\max}} = \frac{\Omega_{\max} + \sqrt{\Omega_{\max}^2 - 4\Delta}}{2},$$

we have

$$\sqrt{\lambda_{2,\max}} \geq \frac{1 + \Delta - c_1(1-\Delta) + \sqrt{1 + \Delta - c_1(1-\Delta)^2 - 4\Delta}}{2} =: f_1(\Delta),$$

where $c_1 \overset{\text{def}}{=} \frac{\zeta}{(1-\zeta)\bar{\kappa}} < 1$. One can easily verify that f_1 is an increasing function of $\Delta, 0 \leq \Delta < 1$, so we conclude that

$$\sqrt{\lambda_{2,\max}} \geq \sqrt{\lambda_{2,\max}}|_{\Delta=0} = 1 - c_1,$$

and we obtain the claim with the condition $\zeta \leq 1/2$.

Second, now we assume that $\gamma \leq \gamma_1 \leq \gamma_3$. Then $\lambda_{2,\max}$ attains its minimum at $\gamma = \gamma_1$ and

$$\Omega_{\max} \geq 1 + \Delta - \frac{1+\Delta}{\sigma_{\max}^2} \cdot \sigma_{\min}^2 = (1+\Delta)(1 - \frac{1}{\kappa}).$$

Therefore, for the same argument as above, we have

$$\sqrt{\lambda_{2,\max}} \geq \frac{(1+\Delta)(1-c_2) + \sqrt{(1+\Delta)^2(1-c_2)^2 - 4\Delta}}{2} =: f_2(\Delta),$$

where $c_2 \overset{\text{def}}{=} 1/\kappa$. On the other hand, the condition $\gamma_1 \leq \gamma_3$ gives

$$\begin{aligned} \frac{1+\Delta}{\zeta\sigma_{\max}^2} \leq \frac{1}{\bar{\kappa}\sigma_{\min}^2} \cdot \frac{1-\Delta}{1-\zeta} \Rightarrow \frac{1-\Delta}{1+\Delta} \geq \frac{\bar{\kappa}}{\kappa} \cdot \frac{1-\zeta}{\zeta}, \text{ or} \\ \Delta \leq \frac{1 - \frac{\bar{\kappa}}{\kappa} \cdot \frac{1-\zeta}{\zeta}}{1 + \frac{\bar{\kappa}}{\kappa} \cdot \frac{1-\zeta}{\zeta}} =: \Delta_*. \end{aligned} \tag{50}$$

Let us define $c_3 \overset{\text{def}}{=} \frac{\bar{\kappa}}{\kappa} \cdot \frac{1-\zeta}{\zeta} < 1$. Then it suffices to show that $\sqrt{\lambda_{2,\max}} \geq 1 - D\frac{c_2}{c_3}$ for some $D > 0$.

Simple algebra shows that f_2 is a concave function on $[0, \Delta_u]$ where $\Delta_u \overset{\text{def}}{=} \frac{1-\sqrt{2c_2 - c_2^2}}{1-c_2}$ makes the radical in the numerator of f_2 vanish. Also, one can verify that $f_2(0) = 1 - c_2 \geq \frac{1-c_2+\sqrt{-2c_2+c_2^2+c_3^2}}{1+c_3} = f_2(\Delta_*)$ and $\Delta_* \leq \Delta_u$, so that $f_2(\Delta) \geq f_2(\Delta_*)$ on $[0, \Delta_*]$. Hence, it suffices to show that $f_2(\Delta_*) \geq 1 - D\frac{c_2}{c_3}$ for some $D > 0$. Observe,

$$\begin{aligned} f_2(\Delta_*) &= \frac{1 - c_2 + \sqrt{-2c_2 + c_2^2 + c_3^2}}{1 + c_3} \\ &= \frac{1 - c_2 + c_3\sqrt{1 - \frac{2c_2 - c_2^2}{c_3^2}}}{1 + c_3} \\ &\geq \frac{1 - c_2 + c_3(1 - \frac{2c_2 - c_2^2}{c_3^2})}{1 + c_3} \\ &= 1 - \frac{\frac{c_2}{c_3}(2 + c_3 - c_2)}{1 + c_3} \\ &\geq 1 - 3\frac{c_2}{c_3}, \end{aligned}$$

and we finish the proof. □

Proof of Proposition 7

Proof. First, you could easily verify the following (by just setting $\Delta = 0$ in the proof of the proposition), which is an analogue of Proposition 3 for SGD without momentum.

