A EXPERIMENTS

In this section, we present numerical results to support the theorems in Section 3, to backup the hypotheses discussed in the introduction, and provide further insights into the impact of the mini-batch size on the dynamics of SGD. The experiments are conducted on four datasets and models that are relatively small due to the computational cost of using large models and datasets.

Remark: We cannot present the complete numerical results in the main paper due to the space limit. Therefore, we move the whole experimental section to Appendix. In order to keep a smooth reading, some of the content is overlapping with Section 4.

A.1 DATASETS AND SETTINGS

For all experiments, we perform mini-batch SGD multiple times starting from the same initial weights and following the same choice of the learning rates and other hyper-parameters, if applicable. This enables us to calculate the variance of the gradient estimators and other statistics in each iteration, where the randomness comes only from different samples of SGD. The learning rate α_t is selected to be inversely proportional to iteration t, or fixed, depending on the task at hand.

All models are implemented using PyTorch version 1.4 (Paszke et al., 2019) and trained on NVIDIA 2080Ti/1080 GPUs. We have also tested several other random initial weights and ground-truth weights, and learning rates, and the results and conclusions are similar and not presented.

A.1.1 GRADUATE ADMISSION DATASET

The Graduate Admission dataset[1] (Acharya et al., 2019) is to predict the chance of a graduate admission using linear regression. The dataset contains 500 samples with 6 features and is normalized by mean and variance of each feature. This is a popular regression dataset with clean data. We build a linear regression model to predict the chance of acceptance (we include the intercept term in the model) and minimize the empirical L_2 loss using mini-batch SGD, as stated in Section 3.1.

For the experiment in Figure 2(a), we randomly select an initial weight vectors w_0 and run SGD for 2,000 iterations where it appears to converge. We record all statistics at every iteration. There are in total 1,000 runs behind each observation which yields a p-value lower than 0.05. As for Figure 2(b), we select 20 different b's and run SGD from the same initial point for 40 iterations. There are in total of 200,000 runs to make sure the p-value of all statistics are lower than 0.05. In all experiments, the learning rate is chosen to be $\alpha_t = \frac{1}{2t}, t \in [2000]$ because this rate yields a theoretical convergence guaranteed (factor 1/2 has been fine tuned). The purpose of this experiment is to empirically study the rate of decrease of the variance. The theoretical study exhibited in Section 3.1 establishes the non-increasing property but it does not state anything about the rate of decrease.

A.1.2 SYNTHETIC DATASET

We build a synthetic dataset of standard normal samples to study the setting in Section 3.2. We fix the teacher network with 64 input neurons, 256 hidden neurons and 128 output neurons. We optimize the population L_2 loss by updating the two parameter matrices of the student network using online SGD, as stated in Section 3.2. In this case we have proved the functional form of the variance as a function of b and show the decreasing property of the variance of the stochastic gradient estimators for large mini-batch sizes. However, we do not show the decreasing property for every b. With this experiment we confirm that the conjecture likely holds. In the experiment, we randomly select two initial weight matrices $W_{0,1}, W_{0,2}$ and the ground-truth weight matrices W_1^*, W_2^*. We run SGD for 1,000 iterations which appears to be a good number for convergence while there are 1,000 runs of SGD in total to again give a p-value below 0.05. We record all statistics at every iteration. The learning rate is chosen to be $\alpha_t = \frac{1}{10t}, t \in [1000]$ for the same reason as in the regression experiment.

[1] `https://www.kaggle.com/mohansacharya/graduate-admissions`

A.1.3 MNIST DATASET

The MNIST dataset is to recognize digits in handwritten images of digits. We use all 60,000 training samples and 10,000 validation samples of MNIST. The images are normalized by mapping each entry to $[-1, 1]$. We build a three-layer fully connected neural network with 1024, 512 and 10 neurons in each layer. For the two hidden layers, we use the ReLU activation function. The last layer is the softmax layer which gives the prediction probabilities for the 10 digits. We use mini-batch SGD to optimize the cross-entropy loss of the model. The model deviates from our analytical setting since it has non-linear activations, it has the cross-entropy loss function (instead of L_2), and empirical loss (as opposed to population). MNIST is selected due to its fast training and popularity in deep learning experiments. The goal is to verify the results in this different setting and to back up our hypotheses.

We run SGD for 1,000 epochs on the training set which is enough for convergence. The learning rate is a constant set to $3 \cdot 10^{-3}$ (which has been tuned). For the experiment in Figure 5, there are in total 100 runs to give us the p-value below 0.05. For the experiment in Figure 4(a), we randomly select five different initial points and we have 50 runs for each initial point. For the experiment corresponding to Figure 4(b), we choose $\alpha = 8$ and $\sigma = 2$ as in Simard et al. (2013). The initial weights and other hyper-parameters are chosen to be the same as in Figure 5.

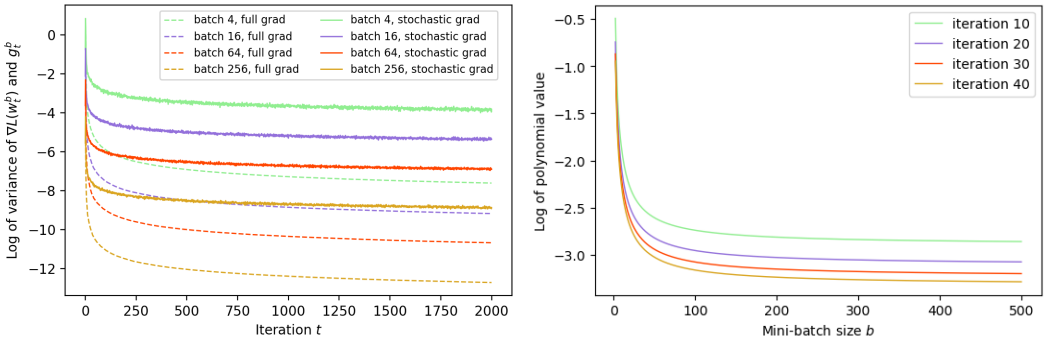
A.1.4 YELP REVIEW DATASET

The Yelp Review dataset from the Yelp Dataset Challenge (Zhang et al., 2015) contains 1,569,264 samples of customer reviews with positive/negative sentiment labels. We use 10,000 samples as our training set and 1,000 samples as the validation set. We use XLNet (Yang et al., 2019) to perform sentiment classification on this dataset. Our XLNet has 6 layers, the hidden size of 384, and 12 attention heads. There are in total 35,493,122 parameters. We intentionally reduce the number of layers and hidden size of XLNet and select a relatively small size of the training and validation sets since training of XLNet is very time-consuming (Yang et al. (2019) train on 512 TPU v3 chips for 5.5 days) and we need to train the model for multiple runs. This setting allows us to train our model in several hours on a single GPU card. We train the model using the Adam weight decay optimizer, and some other techniques, as suggested in Table 8 of Yang et al. (2019). This dataset represents sequential data where we further consider the hypotheses.

We randomly select a set of initial parameters and run Adam with two different mini-batch sizes of 32 and 64. For computational tractability reasons, for each mini-batch size there are in total of 100 runs and each run corresponds to 20 epochs. We record the variance of the stochastic gradient, loss and accuracy in every step of Adam. The statistics reported in Figure 6 are averaged through each epoch. In all experiments, the learning rate is set to be $4 \cdot 10^{-5}$ and the ϵ parameter of Adam is set to be 10^{-8} (these two have been tuned). The stochastic gradients of all parameter matrices are clipped with threshold 1 in each iteration. We use the same setup for the learning rate warm-up strategy as suggested in Yang et al. (2019). The maximum sequence length is set to be 128 and we pad the sequences with length smaller than 128 with zeros.

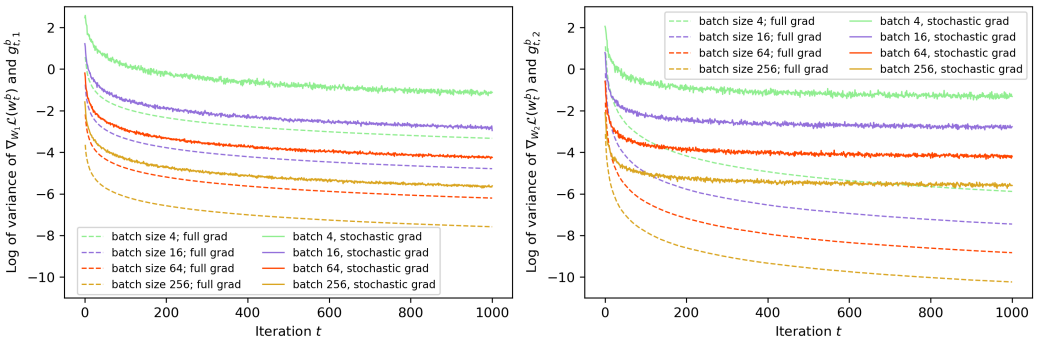
A.2 DISCUSSION

As observed in Figure 2(a), under the linear regression setting with the Graduate Admission dataset, the variance of the stochastic gradient estimators and full gradients are all strictly decreasing functions of b for all iterations. This result verifies the theorems in Section 3.1. Figure 2(b) further studies the rate of decrease of the variance. From the proofs in Section 3.1 we see that $\mathsf{var}\left(g_t^b \middle| \mathcal{F}_0\right)$ is a polynomial of $\frac{1}{b}$ with degree $t+1$. Therefore, for every t, we can approximate this polynomial by sampling many different b's and calculate the corresponding variances. We pick b to cover all numbers that are either a power of 2 or multiple of 40 in $[2, 500]$ (there are a total of 21 such values) and fit a polynomial with degree 6 (an estimate from the analyses) at $t = 10, 20, 30, 40$. Figure 2(b) shows the fitted polynomials. As we observe, the value $\mathsf{var}\left(g_t^b \middle| \mathcal{F}_0\right)$ (approximated by the value of the polynomial) is both decreasing with respect to the mini-batch size b and iteration t. Further, the rate of decrease in b is slower as the b increasing. This provides a further insight into the dynamics of training a linear regression problem with SGD.



(a) Variance of stochastic gradients and full gradients (b) Fitting polynomials of mini-batch size b

Figure 2: Experimental results for the Graduate Admission dataset. **Left:** $\log\left(\mathsf{var}\left(g_t^b \middle| \mathcal{F}_0\right)\right)$ and $\log\left(\mathsf{var}\left(\nabla L(w_t^b) \middle| \mathcal{F}_0\right)\right)$ vs iteration t for 4 different mini-batch sizes. **Right:** The log of polynomial values when fitting polynomials on selected mini-batch sizes at certain iterations.