Corollary 2. *If the learning rate $\gamma \leq \frac{1}{\zeta \sigma_{max}^2}$, with the trace condition $8(1-\zeta)\gamma \cdot \frac{1}{n}\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A}) < 1$, then γ is in the problem constrained regime with $\varepsilon = 1/2$.*

Note that when $\Delta = 0$, $\lambda_{1,j} = 0, \lambda_{2,j} = \Omega_j^2 = (1-\gamma\zeta\sigma_j)^2, \lambda_{3,j} = 0$ so that $\lambda_{2,\max} = \Omega_{\max}^2 = (1-\gamma\zeta\sigma_{\min})^2$. Besides, the assumption on γ makes the learning rate reside in the problem constrained region by Corollary 2. Observe, when $\gamma = 1/(\zeta\sigma_{\max}^2)$,

$$\lambda_{2,\max} = (1-\gamma\zeta\sigma_{\min}^2)^2 = \left(1 - \frac{1}{\kappa}\right)^2.$$

On the other hand, when $\gamma = \left(8(1-\zeta)\cdot\frac{1}{n}\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A})\right)^{-1}$, we have

$$\lambda_{2,\max} = (1-\gamma\zeta\sigma_{\min}^2)^2 = \left(1 - \frac{\zeta}{8(1-\zeta)\bar{\kappa}}\right)^2 = \left(1 - \frac{\mathcal{C}}{\bar{\kappa}}\right)^2.$$

□

D Numerical Simulations

To illustrate our theoretical results, we compare SGD+M's dynamics to (29) on moderately sized problems ($n \approx 1000$) under the setting of section 1. Moreover, the dynamics were also compared using the MNIST data set. Finally, heat maps were displayed to illustrate the interplay between the algorithmic and problem constraints. For all MNIST experiments the hyperparameters R and $\tilde{R}$ were found by running a grid-search. For simulated data experiments, we fixed R and $\tilde{R}$ and generated the data according to assumption 1.1.

Random least squares. In all simulations of the Gaussian random least squares problem, the initial weight vector $\boldsymbol{x}_0$ is set to zero and the signal and noise vectors $\tilde{\boldsymbol{x}}$ and $\boldsymbol{\eta}$ are set to $N(0, \frac{R}{n}\boldsymbol{I})$ and $N(0, \frac{\tilde{R}}{n}\boldsymbol{I})$ respectively with $\tilde{R} = R = 1$. Moreover, $\boldsymbol{A}$ is constructed by independently sampling its entries $A_{ij} \sim N(0,1)$ then row-normalized. Similarly, $\boldsymbol{b}$ is first sampled $\boldsymbol{b} \sim N(0, \frac{\tilde{R}d}{n}\boldsymbol{I})$ then the i-th entry of $\boldsymbol{b}$ is divided by the norm of the i-th row of $\boldsymbol{A}$. The objective function in which we run SGD+M in all cases is the least squares objective function $f(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2$.

Random features (RF). In Figure 5, we generate the data matrix $\boldsymbol{A}$ using a random features set-up. The model was introduced by [30] as a randomized approach for scaling kernel methods to large data sets, and has seen a surge of interest in recent years as a way to study the generalization properties of neural networks [2, 12, 17, 23]. RF is a way to increase the number of parameters without changing the data set for a least-squares problem.

In this model, the entries of $\boldsymbol{A}$ are the result of a matrix multiplication composed with a (potentially non-linear) activation function $\sigma : \mathbb{R} \to \mathbb{R}$:

$$A_{ij} \overset{\text{def}}{=} \sigma\left(\frac{[\boldsymbol{X}\boldsymbol{W}]_{ij}}{\sqrt{n_0}}\right), \quad \text{where } \boldsymbol{X} \in \mathbb{R}^{n\times n_0} \text{ and } \boldsymbol{W} \in \mathbb{R}^{n_0\times d}. \tag{51}$$

The entries of $\boldsymbol{W}$ (in Fig 5) are i.i.d. with zero mean and variance 1. The data matrix $\boldsymbol{X}$ is the MNIST data set where each row of $\boldsymbol{X}$ is an image (i.e., $n_0 = 784$). In these experiments, the activation function σ is the normalized ReLU function $\sigma(\cdot) = (\max\{0,\cdot\} - 1/\sqrt{2\pi})/\sqrt{0.5 - 1/(2\pi)})$; it is normalized so that σ applied to a standard Gaussian outputs a mean 0 and variance 1 random variable (not necessarily Gaussian).

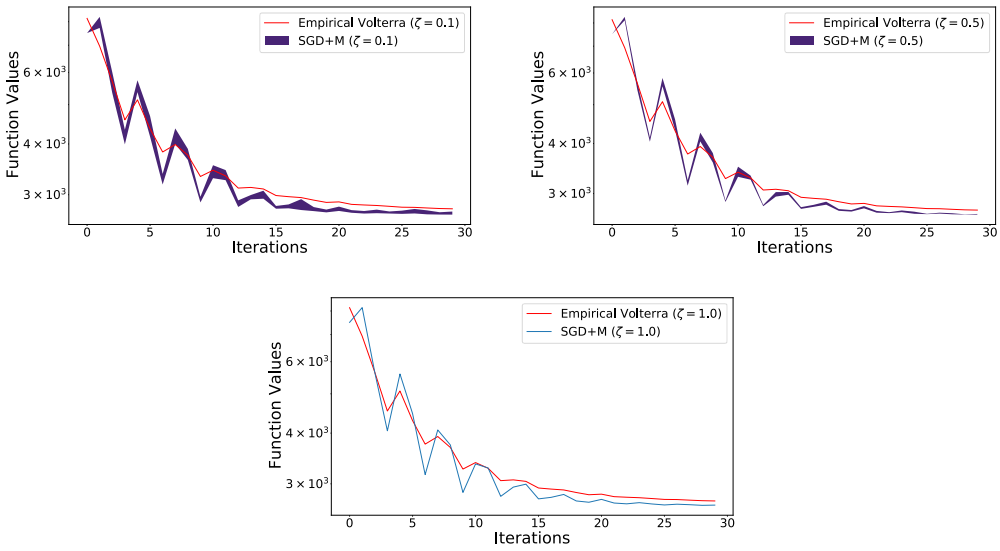

Figure 6: **SGD+M vs. Theory on even/odd MNIST.** MNIST ($60,000 \times 28 \times 28$ images) [16] is reshaped into a single matrix of dimension $60,000 \times 784$ (preconditioned to have centered rows of norm-1), representing 60,000 samples of 10 digits. The target $\boldsymbol{b}$ satisfies $\boldsymbol{b}_i = 0.5$ if the i^{th} sample is an odd digit and $\boldsymbol{b}_i = -0.5$ otherwise. SGD+M was run 10 times with $\Delta = 0.8$, various values of ζ, and learning rates $\gamma = 0.005, 0.001, 0.0005$ (left to right, top to bottom) and empirical Volterra was run once with $(R = 11,000,\ \tilde{R} = 5300)$. The R and $\tilde{R}$ values were found by running a grid-search. The 10^{th} to 90^{th} percentile interval is displayed for the loss values of 10 runs of SGD+M. Volterra predicts the convergent behavior of SGD+M in this setting.

Empirical Volterra equation. We assume that we have access to the eigenvalues of the matrix $\boldsymbol{A}\boldsymbol{A}^T$. The empirical Volterra equation (29) were computed using a dynamic programming approach by using as inputs the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$. First, the values of $h_0(t), h_1(t), H_2(t)$ were computed and stored for values of $t \in [T]$. Then a dynamic programming approach is used to compute $\psi(t)$ for values of $t \in [T]$. The discrete convolution operation in (29) is computed by an array reversal and Numpy dot product.