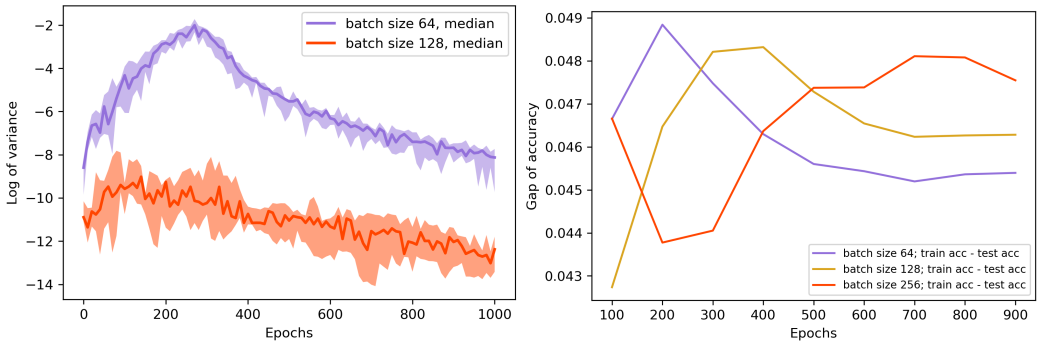
(a) Variance of gradients with respect to W_1 (b) Variance of gradients with respect to W_2

Figure 3: Experimental results for the Synthetic dataset. **Left:** $\log\left(\mathsf{var}\left(g_{t,1}^b \middle| \mathcal{F}_0\right)\right)$ and $\log\left(\mathsf{var}\left(\nabla_{W_1}\mathcal{L}(W_{t,1}^b, W_{t,2}^b) \middle| \mathcal{F}_0\right)\right)$ vs iteration t. **Right:** $\log\left(\mathsf{var}\left(g_{t,2}^b \middle| \mathcal{F}_0\right)\right)$ and $\log\left(\mathsf{var}\left(\nabla_{W_2}\mathcal{L}(W_{t,1}^b, W_{t,2}^b) \middle| \mathcal{F}_0\right)\right)$ vs iteration t.

Under the two-layer linear network setting with the synthetic dataset, Figure 3 verifies that the variance of the stochastic gradient estimators and full gradients are all strictly decreasing functions of b for all iterations. This figure also empirically shows that the constant b_0 in Theorem 5 could be as small as $b_0 = 4$. In fact, we also experiment with the mini-batch size of 1 and 2, and the decreasing property remains to hold. We also test this on multiple choices of initial weights and learning rates and this pattern remains clear.

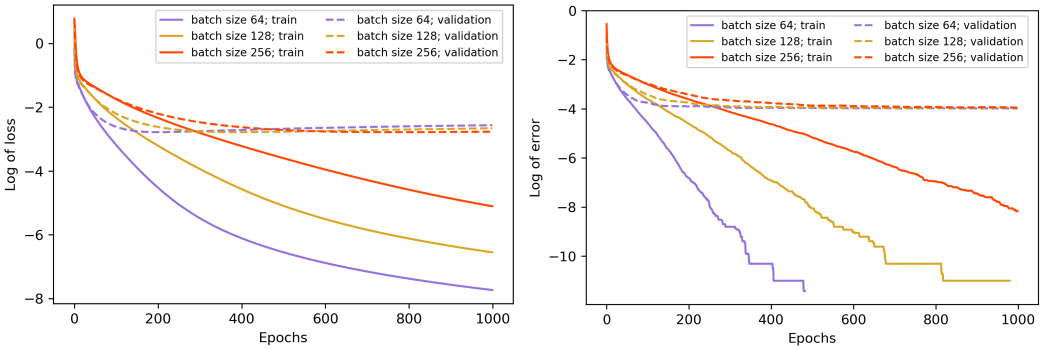
In aforementioned two experiments we use SGD in its original form by randomly sampling mini-batches. In deep learning with large-scale training data such a strategy is computationally prohibitive and thus samples are scanned in a cyclic order which implies fixed mini-batches are processed many times. Therefore, in the next two datasets we perform standard "epoch" based training to empirically study the remaining two hypotheses discussed in the introduction (decreasing loss and error as a function of b) and sensitivity with respect to the initial weights. Note that we are using cross-entropy loss in the MNIST dataset and the Adam optimizer in the Yelp dataset and thus these experiments do not meet all of the assumptions of the analysis in Section 3.

As shown in Figure 4(a), we run SGD with two batch sizes 64 and 128 on five different initial weights. This plot shows that, even the smallest value of the variance among the five different initial weights with a mini-batch size of 64, is still larger than the largest variance of mini-batch size 128. We observe that the sensitivity to the initial weights is not large. This plot also empirically verifies our conjecture in the introduction that the variance of the stochastic gradient estimators is



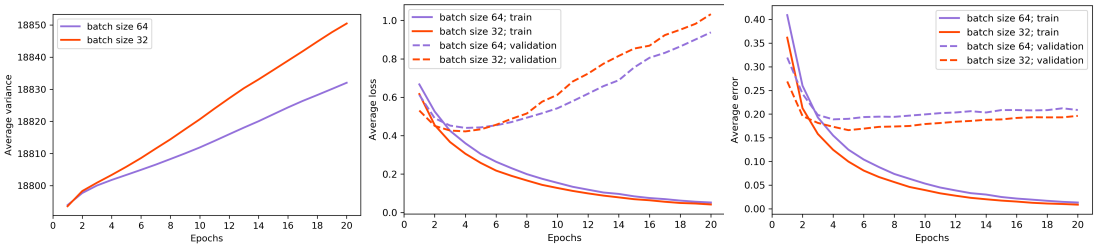
(a) Different initial weights (b) Gap of accuracy (zoomed-in)

Figure 4: Experimental results for the MNIST dataset. **Left:** The median, min, and max of the log of variance of the stochastic gradient estimators for two different mini-batch sizes (distinguished by colors) and five different initial weights. The solid lines show the median of all five initial weights while the highlighted regions show the min and max of the log of variance. **Right:** The gap of accuracy on training and test sets vs epochs starting from epoch 100.



(a) Log of loss for training and validation sets (b) Log of error for training and validation sets

Figure 5: Experimental results for the MNIST dataset. **Left:** The log of the training and validation loss vs epochs. **Right:** The log of training and validation error vs epochs. Here error is defined as one minus predicting accuracy. The plot does not show the epochs if error equals to zero.



(a) Variance of stochastic gradients (b) Training and validation loss (c) Training minus validation error

Figure 6: Experimental results for the XLNet model on the Yelp dataset. **Left:** The variance of stochastic gradient estimators vs epochs. **Middle:** The training and validation loss vs epochs. **Right:** The training and validation error vs epochs.

a decreasing function of the mini-batch size, for all iterations of SGD in a general deep learning model.

In addition, we also conjecture that there exists the decreasing property for the expected loss, error and the generalization ability with respect to the mini-batch size. Figure 5(a) shows that the expected

loss (again, randomness comes from different runs of SGD through the different mini-batches with the same initial weights and learning rates) on the training set is a decreasing function of b. However, this decreasing property does not hold on the validation set when the loss tends to be stable or increasing, in other words, the model starts to be over-fitting. We hypothesize that this is because the learned weights start to bounce around a local minimum when the model is over-fitting. As the larger mini-batch size brings smaller variance, the weights are closer to the local minimum found by SGD, and therefore yield a smaller loss function value. Figure 5(b) shows that both the expected error on training and validation sets are decreasing functions of b.

Figure 4(b) exhibits a relationship between the model's generalization ability and the mini-batch size. As suggested by (Simard et al., 2013), we build a test set by distorting the 10,000 images of the validation set. The prediction accuracy is obtained on both training and test sets and we calculate the gap between these two accuracies every 100 epochs. We use this gap to measure the model generalization ability (the smaller the better). Figure 4(b) shows that the gap is an increasing function of b starting at epoch 500, which partially aligns with our conjecture regarding the relationship between the generalization ability and the mini-batch size. We also test this on multiple choices of the hyper-parameters which control the degree of distortion in the test set and this pattern remains clear.

Figure 6 shows the similar phenomenon that the variance of stochastic estimators and the expected loss and error on both training and validation sets are decreasing functions of b even if we train XLNet using Adam. This example gives us confidence that the decreasing properties are not merely restricted on shallow neural networks or vanilla SGD algorithms. They actually appear in many advanced models and optimization methods.

B Lemmas and Proofs

B.1 Lemmas and Proofs of Results in Section 3.1

For two matrices A, B with the same dimension, we define the inner product $\langle A, B\rangle \triangleq \operatorname{tr}\left(A^T B\right)$.

Lemma 3. *Suppose that $f(x)$ and $g(x)$ are both smooth, non-negative and decreasing functions of $x \in \mathbb{R}$. Then $h(x) = f(x)g(x)$ is also a non-negative and decreasing function of x.*

Proof. It is obvious that $h(x)$ is non-negative for all x. The first-order derivative of h is

$$h'(x) = f'(x)g(x) + f(x)g'(x) \leqslant 0,$$

and thus $h(x)$ is also a decreasing function of x. □

Proof of Lemma 1. Throughout the paper, We use $C_n^k = \frac{n!}{k!(n-k)!}$ to denote the combinatorial number. Note that

$$\begin{aligned}
\mathbb{E}\left[g_t^b \left(g_t^b\right)^T \middle| \mathcal{F}_t^b\right] &= \frac{1}{b^2}\mathbb{E}\left[\sum_{i\in\mathcal{B}_t^b} \nabla L_i\left(w_t^b\right) \sum_{i\in\mathcal{B}_t^b} \nabla L_i\left(w_t^b\right)^T \middle| \mathcal{F}_t^b\right] \\
&= \frac{1}{b^2}\left(\frac{C_{n-1}^{b-1}}{C_n^b}\sum_{i=1}^n \nabla L_i\left(w_t^b\right)\nabla L_i\left(w_t^b\right)^T + \frac{C_{n-2}^{b-2}}{C_n^b}\sum_{i\neq j}\nabla L_i\left(w_t^b\right)\nabla L_j\left(w_t^b\right)^T\right) \\
&= \frac{1}{b^2}\left(\frac{b}{n}\sum_{i=1}^n \nabla L_i\left(w_t^b\right)\nabla L_i\left(w_t^b\right)^T + \frac{b(b-1)}{n(n-1)}\sum_{i\neq j}\nabla L_i\left(w_t^b\right)\nabla L_j\left(w_t^b\right)^T\right) \\
&= \frac{1}{b^2}\left(\frac{b(n-b)}{n(n-1)}\sum_{i=1}^n \nabla L_i\left(w_t^b\right)\nabla L_i\left(w_t^b\right)^T + \frac{b(b-1)}{n(n-1)}\sum_{i=1}^n \nabla L_i\left(w_t^b\right)\sum_{i=1}^n \nabla L_i\left(w_t^b\right)^T\right) \\
&= \frac{n-b}{bn(n-1)}\sum_{i=1}^n \nabla L_i\left(w_t^b\right)\nabla L_i\left(w_t^b\right)^T + \frac{(b-1)n}{b(n-1)}\nabla L\left(w_t^b\right)\nabla L\left(w_t^b\right)^T.
\end{aligned}$$