Volterra equation with Marchenko-Pastur distribution. In this setting, we use the theoretical limiting distribution for a large class of random matrices. In a celebrated work by [21], when the entries of $(n \times d)$ matrix $\boldsymbol{A}$ are drawn from a common, mean 0, variance $1/d$ distribution with fourth moment $\mathcal{O}(d^{-2})$ (e.g., Gaussian $N(0, \frac{1}{d})$), it is known that the distribution of eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$ converges to the Marchenko-Pastur law

$$d\mu_{MP}(\lambda) \stackrel{\text{def}}{=} \delta_0(\lambda)\max\{1-r, 0\} + \frac{r\sqrt{(\lambda-\lambda^-)(\lambda^+-\lambda)}}{2\pi\lambda} 1_{[\lambda^-,\lambda^+]},$$
$$\text{where} \quad \lambda^- \stackrel{\text{def}}{=} (1-\sqrt{\frac{1}{r}})^2 \quad \text{and} \quad \lambda^+ \stackrel{\text{def}}{=} (1+\sqrt{\frac{1}{r}})^2. \tag{52}$$

For these experiments, we generated the data matrix $\boldsymbol{A}$ with entries $N(0, 1/d)$. Instead of using the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$ in the Volterra equation (29), we used the Marchenko-Pastur distribution directly. We used a Chebyshev quadrature rule to approximate the integrals with respect to the Marchenko-Pastur distribution that arise in (29). Similar to the finite case, the integrand is computed using dynamic-programming. However, the implementation of the quadrature rule ignores the point mass at 0 so we manually add this at the end.

Volterra simulations remarks. Despite the numerical approximations to the integral, the resulting solution to the Volterra equation ψ (red line in figure 1) models the true behavior of SGD+M

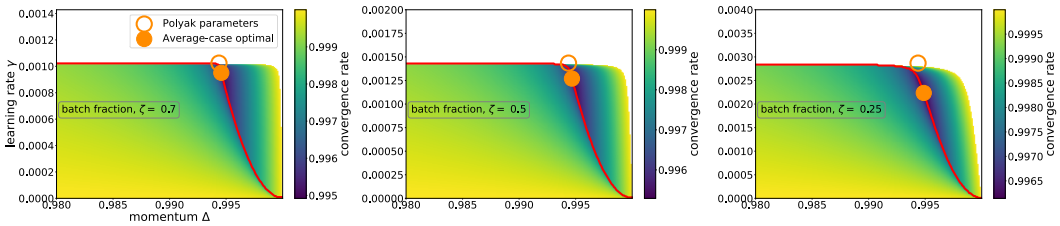

Figure 7: **Different convergence rate regions for MNIST dataset.** Plots are functions of momentum (x-axis) and learning rate (y-axis). Optimal parameters that maximize $\lambda_{2,\max}$ denoted by Polyak parameters (orange circle, (17)) and the optimal parameters for SGD+M (orange dot); below red line is the problem constrained region; otherwise the algorithmic constrained region. The MNIST data set is standardized. As the batch fraction decreases (left $\zeta = 0.7$ to right $\zeta = 0.25$), the optimal parameters of SGD+M and Polyak parameters are quite far from each other. The Malthusian exponent (algorithmically constrained region) starts to control the SGD+M rate as batch fraction $\to 0$.

remarkably well. Notably, the fit of the Volterra equation to SGD+M is extremely accurate across various learning rates, batch sizes, and momentum parameters as long as the learning rate condition is satisfied. In Figure 1, the red line corresponds to the Volterra equation with Marchenko-Pastur distribution with values $R = \tilde{R} = 1$. Also, we opted to shade the 10^{th} to 90^{th} percentile instead of an α-confidence interval for an easier read. One can observe the exact same dynamics in either case.

Heat maps. The heat maps (Figures 4, 7, and 9) illustrate when the convergence rate is dictated by the problem, ($\lambda_{2,\max} \geq \Xi^{-1}$) or by the algorithm ($\lambda_{2,\max} < \Xi^{-1}$). The white regions of the heat maps represent divergent behaviour ($\lambda_{2,\max} > 1$). The threshold, denoted by the red line, describes the boundary for two different regimes. Any non-white point above or to the right of the threshold lies in the algorithmic constraint setting. Conversely, all non-white points lying below or to the left of the threshold lies in the problem constraint setting.

The heat maps are generated by computing $\lambda_{2,\max}$ and Ξ (when it exists) across values of (Δ, γ). Here $\lambda_{2,\max}$ is obtained by calculating

$$\lambda_{2,\max} = \frac{-2\Delta + \Omega_{\max}^2 + \sqrt{\Omega_{\max}^2(\Omega_{\max}^2 - 4\Delta)}}{2}, \; \Omega_{\max} = 1 - \gamma\zeta\sigma_{\min}^2 + \Delta, \text{ and } \sigma_{\min}^2 = \left(1 - \sqrt{\frac{1}{r}}\right)^2.$$

In order to compute Ξ, recall that Ξ is the solution of

$$\tilde{K}(\Xi) \stackrel{\text{def}}{=} \sum_{t=0}^{\infty} \Xi^t K(t) = 1, \tag{53}$$

when it exists. One can show (53) is equal to (see Appendix C.3 for detail)

$$\sum_{j=1}^{n} \frac{\zeta(1-\zeta)\gamma^2\sigma_j^4}{n} \left(\frac{(1+\Xi\Delta)}{(1-\Xi\Delta)(1+\Xi\Delta+\sqrt{\Xi}\Omega_j)(1+\Xi\Delta-\sqrt{\Xi}\Omega_j)} \right) = 1, \tag{54}$$

which is computed using the Chebyshev quadrature rule.

For a given (Δ, γ), we are interested in the algorithmic case ($1 \leq \Xi \leq \lambda_{2,\max}^{-1}$) so if $\lambda_{2,\max}^{-1} < 1$ we assign a Nan value to Ξ. Otherwise, because of monotonicity of $\tilde{K}$ in (53), we perform a binary search starting with initial endpoints 1 and $\lambda_{2,\max}^{-1}$ to find the solution Ξ satisfying (53). Finally, with Ξ^{-1} and $\lambda_{2,\max}$ computed for a given (Δ, γ), we plot the maximum of the two.

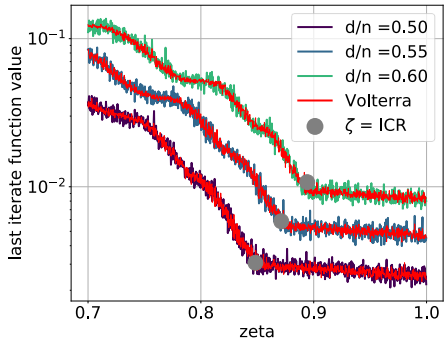

Figure 8: **Convergence behavior near the ICR.** For each value of batch fraction ζ, run SGD+M for 20 (left) and 50 (right) iterations (colored lines – blue, green, and purple) and record the function value of the last iterate. The momentum and learning rate parameters are set to be near-optimal (see (22)). Gray dot is the computed ICR (24), ζ value. Data matrix $\boldsymbol{A} \in \mathbb{R}^{d,n}$ Gaussian entries, $\tilde{\boldsymbol{x}} \sim N(\mathbf{0}, 1/n\boldsymbol{I}_d)$, $\boldsymbol{x}_0 = \mathbf{0}$ (R = 1.0), and $\boldsymbol{\eta} \sim N(\mathbf{0}, 0.0001/n\boldsymbol{I}_n)$ ($\tilde{R} = 0.0001$). Different colored lines (blue, green, purple) correspond to running SGD+M with different values of the ratio d/n. At the predicted $\zeta =$ ICR (gray dot), there is a noticeable change in the behavior of the last iterate. For ζ values less than the ICR, the value of the last iterate gets smaller as ζ increases. Then the batch fraction ζ hits the ICR and we see little to no improvement in the value of the last iterate. This agrees *exactly* with our theory for batch fraction saturation (Prop 5 and Prop. 6). For $\zeta \geq$ ICR, the convergence rate does not change; thus the values of the last iterates are approximately all equal in this regime. For $\zeta <$ ICR, our theory predicts the convergence rate improves as $\zeta \to$ ICR, $\mathcal{O}(\zeta/\bar{\kappa})$. Hence the value of the last iterate decreases here. Moreover (left), SGD+M dynamics match the predicted last value given by the Volterra equation (red) (see Thm 1).