For any $A \in \mathbb{R}^{p\times p}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|Ag_t^b\right\|^2\middle|\mathcal{F}_t^b\right] &= \mathbb{E}\left[\left(g_t^b\right)^T A^T A g_t^b\middle|\mathcal{F}_t^b\right] = \mathbb{E}\left[\mathrm{tr}\left(\left(g_t^b\right)^T A^T A g_t^b\right)\middle|\mathcal{F}_t^b\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(A^T A g_t^b \left(g_t^b\right)^T\right)\middle|\mathcal{F}_t^b\right] \\
&= \mathrm{tr}\left(A^T A \mathbb{E}\left[g_t^b \left(g_t^b\right)^T\middle|\mathcal{F}_t^b\right]\right) \\
&= \mathrm{tr}\left(\frac{n-b}{bn(n-1)}\sum_{i=1}^n A^T A \nabla L_i\left(w_t^b\right)\nabla L_i\left(w_t^b\right)^T + \frac{(b-1)n}{b(n-1)} A^T A \nabla L\left(w_t^b\right)\nabla L\left(w_t^b\right)^T\right) \\
&= \frac{n-b}{bn(n-1)}\sum_{i=1}^n \left\|A\nabla L_i\left(w_t^b\right)\right\|^2 + \frac{(b-1)n}{b(n-1)}\left\|A\nabla L\left(w_t^b\right)\right\|^2 \\
&= c_b\left(\frac{1}{n}\sum_{i=1}^n \left\|A\nabla L_i\left(w_t^b\right)\right\|^2 - \left\|A\nabla L\left(w_t^b\right)\right\|^2\right) + \left\|A\nabla L\left(w_t^b\right)\right\|^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathsf{var}\left(Ag_t^b|\mathcal{F}_t^b\right) &= \mathbb{E}\left[\left\|Ag_t^b\right\|^2\middle|\mathcal{F}_t^b\right] - \left\|\mathbb{E}\left[Ag_t^b|\mathcal{F}_t^b\right]\right\|^2 \\
&= \mathbb{E}\left[\left\|Ag_t^b\right\|^2\middle|\mathcal{F}_t^b\right] - \left\|A\nabla L\left(w_t^b\right)\right\|^2 \\
&= c_b\left(\frac{1}{n}\sum_{i=1}^n \left\|A\nabla L_i\left(w_t^b\right)\right\|^2 - \left\|A\nabla L\left(w_t^b\right)\right\|^2\right).
\end{aligned}$$

□

Lemma 4. *For any set of square matrices $\{A_1, \cdots, A_n\} \in \mathbb{R}^{p\times p}$, if we denote $A = \sum_{i=1}^n A_i x_i x_i^T$, then we have*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_{t+1}^b\right)\right\|^2\middle|\mathcal{F}_0\right] = \mathbb{E}\left[\left\|\sum_{i=1}^n B_i \nabla L_i\left(w_t^b\right)\right\|^2\middle|\mathcal{F}_0\right] + \frac{\alpha_t^2 c_b}{n^2}\sum_{k=1}^n\sum_{l=1}^n \mathbb{E}\left[\left\|\sum_{i=1}^n B_i^{kl}\nabla L_i\left(w_t^b\right)\right\|^2\middle|\mathcal{F}_0\right].$$

Here $B_i = A_i - \frac{\alpha_t}{n}A$; $B_i^{kl} = A$ if $i = k, i \neq l$, $B_i^{kl} = A$ if $i = l, i \neq k$, and B_i^{kl} equals the zero matrix, otherwise.

Proof of Lemma 4. Let $C_i = x_i x_i^T$ and $C = \frac{1}{n}\sum_{i=1}^n C_i$. For the given $A_1, \ldots, A_n$, we denote $A = \sum_{i=1}^n A_i C_i$. Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_{t+1}^b\right)\right\|^2 \middle| \mathcal{F}_0\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_{t+1}^b\right)\right\|^2 \middle| \mathcal{F}_t^b\right] \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \left(x_i^T w_{t+1}^b - y_i\right) x_i\right\|^2 \middle| \mathcal{F}_t^b\right] \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \left(x_i^T \left(w_t^b - \alpha_t g_t^b\right) - y_i\right) x_i\right\|^2 \middle| \mathcal{F}_t^b\right] \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_t^b\right) - \alpha_t A g_t^b\right\|^2 \middle| \mathcal{F}_t^b\right] \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right] - 2\alpha_t \mathbb{E}\left[\mathbb{E}\left[\left\langle \sum_{i=1}^n A_i \nabla L_i\left(w_t^b\right), A g_t^b \right\rangle \middle| \mathcal{F}_t^b\right] \middle| \mathcal{F}_0\right] \\
&\quad + \alpha_t^2 \mathbb{E}\left[\mathbb{E}\left[\left\|A g_t^b\right\|^2 \middle| \mathcal{F}_t^b\right] \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right] - 2\alpha_t \mathbb{E}\left[\left\langle \sum_{i=1}^n A_i \nabla L_i\left(w_t^b\right), A \nabla L\left(w_t^b\right) \right\rangle \middle| \mathcal{F}_0\right] \\
&\quad + \alpha_t^2 \mathbb{E}\left[c_b\left(\frac{1}{n}\sum_{i=1}^n \left\|A \nabla L_i(w_t^b)\right\|^2 - \left\|A \nabla L(w_t^b)\right\|^2\right) + \left\|A \nabla L(w_t^b)\right\|^2 \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_t^b\right) - \alpha_t A \nabla L(w_t^b)\right\|^2 \middle| \mathcal{F}_0\right] + \alpha_t^2 c_b \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|A \nabla L_i(w_t^b)\right\|^2 - \left\|A \nabla L(w_t^b)\right\|^2 \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_t^b\right) - \alpha_t A \nabla L(w_t^b)\right\|^2 \middle| \mathcal{F}_0\right] + \frac{\alpha_t^2 c_b}{n^2}\sum_{i\neq j} \mathbb{E}\left[\left\|A \nabla L_i\left(w_t^b\right) - A \nabla L_j\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^n \left(A_i - \frac{\alpha_t}{n} A\right) \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right] + \frac{\alpha_t^2 c_b}{n^2}\sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[\left\|A \nabla L_i\left(w_t^b\right) - A \nabla L_j\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right].
\end{aligned}
$$

Therefore, if we set $B_i = A_i - \frac{\alpha_t}{n} A$ and

$$
B_i^{kl} = \begin{cases} A & i = k, i \neq l, \\ -A & i = l, i \neq k, \\ 0 & \text{otherwise}, \end{cases}
$$

we have

$$
\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_{t+1}^b\right)\right\|^2 \middle| \mathcal{F}_0\right] = \mathbb{E}\left[\left\|\sum_{i=1}^n B_i \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right] + \frac{\alpha_t^2 c_b}{n^2}\sum_{k=1}^n \sum_{l=1}^n \mathbb{E}\left[\left\|\sum_{i=1}^n B_i^{kl} \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right].
$$

□

Proof of Theorem 1. We use induction to show this statement.

When $t = 0$, $\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right] = \left\|\sum_{i=1}^n A_i \nabla L_i\left(w_0\right)\right\|^2$ which is invariant of b. Therefore, it is a decreasing function of b.

Suppose the statement holds for t. For any set of matrices $\{A_1, \ldots, A_n\}$ in $\mathbb{R}^{p\times p}$, by Lemma 2 we know that there exist matrices $\{B_1, \cdots, B_n\}$ and $\left\{B_i^{kl} : i, k, l \in [n]\right\}$ such that

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} A_i \nabla L_i\left(w_{t+1}^b\right)\right\|^2 \middle| \mathcal{F}_0\right] = \mathbb{E}\left[\left\|\sum_{i=1}^{n} B_i \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right] + \frac{\alpha_t^2 c_b}{n^2} \sum_{k=1}^{n}\sum_{l=1}^{n} \mathbb{E}\left[\left\|\sum_{i=1}^{n} B_i^{kl} \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right].$$

By induction, we know that $\mathbb{E}\left[\left\|\sum_{i=1}^n B_i \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right]$ and all $\mathbb{E}\left[\left\|\sum_{i=1}^n B_i^{kl} \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right]$ are non-negative and decreasing functions of b. Besides, clearly $\frac{\alpha_t^2 c_b}{n^2} = \frac{\alpha_t^2 (n-b)}{bn^3(n-1)}$ is a non-negative and decreasing function of b. By Lemma 3, we know that $\frac{\alpha_t^2 c_b}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n B_i^{kl} \nabla L_i\left(w_t^b\right)\right\|^2 \middle| \mathcal{F}_0\right]$ is also a non-negative and decreasing function of b. Finally, $\mathbb{E}\left[\left\|\sum_{i=1}^n A_i \nabla L_i\left(w_{t+1}^b\right)\right\|^2 \middle| \mathcal{F}_0\right]$, as the sum of non-negative and decreasing functions in b, is a non-negative and decreasing function of b.

□

In order to prove Theorem 2, we split the task to two separate theorems about the full gradient and the stochastic gradient and prove them one by one.

Theorem 6. *Fixing initial weights w_0, $\mathsf{var}\left(B\nabla L\left(w_t^b\right)\middle|\mathcal{F}_0\right)$ is a decreasing function of mini-batch size b for all $b \in [n]$, $t \in \mathbb{N}$, and all square matrices $B \in \mathbb{R}^{p\times p}$.*

Theorem 7. *Fixing initial weights w_0, $\mathsf{var}\left(Bg_t^b\middle|\mathcal{F}_0\right)$ is a decreasing function of mini-batch size b for all $b \in [n]$, $t \in \mathbb{N}$, and all square matrices $B \in \mathbb{R}^{p\times p}$.*

Proof of Theorem 6. We induct on t to show that the statement holds. For $t = 0$, we have $\mathsf{var}\left(B\nabla L\left(w_t^b\right)\middle|\mathcal{F}_0\right) = 0$ for any matrix B. Suppose the statement holds for $t - 1 \geqslant 0$. Note that from

$$\begin{aligned}
\nabla L\left(w_t^b\right) &= \frac{1}{n}\sum_{i=1}^{n} x_i\left(x_i^T w_t^b - y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i\left(x_i^T\left(w_{t-1}^b - \alpha_t g_{t-1}^b\right) - y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i\left(x_i^T w_{t-1}^b - y_i\right) - \frac{\alpha_t}{n}\sum_{i=1}^{n} x_i x_i^T g_{t-1}^b \\
&= \nabla L\left(w_{t-1}^b\right) - \alpha_t C g_{t-1}^b,
\end{aligned}$$

we have

$$
\begin{aligned}
&\mathsf{var}\left(B\nabla L\left(w_t^b\right)|\mathcal{F}_0\right)\\
&=\mathsf{var}\left(B\nabla L\left(w_{t-1}^b\right)-\alpha_t BCg_{t-1}^b\,|\mathcal{F}_0\right)\\
&=\mathbb{E}\left[\left\|B\nabla L\left(w_{t-1}^b\right)-\alpha_t BCg_{t-1}^b\right\|^2\Big|\mathcal{F}_0^b\right]-\left\|\mathbb{E}\left[B\nabla L\left(w_{t-1}^b\right)-\alpha_t BCg_{t-1}^b|\mathcal{F}_0^b\right]\right\|^2\\
&=\mathbb{E}\left[\left\|B\nabla L\left(w_{t-1}^b\right)\right\|^2-2\alpha_t\left\langle B\nabla L\left(w_{t-1}^b\right),BCg_{t-1}^b\right\rangle+\alpha_t^2\left\|BCg_{t-1}^b\right\|^2\Big|\mathcal{F}_0^b\right]-\left\|\mathbb{E}\left[B\nabla L\left(w_{t-1}^b\right)-\alpha_t BCg_{t-1}^b|\mathcal{F}_0^b\right]\right\|^2\\
&=\mathbb{E}\left[\left\|B\nabla L\left(w_{t-1}^b\right)\right\|^2\Big|\mathcal{F}_0\right]+\alpha_t^2\mathbb{E}\left[\mathbb{E}\left[\left\|BCg_{t-1}^b\right\|^2\Big|\mathcal{F}_{t-1}^b\right]\Big|\mathcal{F}_0^b\right]-2\alpha_t\mathbb{E}\left[\mathbb{E}\left[\left\langle B\nabla L\left(w_{t-1}^b\right),BCg_{t-1}^b\right\rangle|\mathcal{F}_{t-1}^b\right]|\mathcal{F}_0\right]\\
&\quad-\left\|\mathbb{E}\left[\mathbb{E}\left[B\nabla L\left(w_{t-1}^b\right)-\alpha_t BCg_{t-1}^b|\mathcal{F}_{t-1}^b\right]|\mathcal{F}_0^b\right]\right\|^2\\
&=\mathbb{E}\left[\left\|B\nabla L\left(w_{t-1}^b\right)\right\|^2\Big|\mathcal{F}_0\right]+\alpha_t^2\mathbb{E}\left[c_b\left(\frac{1}{n}\sum_{i=1}^n\left\|BC\nabla L_i\left(w_{t-1}^b\right)\right\|^2-\left\|BC\nabla L\left(w_{t-1}^b\right)\right\|^2\right)+\left\|BC\nabla L\left(w_{t-1}^b\right)\right\|^2\Big|\mathcal{F}_0\right]\\
&\quad-2\alpha_t\mathbb{E}\left[\left\langle B\nabla L\left(w_{t-1}^b\right),BC\nabla L\left(w_{t-1}^b\right)\right\rangle|\mathcal{F}_0\right]-\left\|\mathbb{E}\left[B\nabla L\left(w_{t-1}^b\right)-\alpha_t BC\nabla L\left(w_{t-1}^b\right)|\mathcal{F}_0^b\right]\right\|^2 \qquad (3)\\
&=\mathbb{E}\left[\left\|B\left(I-\alpha_t C\right)\nabla L\left(w_{t-1}^b\right)\right\|^2\Big|\mathcal{F}_0^b\right]+\alpha_t^2c_b\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n\left\|BC\nabla L_i\left(w_{t-1}^b\right)\right\|^2-\left\|BC\nabla L\left(w_{t-1}^b\right)\right\|^2\right)\Big|\mathcal{F}_0\right]\\
&\quad-\left\|\mathbb{E}\left[B\left(I-\alpha_t C\right)\nabla L\left(w_{t-1}^b\right)|\mathcal{F}_0^b\right]\right\|^2\\
&=\mathsf{var}\left(B\left(I-\alpha_t C\right)\nabla L\left(w_{t-1}^b\right)|\mathcal{F}_0\right)+\alpha_t^2c_b\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|BC\nabla L_i\left(w_{t-1}^b\right)\right\|^2\Big|\mathcal{F}_0\right]-\mathbb{E}\left[\left\|BC\nabla L\left(w_{t-1}^b\right)\right\|^2\Big|\mathcal{F}_0\right]\right)\\
&=\mathsf{var}\left(B\left(I-\alpha_t C\right)\nabla L\left(w_{t-1}^b\right)|\mathcal{F}_0\right)+\frac{\alpha_t^2c_b}{n^2}\sum_{i\neq j}\mathbb{E}\left[\left\|BC\nabla L_i\left(w_{t-1}^b\right)-BC\nabla L_j\left(w_{t-1}^b\right)\right\|^2\Big|\mathcal{F}_0\right], \qquad (4)
\end{aligned}
$$

where (3) is by Lemma 1. By induction, we know that the first term of (4) is a decreasing function of b. Taking $A_i=BC, A_j=-BC, A_k=0, k\in[n]\backslash\{i,j\}$ in Theorem 1, we know that

$$\mathbb{E}\left[\left\|BC\nabla L_i\left(w_{t-1}^b\right)-BC\nabla L_j\left(w_{t-1}^b\right)\right\|^2\Big|\mathcal{F}_0\right]$$

is also a decreasing function of b. Note that $\frac{\alpha_t^2c_b}{n^2}$ decreases as b increases. By Lemma 3 we learn that (4) is a decreasing function of b and hence we have completed the induction.

□

Proof of Theorem 7. We have

$$
\begin{aligned}
\mathsf{var}\left(Bg_t^b|\mathcal{F}_0\right)&=\mathbb{E}\left[\left\|Bg_t^b\right\|^2\Big|\mathcal{F}_0\right]-\left\|\mathbb{E}\left[Bg_t^b|\mathcal{F}_0\right]\right\|^2\\
&=\mathbb{E}\left[\mathbb{E}\left[\left\|Bg_t^b\right\|^2\Big|\mathcal{F}_t^b\right]\Big|\mathcal{F}_0\right]-\left\|\mathbb{E}\left[\mathbb{E}\left[Bg_t^b|\mathcal{F}_t^b\right]|\mathcal{F}_0\right]\right\|^2\\
&=c_b\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|B\nabla L_i\left(w_t^b\right)\right\|^2\Big|\mathcal{F}_0\right]-\mathbb{E}\left[\left\|B\nabla L\left(w_t^b\right)\right\|^2\Big|\mathcal{F}_0\right]\right)\\
&\quad+\mathbb{E}\left[\left\|B\nabla L\left(w_t^b\right)\right\|^2\Big|\mathcal{F}_0\right]-\left\|\mathbb{E}\left[B\nabla L\left(w_t^b\right)|\mathcal{F}_0\right]\right\|^2\\
&=\frac{c_b}{n^2}\sum_{i\neq j}\mathbb{E}\left[\left\|B\nabla L_i\left(w_t^b\right)-B\nabla L_j\left(w_t^b\right)\right\|^2\Big|\mathcal{F}_0\right]+\mathsf{var}\left(B\nabla L\left(w_t^b\right)|\mathcal{F}_0\right).
\end{aligned}
$$

Taking $A_i=B, A_j=-B, A_k=0, k\in[n]\backslash\{i,j\}$ in Theorem 1, we know that

$$\mathbb{E}\left[\left\|B\nabla L_i\left(w_t^b\right)-B\nabla L_j\left(w_t^b\right)\right\|^2\Big|\mathcal{F}_0\right]$$

is a decreasing and non-negative function of b for all $i,j\in[n]$. By Theorem 6, we know that $\mathsf{var}\left(B\nabla L\left(w_t^b\right)|\mathcal{F}_0\right)$ is also a decreasing function of b. Therefore, $\mathsf{var}\left(Bg_t^b|\mathcal{F}_0\right)$, as the sum of two decreasing functions of b, is also a decreasing function of b. □

Proof of Corollary 1. Simply taking $B = I_p$ in Theorem 1 yields the proof. □

B.2 PROOFS FOR RESULTS IN 3.2

Remark. We often rely on the trivial facts that $x_1 x_2^T = x_1 I_p x_2^T$ and $x_1 x_2^T x_3 x_4^T = x_1 x_2^T I_p x_3 x_4^T$.

Lemma 5. *Given a multiplicative term of parameter matrices* $\left\{u_i v_i^T : u_i, v_i \in \mathbb{R}^p, i \in [n_1]\right\} \cup \{A_j : A_j \in \mathbb{R}^{p\times p}, j \in [n_2]\}$ *and constant matrix* $\{I_p\}$ *such that* $\deg(u_1 v_1^T; M) \geqslant 1$*, we have*

$$\operatorname{tr}(M) = v_1^T M' u_1,$$

where M' *is a multiplicative term of parameter matrices* $\left\{u_i v_i^T : u_i, v_i \in \mathbb{R}^p, i \in [n_1]\right\} \cup \{A_j : A_j \in \mathbb{R}^{p\times p}, j \in [n_2]\}$ *and constant matrix* $\{I_p\}$ *such that* $\deg(M) = \deg(M') + 1, \deg(A_j; M) = \deg(A_j; M'), j \in [n_2], \deg(u_i v_i^T; M) = \deg(u_i v_i^T; M'), i \in [2 : n_1]$ *and* $\deg(u_1 v_1^T; M) = \deg(u_1 v_1^T; M') + 1$.

Proof. By the definition of multiplicative terms, we know that there exist two multiplicative terms M_1, M_2 of parameter matrices $\left\{u_i v_i^T : u_i, v_i \in \mathbb{R}^p, i \in [n_1]\right\} \cup \{A_j : A_j \in \mathbb{R}^{p\times p}, j \in [n_2]\}$ and constant matrix $\{I_p\}$ such that

$$M = M_1 u_1 v_1^T M_2,$$

where $\deg(M) = \deg(M_1) + \deg(M_2) + 1, \deg(A_j; M) = \deg(A_j; M_1) + \deg(A_j; M_2), j \in [n_2], \deg(u_i v_i^T; M) = \deg(u_i v_i^T; M_1) + \deg(u_i v_i^T; M_2), i \in [2 : n_1]$ and $\deg(u_1 v_1^T; M) = \deg(u_1 v_1^T; M_1) + \deg(u_1 v_1^T; M_2) + 1$. Therefore we have

$$\operatorname{tr}(M) = \operatorname{tr}\left(M_1 u_1 v_1^T M_2\right) = \operatorname{tr}\left(v_1^T M_2 M_1 u_1\right) = v_1^T M_2 M_1 u_1.$$

Note that $M' = M_2 M_1$ satisfies that $\deg(M') = \deg(M_1) + \deg(M_2), \deg(A_j, M') = \deg(A_j; M_1) + \deg(A_j; M_2), j \in [n_2], \deg(u_i v_i^T; M) = \deg(u_i v_i^T; M_1) + \deg(u_i v_i^T; M_2), i \in [2 : n_1]$ and $\deg(u_1 v_1^T; M') = \deg(u_1 v_1^T; M_1) + \deg(u_1 v_1^T; M_2) + 1$. We have finished the proof. □

The following two lemmas focus on the expectation of the product of quadratic forms of the standard normal samples. Lemma 6 focuses on single sample while 7 focuses on the same form with b i.i.d. samples drawn from the standard normal distribution.

Lemma 6. *Given matrices* $A_j \in \mathbb{R}^{p\times p}, j \in [m-1]$*, we have*

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_p)}\left[x x^T A_1 x x^T A_2 \cdots A_{m-1} x x^T\right] = \sum_{i=1}^{N_m} \prod_{k=1}^{n_i} \operatorname{tr}(M_{ik})\, M_{i0},$$

where N_m *and* $n_i, i \in [N_m]$ *are constants depending on* m *and* $\{M_{ik}, k \in [0 : n_i], i \in [N_m]\}$ *are multiplicative terms of parameter matrices* $\{A_j, j \in [m-1]\}$ *and constant matrix* $\{I_p\}$*. Furthermore, for every* $i \in [N_m]$*, we have* $\sum_{k=0}^{n_i} \deg(A_j; M_{ik}) = 1, j \in [m-1]$ *and therefore* $\sum_{k=0}^{n_i} \deg(M_{ik}) = m - 1$.