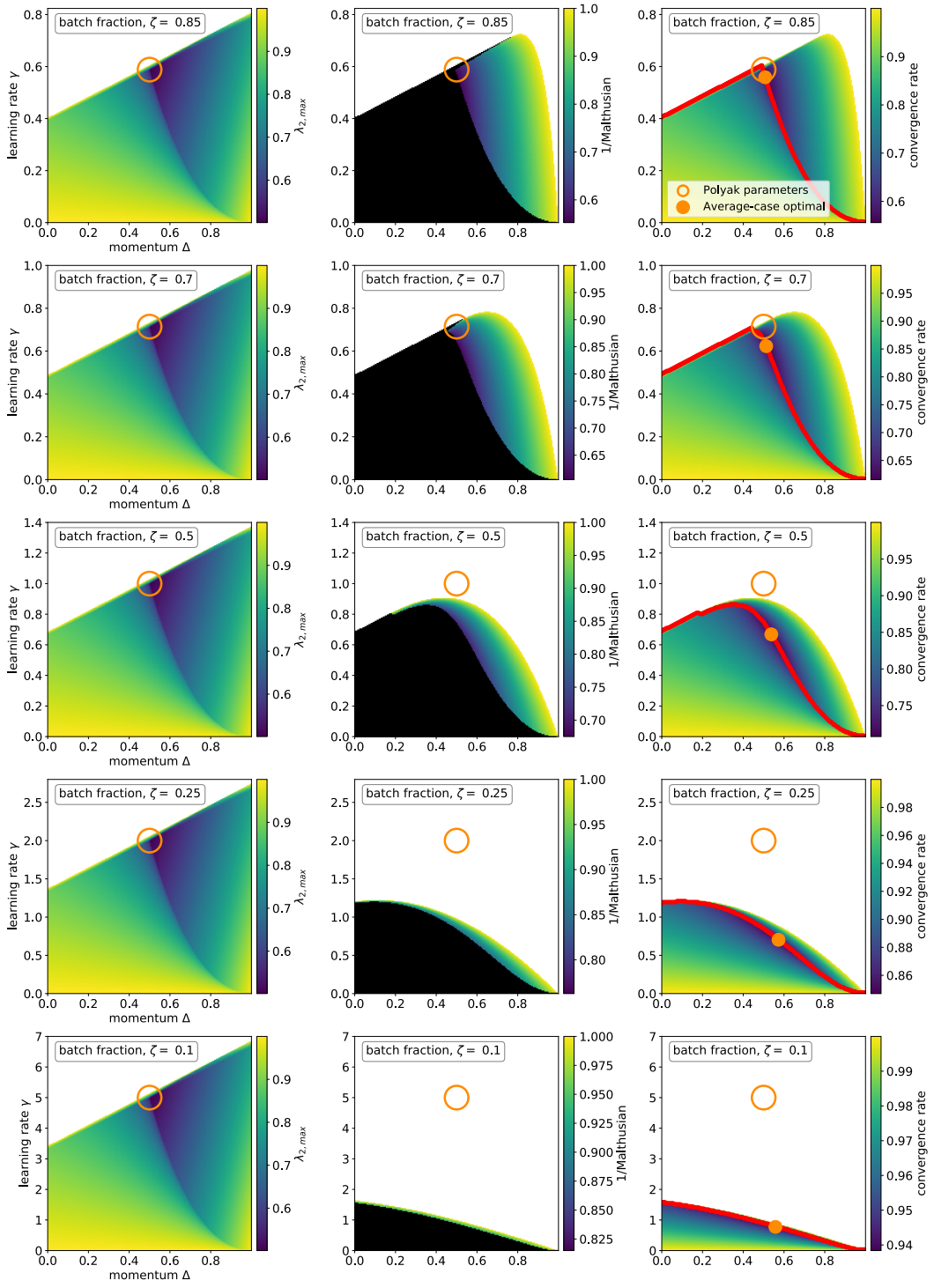

Figure 9: **Convergence rate regions for Gaussian random least squares.** Same set-up as in Figure 4 but for a wider range of batch fractions.