Proof. See Magnus (1978). □

Lemma 7. *We are given matrices* $A_j \in \mathbb{R}^{p\times p}, j \in [m-1]$ *and random vectors* $x_i, i \in [b]$ *independently and identically drawn from* $\mathcal{N}(0, I_p)$*. We assume that the multi-set* $\mathcal{S} = \left\{i_j, i'_j : j \in [m]\right\}$ *satisfies that for every* $i \in \mathcal{S}$*,* i *is an element of* $[b]$ *and the number of appearance of* i *in* $\mathcal{S}$ *is even. Then*

$$\mathbb{E}_{x_i\sim\mathcal{N}(0,I_p)}\left[x_{i_1} x_{i'_1}^T A_1 x_{i_2} x_{i'_2}^T A_2 \cdots A_{m-1} x_{i_m} x_{i'_m}^T\right] = \sum_{i=1}^{N_m} \prod_{k=1}^{n_i} \operatorname{tr}(M_{ik})\, M_{i0}, \tag{5}$$

where N_m *and* n_i *are constants depending on* m *(and independent of* b*) and* $M_{ik}, k \in [0 : n_i], i \in [N_m]$ *are multiplicative terms of parameter matrices* $\{A_j, j \in [m-1]\}$ *and constant matrix* $\{I_p\}$*. Furthermore, for every* $i \in [N_m]$*, we have* $\sum_{k=0}^{n_i} \deg(A_j; M_{ik}) = 1, j \in [m-1]$ *and therefore* $\sum_{k=0}^{n_i} \deg(M_{ik}) = m - 1$.

Proof. Let $\beta_i, i \in [b]$ be the number of appearances of i in $\mathcal{S}$, which are even by assumption. We induct on the quantity $N = \sum_{i=1}^{b} \mathbb{1}\{\beta_i \neq 0\}$.

For the base case of $N = 1$, all elements in the multi-set $\mathcal{S}$ have the same value. Without loss of generality, we assume $i_j = i'_j = 1, j \in [m]$. Then

$$\mathbb{E}_{x_i \sim \mathcal{N}(0,I_p)} \left[x_{i_1} x_{i'_1}^T A_1 x_{i_2} x_{i'_2}^T \cdots A_{m-1} x_{i_m} x_{i'_m}^T \right] = \mathbb{E}_{x_1 \sim \mathcal{N}(0,I_p)} \left[x_1 x_1^T A_1 x_1 x_1^T \cdots A_{m-1} x_1 x_1^T \right],$$

which is the statement of Lemma 6.

Suppose the statement holds for $N \geqslant 1$, and we consider the case of $N+1$. Note that $x_{i'_j}^T A_j x_{i_{j+1}} = x_{i_{j+1}}^T A_j x_{i'_j}$ is a scalar so that we can move it around without changing the value of the expression[2]. We distinguish two cases.

- Let $i_1 \neq i'_m$. Without loss of generality, we assume $i_1 = 1$. We can always change the order of $x_{i'_j}^T A_j x_{i_{j+1}}, j \in [m-1]$ (and flip it to be $x_{i_{j+1}}^T A_j x_{i'_j}$ if necessary) such that all x_1's appear in the form of $x_1 x_1^T$:

$$\begin{aligned} x_{i_1} x_{i'_1}^T A_1 x_{i_2} x_{i'_2}^T A_2 \cdots A_{m-1} x_{i_m} x_{i'_m}^T &= x_1 \left(x_{i'_1}^T A_1 x_{i_2} x_{i'_2}^T A_2 \cdots A_{m-1} x_{i_m} \right) x_{i'_m}^T \\ &= x_1 x_1^T \widetilde{A}_1 x_1 x_1^T \widetilde{A}_2 \cdots \widetilde{A}_{\frac{\beta_1}{2}-1} x_1 x_1^T \widetilde{A}_{\frac{\beta_1}{2}} \widetilde{x} x_{i'_m}^T \end{aligned}$$

  where $\widetilde{x} \in \{x_i, i \in [b]\}, \widetilde{x} \neq x_1$ and $\widetilde{A}_i$'s are multiplicative terms of parameter matrices $\{x_u x_v^T : u, v \in [2:b]\} \cup \{A_j : j \in [m-1]\}$ and constant matrix $\{I_p\}$ such that $\sum_{u,v \in [2:b]} \sum_{k=1}^{\frac{\beta_1}{2}} \deg(x_u x_v^T; \widetilde{A}_k) = m - \frac{\beta_1}{2} - 1$ and $\sum_{k=1}^{\frac{\beta_1}{2}} \deg(A_j; \widetilde{A}_k) = 1, j \in [m-1]$[3].

  Applying Lemma 6 and the law of iterative expectations, we have

$$\begin{aligned} \mathbb{E}_{x_i \sim \mathcal{N}(0,I_p)} \left[x_{i_1} x_{i'_1}^T A_1 x_{i_2} x_{i'_2}^T \cdots A_{m-1} x_{i_m} x_{i'_m}^T \right] &= \mathbb{E}_{x_1, \cdots, x_b} \left[x_1 x_1^T \widetilde{A}_1 x_1 x_1^T \widetilde{A}_2 \cdots \widetilde{A}_{\frac{\beta_1}{2}-1} x_1 x_1^T \widetilde{A}_{\frac{\beta_1}{2}} \widetilde{x} x_{i'_m}^T \right] \\ &= \mathbb{E}_{x_2, \cdots, x_b} \left[\left(\sum_{i=1}^{N_m} \prod_{k=1}^{n_i} \operatorname{tr}(M_{ik}) M_{i0} \right) \widetilde{A}_{\frac{\beta_1}{2}} \widetilde{x} x_{i'_m}^T \right] \\ &= \sum_{i=1}^{N_m} \mathbb{E}_{x_2, \cdots, x_b} \left[\left(\prod_{k=1}^{n_i} \operatorname{tr}(M_{ik}) M_{i0} \right) \widetilde{A}_{\frac{\beta_1}{2}} \widetilde{x} x_{i'_m}^T \right], \end{aligned}$$

  where N_m and n_i are constant depending on m (and independent of b) and $M_{ik}, k \in [0:n_i], i \in [N_m]$ are multiplicative terms of parameter matrices $\left\{ \widetilde{A}_j, j \in [\frac{\beta_1}{2} - 1] \right\}$ and constant matrix $\{I_p\}$. Furthermore, for every $i \in [N_m]$, we have $\sum_{k=0}^{n_i} \deg(\widetilde{A}_j; M_{ik}) = 1, j \in [\frac{\beta_1}{2} - 1]$ and therefore $\sum_{k=0}^{n_i} \deg(M_{ik}) = \frac{\beta_1}{2} - 1$.

  Combining the definition of $\widetilde{A}_j$'s, we know that $M_{ik}, k \in [0:n_i], i \in [N_m]$ are multiplicative terms of parameter matrices $\{x_u x_v^T : u, v \in [2:b]\} \cup \{A_j : j \in [m-1]\}$ and constant

[2] For example, we can rewrite

$$\begin{aligned} x_{i_1} x_{i'_1}^T A_1 x_{i_2} x_{i'_2}^T A_2 x_{i_3} x_{i'_3}^T &= x_{i_1} \left(x_{i'_1}^T A_1 x_{i_2} \right) \left[x_{i'_2}^T A_2 x_{i_3} \right] x_{i'_3}^T = x_{i_1} \left[x_{i'_2}^T A_2 x_{i_3} \right] \left(x_{i'_1}^T A_1 x_{i_2} \right) x_{i'_3}^T \\ &= x_{i_1} \left[x_{i'_2}^T \left(x_{i'_1}^T A_1 x_{i_2} \right) A_2 x_{i_3} \right] x_{i'_3}^T = x_{i_1} \left[x_{i'_2}^T A_2 \left(x_{i'_1}^T A_1 x_{i_2} \right) x_{i_3} \right] x_{i'_3}^T. \end{aligned}$$

[3] For example, we can rewrite

$$x_1 x_2^T A_1 x_1 x_1^T A_2 x_3 x_3^T A_3 x_1 x_2 = x_1 \left(x_2^T A_1 x_1 \right) \left[x_1^T A_2 x_3 \right] \left\{ x_3^T A_3 x_1 \right\} x_2 = x_1 \left(x_1^T A_1 x_2 \right) \left[x_3^T A_2 x_1 \right] \left\{ x_1^T A_3 x_3 \right\} x_2$$

$$= x_1 x_1^T A_1 x_2 x_3^T A_2 x_1 x_1^T A_3 x_3 x_2 = x_1 x_1^T \widetilde{A}_1 x_1 x_1^T \widetilde{A}_2 \widetilde{x} x_2,$$

where $\widetilde{A}_1 = A_1 x_2 x_3^T A_2$, $\widetilde{A}_2 = A_3$ and $\widetilde{x} = x_3$. Besides, $m = 4, \beta_1 = 4$, thus the degree of $x_u x_v^T$ in all $\widetilde{A}_k$ sum up to $m - \frac{\beta_1}{2} - 1 = 1$

matrix $\{I_p\}$ such that for every $i \in [N_m]$, we have $\sum_{u,v\in[2:b]} \sum_{k=0}^{n_i} \deg(x_u x_v^T; M_{ik}) = m - \frac{\beta_1}{2} - 1$ and $\sum_{k=0}^{n_i} \deg(A_j; M_{ik}) = 1, j \in [m-1]$.

Applying Lemma 5, for every $k \in [0:n_i]$ and every $i \in [N_m]$, there exists $u_{ik}, v_{ik} \in \{x_j : j \in [2:b]\}$ and multiplicative term M'_{ik} of parameter matrices $\{x_u x_v^T : u, v \in [2:b]\} \cup \{A_j : j \in [m-1]\}$ and constant matrix $\{I_p\}$ such that

$$\operatorname{tr}(M_{ik}) = u_{ik}^T M'_{ik} v_{ik}.$$

Therefore, we have

$$\left(\prod_{k=1}^{n_i} \operatorname{tr}(M_{ik})\, M_{i0}\right) \widetilde{A}_{\frac{\beta_1}{2}} \widetilde{x} x_{i'_m}^T = \prod_{k=1}^{n_i} \left(u_{ik}^T M'_{ik} v_{ik}\right) M_{i0} \widetilde{A}_{\frac{\beta_1}{2}} \widetilde{x} x_{i'_m}^T = M_{i0} \widetilde{A}_{\frac{\beta_1}{2}} \widetilde{x} \prod_{k=1}^{n_i} \left(u_{ik}^T M'_{ik} v_{ik}\right) x_{i'_m}^T \triangleq U_i.$$

Note that for every $i \in [N_m]$, we have

$$\sum_{j=1}^{m-1} \deg(x_i; A_j) = \sum_{k=1}^{n_i} \deg(x_i; M'_{ik}) + \deg(x_i; M_{i0}) + \deg\left(x_i; \widetilde{A}_{\frac{\beta_1}{2}}\right) + \deg(x_i; \widetilde{x}) + \deg\left(x_i; x_{i'_m}^T\right),$$

and for every $j \in [m-1]$, we have

$$\sum_{k=1}^{n_i} \deg(A_j; M'_{ik}) + \deg(A_j; M_{i0}) + \deg\left(A_j; \widetilde{A}_{\frac{\beta_1}{2}}\right) = 1.$$

In other words, for every $i \in [N_m]$, U_i has the form of $\widehat{A}_0 x_{\widehat{i}_1} x_{\widehat{i}'_1}^T \widehat{A}_1 x_{\widehat{i}_2} x_{\widehat{i}'_2}^T \cdots \widehat{A}_{m-1} x_{\widehat{i}_{m'}} x_{i'_m}^T \widehat{A}_{m'}$ but there is no appearance of x_1. Here $x_{\widehat{i}_j}, x_{\widehat{i}_j} \in \{x_j, j \in [2:b]\}$, and $\widehat{A}_i, i \in [0:m]$ are multiplicative terms of parameter matrices $\{A_j, j \in [m-1]\}$ and constant matrix $\{I_p\}$. Furthermore, for every $j \in [m-1]$, we have $\sum_{k=0}^{n_i} \deg(A_j; \widehat{A}_i) = 1$. Note that here we use the liberty of adding identity matrices if more than two consecutive x's appear. Since we have reduced $N+1$ by one, we can use induction on $x_{\widehat{i}_1} x_{\widehat{i}'_1}^T \widehat{A}_1 x_{\widehat{i}_2} x_{\widehat{i}'_2}^T \cdots \widehat{A}_{m-1} x_{\widehat{i}_{m'}} x_{i'_m}^T$ and finish the proof. The two constant matrices $\widehat{A}_0$ and $\widehat{A}_m$ do not change the result of expectation since $\mathbb{E}\left(\widehat{A}_0 X \widehat{A}_{m'}\right) = \widehat{A}_0 \mathbb{E}(X) \widehat{A}_{m'}$.

- If $i_1 = i'_m$, without loss of generality we assume, $i'_1 = 1$ and $i'_1 \neq i_1$ (note that all $x_{i'_j}^T A_j x_{i_{j+1}}, j \in [m-1]$ are inter-changeable and there is at least one element in $\mathcal{S}$ that is not equal to i_1). We change the orders of $x_{i'_j}^T A_j x_{i_{j+1}}, j \in [m-1]$ (and flip it to be $x_{i_{j+1}}^T A_j x_{i'_j}$ if necessary) such that all x_1's appear in a consecutive form of $x_1 x_1^T$:

$$\begin{aligned} x_{i_1} x_{i'_1}^T A_1 x_{i_2} x_{i'_2}^T A_2 \cdots A_{m-1} x_{i_m} x_{i'_m}^T &= x_{i_1} \left(x_{i'_1}^T A_1 x_{i_2} x_{i'_2}^T A_2 \cdots A_{m-1} x_{i_m}\right) x_{i'_m}^T \\ &= x_{i_1} \left(\widetilde{x}_1^T \widetilde{A}_0 \left[x_1 x_1^T \widetilde{A}_1 \cdots \widetilde{A}_{\frac{\beta_1}{2}-1} x_1 x_1^T\right] \widetilde{A}_{\frac{\beta_1}{2}} \widetilde{x}_2\right) x_{i'_m}^T, \end{aligned}$$

where $\widetilde{x}_1, \widetilde{x}_2 \in \{x_i, i \in [b]\}, \widetilde{x}_1, \widetilde{x}_2 \neq x_1$ and $\widetilde{A}_i$'s are multiplicative terms of parameter matrices $\{x_u x_v^T : u, v \in [2:b]\} \cup \{A_j : j \in [m-1]\}$ and constant matrix $\{I_p\}$ such that

$$\sum_{u,v\in[2:b]} \sum_{k=0}^{\frac{\beta_1}{2}} \deg(x_u x_v^T; \widetilde{A}_k) = m - \frac{\beta_1}{2} - 2$$

and $\sum_{k=0}^{\frac{\beta_1}{2}} \deg(A_j; \widetilde{A}_k) = 1, j \in [m-1]$. The remaining reasoning is the same as the previous case.

□

Remark. If one of the β_i numbers of appearance of $x_j, j \in [b]$ is odd, then it is easy to see that the result in (5) is the zero matrix.

As pointed out in the Section 1, the difficulty of studying the dynamics of SGD is how to connect the quantities in iteration t with fixed variables, like initial weights $W_{0,1}, W_{0,2}$ and mini-batch size b. We overcome this challenge by the following two lemmas. Lemma 8 provides the relationship between $g_{t,i}^b, i = 1, 2$ and $W_{t,i}^b, i = 1, 2$ by taking expectation over the distribution of random samples in $\mathcal{B}_t^b$. Lemma 9 shows the relationship between $W_{t,i}^b, i = 1, 2$ and $g_{t-1,i}^b, i = 1, 2$ using (1) and (2).

Lemma 8. *For multiplicative terms $M_i, i \in [0:m]$ of parameter matrices $\left\{g_{t,1}^b, g_{t,2}^b\right\}$ and constant matrices $\left\{W_{t,1}^b, W_{t,2}^b, W_1^*, W_2^*\right\}$ with degree d_i, respectively, we denote $M = \prod_{i=1}^m \operatorname{tr}(M_i) M_0$ and $d = \sum_{i=0}^m d_i$. There exists a set of multiplicative terms $\left\{M_{ij}^k, i \in [m_k], j \in [0 : m_{ki}], k \in [0:q]\right\}$ of parameter matrices $\left\{W_{t,1}^b, W_{t,2}^b\right\}$ and constant matrices $\{W_1^*, W_2^*\}$ such that*

$$\mathbb{E}\left[M \middle| \mathcal{F}_t^b\right] = N_0 + N_1 \frac{1}{b} + \cdots + N_d \frac{1}{b^d},$$

where $N_k = \sum_{i=1}^{m_k} \prod_{j=1}^{m_{ki}} \operatorname{tr}\left(M_{ij}^k\right) M_{i0}^k, k \in [0:d]$. Here m_k, m_{ki} are constants independent of b, and $\sum_{j=0}^{m_{ki}} \deg\left(M_{ij}^k\right) \leqslant 3d + \sum_{i=0}^m \left(\deg\left(W_{t,1}^b; M_i\right) + \deg(W_{t,2}^b; M_i)\right)$.

Lemma 9. *For multiplicative term $M_i, i \in [0:m]$ of parameter matrices $\left\{W_{t,1}^b, W_{t,2}^b\right\}$ and constant matrices $\{W_1^*, W_2^*\}$ of degree d_i, let $d = 2^{d_0 + \cdots + d_m}$. There exists a set of multiplicative terms $\{M_{ik}, i \in [0:m], k \in [d]\}$ of parameter matrices $\left\{g_{t,1}^b, g_{t,2}^b\right\}$ and constant matrices $\left\{W_{t,1}^b, W_{t,2}^b, W_1^*, W_2^*\right\}$ such that*

$$\prod_{i=1}^m \operatorname{tr}(M_i) M_0 = \sum_{k=1}^d \prod_{i=1}^m \operatorname{tr}(M_{ik}) M_{0k},$$

where $\sum_{i=0}^m \deg(M_{ik}) \leqslant d$.

Proof of Lemma 8. By (1) and (2) we have

$$M = \prod_{i=1}^m \operatorname{tr}(M_i) M_0 = \frac{1}{b^d} \sum_{k=1}^{b^d} \prod_{i=1}^m \operatorname{tr}(M_{ki}) M_{k0}, \tag{6}$$

where each $M_{ki}, k \in [b^d], i \in [0:m]$ is a multiplicative term of parameter matrices $\left\{x_{t,i} x_{t,i}^T, i \in [b]\right\}$ and constant matrices $\left\{W_{t,1}^b, W_{t,2}^b, \mathcal{W}_t^b\right\}$. Let $\widetilde{M}_k = \prod_{i=1}^m \operatorname{tr}(M_{ki}) M_{k0}, k \in \left[b^d\right]$. We split set $\left\{\widetilde{M}_k : k \in \left[b^d\right]\right\}$ into disjoint and non-empty sets (equivalent classes) $S_1, \ldots, S_{n_M}$ such that

1. for every $i \in [n_M]$ and every $M_1, M_2 \in S_i$, we have $\mathbb{E}\left[M_1 \middle| \mathcal{F}_t^b\right] = \mathbb{E}\left[M_2 \middle| \mathcal{F}_t^b\right]$,
2. for every $i, j \in [n_M], i \neq j$ and every $M_1 \in S_i$ and $M_2 \in S_j$, we have $\mathbb{E}\left[M_1 \middle| \mathcal{F}_t^b\right] \neq \mathbb{E}\left[M_2 \middle| \mathcal{F}_t^b\right]$.

Note that $\cup_{i=1}^{n_M} S_i = \left\{\widetilde{M}_k : k \in \left[b^d\right]\right\}$. Let $\widehat{M}_k \in S_k$ represent the equivalent class S_k (it can be any member of S_k). For every $i \in [n_M]$, we can always write $|S_i| = e_{i,0} + e_{i,1} b + \cdots + e_{i,d} b^d$ such that $e_{i,j} \in \mathbb{N}, e_{i,j} < b, j \in [0:d]$ (actually $e_{i,j}$'s are the digits of the base-b representation of $|S_i|$). Then

we have

$$\begin{aligned}\mathbb{E}\left[M\middle|\mathcal{F}_t^b\right] &= \mathbb{E}\left[\frac{1}{b^d}\sum_{k=1}^{b^d}\widetilde{M}_k\middle|\mathcal{F}_t^b\right] = \frac{1}{b^d}\mathbb{E}\left[\sum_{i=1}^{n_M}\left(e_{i,0}+e_{i,1}b+\cdots+e_{i,d}b^d\right)\widehat{M}_i\middle|\mathcal{F}_t^b\right] \\ &= \frac{1}{b^d}\sum_{i=1}^{n_M}\left(e_{i,0}+e_{i,1}b+\cdots+e_{i,d}b^d\right)\mathbb{E}\left[\widehat{M}_i\middle|\mathcal{F}_t^b\right] \\ &= \sum_{i=1}^{n_M}\left(e_{i,d}+e_{i,d-1}\frac{1}{b}+\cdots+e_{i,0}\frac{1}{b^d}\right)\mathbb{E}\left[\widehat{M}_i\middle|\mathcal{F}_t^b\right].\end{aligned} \tag{7}$$

It is important to note that n_M, the number of different equivalent classes, is independent of b. This follows from the fact that each $\mathbb{E}\left[\widetilde{M}_k\middle|\mathcal{F}_t^b\right]$ (and so as $\mathbb{E}\left[\widehat{M}_k\middle|\mathcal{F}_t^b\right]$) includes a finite number of weight matrices $W_{t,1}^b$ and $W_{t,2}^b$ with degree less than or equal to $3d + \sum_{i=0}^{m}\left(\deg\left(W_{t,1}^b;M_i\right)+\deg(W_{t,2}^b;M_i)\right)$ (see Lemma 7). Thus the number of partition sets is bounded by a quantity independent of b.

Note that each M_{ki} can be represented as

$$M_{ki} = A_0^{ki}x_{t,i_1}^{ki}{x_{t,i_1}^{ki}}^T A_1^{ki}\cdots A_{d_i-1}^{ki}x_{t,i_{d_i}}^{ki}{x_{t,i_{d_i}}^{ki}}^T A_{d_i}^{ki}$$

for some matrices $A_0^{ki},\ldots,A_{d_i}^{ki}$ that are multiplicative term of parameter matrices $\left\{W_{t,1}^b, W_{t,2}^b and \mathcal{W}_t^b\right\}$ constant matrix $\{I_p\}$ (we stress again that some A matrices can be identities, based on the definition of multiplicative terms), and $x_{t,i_1}^{ki},\ldots,x_{t,i_{d_i}}^{ki}\in\{x_{t,1},\ldots,x_{t,b}\}$. We have

$$\begin{aligned}\operatorname{tr}\left(M_{ki}\right) &= \operatorname{tr}\left(A_0^{ki}x_{t,i_1}^{ki}{x_{t,i_1}^{ki}}^T A_1^{ki}\cdots A_{d_i-1}^{ki}x_{t,i_{d_i}}^{ki}{x_{t,i_{d_i}}^{ki}}^T A_{d_i}^{ki}\right) \\ &= {x_{t,i_{d_i}}^{ki}}^T A_{d_i}^{ki}A_0^{ki}x_{t,i_1}^{ki}{x_{t,i_1}^{ki}}^T A_1^{ki}\cdots A_{d_i-1}^{ki}x_{t,i_{d_i}}^{ki}.\end{aligned}$$

For every $k\in\left[b^d\right]$, we have

$$\begin{aligned}\prod_{i=1}^{m}\operatorname{tr}\left(M_{ki}\right)M_{k0} &= \left[\prod_{i=1}^{m}{x_{t,i_{d_i}}^{ki}}^T A_{d_i}^{ki}A_0^{ki}x_{t,i_1}^{ki}{x_{t,i_1}^{ki}}^T A_1^{ki}\cdots A_{d_i-1}^{ki}x_{t,i_{d_i}}^{ki}\right]A_0^{k0}x_{t,i_1}^{k0}{x_{t,i_1}^{k0}}^T A_1^{k0}\cdots A_{d_0-1}^{k0}x_{t,i_{d_0}}^{k0}{x_{t,i_{d_0}}^{k0}}^T A_{d_0}^{k0} \\ &= \left[\prod_{i=1}^{m}{x_{t,i_{d_i}}^{ki}}^T A_{d_i}^{ki}A_0^{ki}x_{t,i_1}^{ki}{x_{t,i_1}^{ki}}^T A_1^{ki}\cdots A_{d_i-1}^{ki}x_{t,i_{d_i}}^{ki}\right]\left[{x_{t,i_1}^{k0}}^T A_1^{k0}\cdots A_{d_0-1}^{k0}x_{t,i_{d_0}}^{k0}\right]A_0^{k0}x_{t,i_1}^{k0}{x_{t,i_{d_0}}^{k0}}^T A_{d_0}^{k0},\end{aligned}$$

which can be rewritten as

$$\widetilde{M}_k = \prod_{i=1}^{m}\operatorname{tr}\left(M_{ki}\right)M_{k0} = \left(\prod_{j=1}^{d}x_{t,\bar{i}_j}^T A_j^k x_{t,\bar{i}'_j}\right)A_0^{k0}x_{t,i_1}^{k0}{x_{t,i_{d_0}}^{k0}}^T A_{d_0}^{k0}.$$

Note that the randomness of each $\widetilde{M}_k$ given $\mathcal{F}_t^b$ only comes from the randomness of $x_{t,j}$'s, i.e. for all $k\in\left[b^d\right]$ we have

$$\begin{aligned}\mathbb{E}\left[\widetilde{M}_k\middle|\mathcal{F}_t^b\right] &= \mathbb{E}_{x_{t,j}\sim\mathcal{N}(0,I)}\left[\left(\prod_{j=1}^{d}x_{t,i_j}^T A_j^k x_{t,i'_j}\right)A_0^k x_{t,i'_0}x_{t,i_0}^T A_0^{k'}\right] \\ &= \mathbb{E}_{x_{t,j}\sim\mathcal{N}(0,I)}\left[A_0^k x_{t,i'_0}\left(\prod_{j=1}^{d}x_{t,i_j}^T A_j^k x_{t,i'_j}\right)x_{t,i_0}^T A_0^{k'}\right] \\ &= \sum_{i=1}^{n_M^k}\prod_{j=1}^{n_i^k}\operatorname{tr}\left(\widetilde{M}_{ij}^k\right)\widetilde{M}_{i0}^k,\end{aligned} \tag{8}$$

where the last equation comes from Lemma 7. Here $n_M^k, n_i^k, i \in \left[n_M^k\right], k \in \left[b^d\right]$ are constants independent of b, M_{ij}^k's are multiplicative terms of parameter matrices $\left\{W_{t,1}^b, W_{t,2}^b, \mathcal{W}_t^b\right\}$ and constant matrix $\{I_p\}$ such that for every $i \in \left[n_M^k\right]$, we have

$$\sum_{j=0}^{n_i^k} \deg\left(\mathcal{W}_t^b; \widetilde{M}_{ij}^k\right) = d \tag{9}$$

and

$$\sum_{j=0}^{n_i^k} \left(\deg\left(W_{t,1}^b; \widetilde{M}_{ij}^k\right) + \deg\left(W_{t,2}^b; \widetilde{M}_{ij}^k\right)\right) = d + \sum_{r=0}^{m} \left(\deg\left(W_{t,1}^b; M_r\right) + \deg(W_{t,2}^b; M_r)\right). \tag{10}$$

These degree relationships can be observed from (1), (2), and the fact that each $g_{t,1}^b$ or $g_{t,1}^b$ contributes one $\mathcal{W}_t^b$ and one of $W_{t,1}^b$ or $W_{t,2}^b$ in $\prod_{j=1}^{n_i^k} \operatorname{tr}\left(\widetilde{M}_{ij}^k\right) \widetilde{M}_{i0}^k$. Note that $\mathcal{W}_t = W_{t,2}^b W_{t,2}^b - W_2^* W_1^*$. For every $i \in \left[n_M^k\right]$, if we replace all appearances of $\mathcal{W}_t^b$ in $\prod_{j=1}^{n_i^k} \operatorname{tr}\left(\widetilde{M}_{ij}^k\right) \widetilde{M}_{i0}^k$ and expand all parentheses of $\left(W_{t,2}^b W_{t,2}^b - W_2^* W_1^*\right)$, we have

$$\prod_{j=1}^{n_i^k} \operatorname{tr}\left(\widetilde{M}_{ij}^k\right) \widetilde{M}_{i0}^k = \sum_{l=1}^{2^d} \prod_{j=1}^{n_i^k} \operatorname{tr}\left(\widetilde{M}_{ij}^{kl}\right) \widetilde{M}_{i0}^{kl}, \tag{11}$$

where $\widetilde{M}_{ij}^{kl}$'s are multiplicative terms of parameter matrices $\left\{W_{t,1}^b, W_{t,2}^b\right\}$ and constant matrices $\{W_1^*, W_2^*\}$ such that

$$\sum_{j=0}^{n_i^k} \left(\deg\left(W_{t,1}^b; \widetilde{M}_{ij}^{kl}\right) + \deg\left(W_{t,2}^b; \widetilde{M}_{ij}^{kl}\right)\right) \leqslant 3d + \sum_{r=0}^{m} \left(\deg\left(W_{t,1}^b; M_r\right) + \deg(W_{t,2}^b; M_r)\right), \tag{12}$$

where the inequality comes from (9) and (10) and the fact that each $g_{t,1}^b$ or $g_{t,2}^b$ contributes 2 or 0 degrees in the form of $W_{t,2}^b W_{t,1}^b$ or $W_2^* W_1^*$, respectively.

Combining (7), (8) and (11), we have

$$\begin{aligned}
\mathbb{E}\left[M \middle| \mathcal{F}_t^b\right] &= \sum_{k=1}^{n_M} \left(e_{k,d} + e_{k,d-1}\frac{1}{b} + \cdots + e_{k,0}\frac{1}{b^d}\right) \mathbb{E}\left[\widehat{M}_k \middle| \mathcal{F}_t^b\right] \\
&= \sum_{k=1}^{n_M} \left(e_{k,d} + e_{k,d-1}\frac{1}{b} + \cdots + e_{k,0}\frac{1}{b^d}\right) \sum_{i=1}^{n_M^{s_k}} \sum_{l=1}^{2^d} \prod_{j=1}^{n_i^k} \operatorname{tr}\left(\widetilde{M}_{ij}^{kl}\right) \widetilde{M}_{i0}^{kl} \\
&= N_0 + N_1\frac{1}{b} + \cdots + N_d\frac{1}{b^d},
\end{aligned}$$

where

$$N_r = \sum_{k=1}^{n_M} e_{k,d-r} \left(\sum_{i=1}^{n_M^{s_k}} \sum_{l=1}^{2^d} \prod_{j=1}^{n_i^k} \operatorname{tr}\left(\widetilde{M}_{ij}^{kl}\right) \widetilde{M}_{i0}^{kl}\right). \tag{13}$$

Note that all constants in (13) are independent of b and combining with (12), we have finished the proof.

□

Proof of Lemma 9. Simply using the fact that $W_{t,i}^b = W_{t-1,i}^b - \alpha_t g_{t-1,i}^b, i = 1, 2$, if we replace each $W_{t,i}^b$ in the left-hand-side of (13) by $W_{t-1,i}^b - \alpha_t g_{t-1,i}^b$ and expand all the parentheses, then each $M_i, i \in [0:m]$ becomes the sum of 2^{d_i} multiplicative terms of parameter matrices $\left\{g_{t,1}^b, g_{t,2}^b\right\}$ and constant matrices $\left\{W_{t,1}^b, W_{t,2}^b, W_1^*, W_2^*\right\}$ with degree at most d_i. As a result, $\prod_{i=1}^m \operatorname{tr}(M_i) M_0$ becomes the sum of 2^d terms in the form of $\prod_{i=1}^m \operatorname{tr}(M_{ik}) M_{0k}$ where $\deg(M_{ik}) \leqslant 2^{d_i}$, and therefore $\sum_{i=0}^m \deg(M_{ik}) \leqslant \prod_{i=0}^m 2^{d_i} = d$. □

Proof of Theorem 3. We use induction on t to show this result. The base case of $t = 0$ it is the same as the statement in Lemma 8.

Suppose that the statement holds for $t \geqslant 0$, and we consider the case of $t + 1$. By Lemma 8, there exists a set of multiplicative terms $\left\{M^k_{t+1,i,j}, i \in [m_{t+1,k}], j \in [0 : m_{t+1,k,i}], k \in [0 : d]\right\}$ of parameter matrices $\left\{W^b_{t+1,1}, W^b_{t+1,2}\right\}$ and constant matrices $\{W^*_1, W^*_2\}$ such that

$$\mathbb{E}\left[M\middle|\mathcal{F}^b_{t+1}\right] = N_{t+1,0} + N_{t+1,1}\frac{1}{b} + \cdots + N_{t+1,d}\frac{1}{b^d}, \tag{14}$$

where $N_{t+1,k} = \sum_{i=1}^{m_{t+1,k}} \prod_{j=1}^{m_{t+1,k,i}} \operatorname{tr}\left(M^k_{t+1,i,j}\right) M^k_{t+1,i,0}, k \in [0 : d]$. Here $m_{t+1,k}, m_{t+1,k,i}$ are constants independent of b, and $\sum_{j=0}^{m_{t+1,k,i}} \deg\left(M^k_{t+1,i,j}\right) \leqslant 3d + d'$.

For each $i \in [m_{t+1,k}]$ and each $k \in [0 : d]$, by Lemma 9, there exists a set of multiplicative terms $\{M_{t,i,j,k,l}, j \in [m_{t+1,i,k}], l \in [d_{t,i,k}]\}$ of parameter matrices $\left\{g^b_{t,1}, g^b_{t,2}\right\}$ and constant matrices $\left\{W^b_{t,1}, W^b_{t,2}, W^*_1, W^*_2\right\}$ such that

$$\prod_{j=1}^{m_{t+1,k,i}} \operatorname{tr}\left(M^k_{t+1,i,j}\right) M^k_{t+1,i,0} = \sum_{l=1}^{d_{t,i,k}} \prod_{j=1}^{m_{t+1,k,i}} \operatorname{tr}\left(M_{t,i,j,k,l}\right) M_{t,i,0,k,l}, \tag{15}$$

where $d_{t,i,k} = 2^{\sum_{j=0}^{m_{t+1,k,i}} \left(\deg\left(W^b_{t,1}; M_{t,i,j,k,l}\right) + \deg\left(W^b_{t,2}; M_{t,i,j,k,l}\right)\right)}$ is a constant independent of b and

$$\sum_{j=0}^{m_{t+1,k,i}} \deg\left(M_{t,i,j,k,l}\right) \leqslant 3d + d', \tag{16}$$

and

$$\sum_{j=0}^{m_{t+1,k,i}} \left(\deg\left(W_{t,1}; M_{t,i,j,k,l}\right) + \deg\left(W_{t,2}; M_{t,i,j,k,l}\right)\right) \leqslant 3d + d'. \tag{17}$$

Combining (14) and (15), we have for every $k \in [0 : d]$

$$N_{t+1,k} = \sum_{i=1}^{m_{t+1,k}} \sum_{l=1}^{d_{t,i,k}} \prod_{j=1}^{m_{t+1,k,i}} \operatorname{tr}\left(M_{t,i,j,k,l}\right) M_{t,i,0,k,l}. \tag{18}$$

Note that

$$\begin{aligned}
\mathbb{E}\left[M|\mathcal{F}_0\right] = \mathbb{E}\left[\mathbb{E}\left[M\middle|\mathcal{F}^b_{t+1}\right]\middle|\mathcal{F}_0\right] &= \mathbb{E}\left[N_{t+1,0}|\mathcal{F}_0\right] + \mathbb{E}\left[N_{t+1,1}|\mathcal{F}_0\right]\frac{1}{b} + \cdots + \mathbb{E}\left[N_{t+1,d}|\mathcal{F}_0\right]\frac{1}{b^d} \\
&= \sum_{i=1}^{m_{t+1,0}} \sum_{l=1}^{d_{t,i,0}} \mathbb{E}\left[\prod_{j=1}^{m_{t+1,0,i}} \operatorname{tr}\left(M_{t,i,j,0,l}\right) M_{t,i,0,0,l}\middle|\mathcal{F}_0\right] + \\
&+ \sum_{i=1}^{m_{t+1,1}} \sum_{l=1}^{d_{t,i,1}} \mathbb{E}\left[\prod_{j=1}^{m_{t+1,1,i}} \operatorname{tr}\left(M_{t,i,j,1,l}\right) M_{t,i,0,1,l}\middle|\mathcal{F}_0\right]\frac{1}{b} + \cdots + \\
&+ \sum_{i=1}^{m_{t+1,d}} \sum_{l=1}^{d_{t,i,d}} \mathbb{E}\left[\prod_{j=1}^{m_{t+1,d,i}} \operatorname{tr}\left(M_{t,i,j,d,l}\right) M_{t,i,0,d,l}\middle|\mathcal{F}_0\right]\frac{1}{b^d},
\end{aligned} \tag{19}$$

and each $M_{t,i,j,k,l}$ is a multiplicative term of parameter matrices $\left\{g^b_{t,1}, g^b_{t,2}\right\}$ and constant matrices $\{W^b_{t,1}, W^b_{t,2}, W^*_1, W^*_2\}$ such that the degree is at most 1. Therefore, by induction, for every i, k, l, we have

$$\mathbb{E}\left[\prod_{j=1}^{m_{t+1,k,i}} \operatorname{tr}\left(M_{t,i,j,k,l}\right) M_{t,i,0,k,l}\middle|\mathcal{F}_0\right] = N_{t,i,k,l,0} + N_{t,i,k,l,1}\frac{1}{b} + \cdots N_{t,i,k,l,q_t}\frac{1}{b^{q_t}}, \tag{20}$$

where $q_t \leqslant d' + \frac{1}{2}(3^t - 1)(3d + d')$ and $N_{t,i,k,l,0}, \cdots, N_{t,i,k,l,q_t}$ are sum of multiplicative terms of parameter matrices $\left\{W^b_{0,1}, W^b_{0,2}\right\}$ and constant matrices $\{W^*_1, W^*_2\}$ with degree at most $d \cdot 3^t$.

Combining (19) and (20), we can rewrite

$$\mathbb{E}\left[M|\mathcal{F}_0\right] = N_0 + N_1\frac{1}{b} + \cdots + N_q\frac{1}{b^q},$$

in the same form as in the statement. Here $q \leqslant d + 3q_t \leqslant \frac{1}{2}(3^{t+2}-1)d + \frac{1}{2}(3^{t+1}-1)d'$ and $\sum_{j=0}^{m_{ki}} \deg\left(M_{ij}^k\right) \leqslant 3 \times 3^t(3d+d') = 3^{t+1}(3d+d')$ follow from (16) and (17).

In conclusion, we have shown that the statement holds for $t+1$, and therefore finishes the proof.

□

By changing the role of parameter and constant matrices in Theorem 3, we obtain the following corollary.

Corollary 2. *Given $t \geqslant 0$, for any multiplicative terms $M_i, i \in [0:m]$ of parameter matrices $\left\{W_{t,1}^b, W_{t,2}^b, \mathcal{W}_t^b\right\}$ and constant matrices $\{W_1^*, W_2^*\}$ such that $\sum_{i=1}^2 \deg\left(W_{t,i}^b; M\right) = d$ and $\deg\left(\mathcal{W}_t^b; M\right) = d'$, we denote $M = \prod_{i=1}^m \mathrm{tr}\left(M_i\right) M_0$. There exists a set of multiplicative terms $\left\{M_{ij}^k, i \in [m_k], j \in [0:m_{ki}], k \in [0:q]\right\}$ of parameter matrices $\left\{W_{0,1}^b, W_{0,2}^b\right\}$ and constant matrices $\{W_1^*, W_2^*\}$ such that*

$$\mathbb{E}\left[M|\mathcal{F}_0\right] = N_0 + N_1\frac{1}{b} + \cdots + N_q\frac{1}{b^q},$$

where $N_k = \sum_{i=1}^{m_k}\prod_{j=1}^{m_{ki}} \mathrm{tr}\left(M_{ij}^k\right) M_{i0}^k, k \in [0:q]$. Here m_k, m_{ki} and $q \leqslant 3^t\left(d+2d'\right)$ are constants independent of b, and $\sum_{j=0}^{m_{ki}} \deg\left(M_{ij}^k\right) \leqslant 3^t\left(d+2d'\right)$.

Proof of Corollary 2. We simply note that M can be written as the sum of at most 2^d multiplicative terms of parameter matrices $\left\{W_{t,1}^b, W_{t,2}^b, W_1^*, W_2^*\right\}$ and constant matrix $\{I_0\}$. Then we apply Lemmas 8 and 9 iteratively in the same way as in the proof of Theorem 3 to finish the proof. □

Proof of Theorem 4. We only show the case for $g_{t,1}$ since the proof for $g_{t,2}$ can be tackled similarly. Note that

$$\begin{aligned}
\mathsf{var}\left(g_{t,1}^b\middle|\mathcal{F}_0\right) &= \mathsf{var}\left(\frac{1}{b}\sum_{i=1}^b {W_{t,2}^b}^T\mathcal{W}_t^b x_{t,i}x_{t,i}^T\middle|\mathcal{F}_0\right) = \frac{1}{b^2}\sum_{i=1}^b \mathsf{var}\left({W_{t,2}^b}^T\mathcal{W}_t^b x_{t,i}x_{t,i}^T\middle|\mathcal{F}_0\right) \\
&= \frac{1}{b}\mathsf{var}\left({W_{t,2}^b}^T\mathcal{W}_t^b x_{t,1}x_{t,1}^T\middle|\mathcal{F}_0\right) \\
&= \frac{1}{b}\left(\mathbb{E}\left[\left\|{W_{t,2}^b}^T\mathcal{W}_t^b x_{t,1}x_{t,1}^T\right\|^2\middle|\mathcal{F}_0\right] - \left\|\mathbb{E}\left[{W_{t,2}^b}^T\mathcal{W}_t^b x_{t,1}x_{t,1}^T\middle|\mathcal{F}_0\right]\right\|^2\right) \\
&= \frac{1}{b}\left(\mathbb{E}\left[\mathrm{tr}\left(x_{t,1}x_{t,1}^T\mathcal{W}_t^{bT}W_{t,2}^b{W_{t,2}^b}^T\mathcal{W}_t^b x_{t,1}x_{t,1}^T\right)\middle|\mathcal{F}_0\right] - \left\|\mathbb{E}\left[{W_{t,2}^b}^T\mathcal{W}_t^b x_{t,1}x_{t,1}^T\middle|\mathcal{F}_0\right]\right\|^2\right) \\
&= \frac{1}{b}\left(\mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\left(x_{t,1}x_{t,1}^T\mathcal{W}_t^{bT}W_{t,2}^b{W_{t,2}^b}^T\mathcal{W}_t^b x_{t,1}x_{t,1}^T\right)\middle|\mathcal{F}_t^b\right]\middle|\mathcal{F}_0\right] - \left\|\mathbb{E}\left[\mathbb{E}\left[{W_{t,2}^b}^T\mathcal{W}_t^b x_{t,1}x_{t,1}^T\middle|\mathcal{F}_t^b\right]\middle|\mathcal{F}_0\right]\right\|^2\right) \\
&= \frac{1}{b}\left((p+2)\mathrm{tr}\left(\mathcal{W}_t^{bT}W_{t,2}^b{W_{t,2}^b}^T\mathcal{W}_t^b\right)\middle|\mathcal{F}_0\right] - \left\|\mathbb{E}\left[{W_{t,2}^b}^T\mathcal{W}_t^b\middle|\mathcal{F}_0\right]\right\|^2\right) \\
&= \frac{1}{b}\left((p+2)\mathrm{tr}\left(\mathbb{E}\left[\mathcal{W}_t^{bT}W_{t,2}^b{W_{t,2}^b}^T\mathcal{W}_t^b\middle|\mathcal{F}_0\right]\right) - \left\|\mathbb{E}\left[{W_{t,2}^b}^T\mathcal{W}_t^b\middle|\mathcal{F}_0\right]\right\|^2\right). \\
&= \frac{1}{b}\left((p+2)\mathrm{tr}\left(\mathbb{E}\left[\mathcal{W}_t^{bT}W_{t,2}^b{W_{t,2}^b}^T\mathcal{W}_t^b\middle|\mathcal{F}_0\right]\right) - \left\|\mathbb{E}\left[{W_{t,2}^b}^T\mathcal{W}_t^b\middle|\mathcal{F}_0\right]\right\|^2\right).
\end{aligned}$$

Here we have used the fact that $\mathbb{E}_{x\sim\mathcal{N}(0,I_p)}\mathrm{tr}\left(xx^TAxx^T\right) = (p+2)\mathrm{tr}\left(A\right)$. By Corollary 2 we know that there exists a set of multiplicative terms $\left\{M_{ij}^k, i \in [m_k], j \in [0:m_{ki}], k \in [0:q]\right\}$ of parameter matrices $\left\{W_{0,1}^b, W_{0,2}^b\right\}$ and constant matrices $\{W_1^*, W_2^*\}$ such that

$$\mathrm{tr}\left(\mathbb{E}\left[\mathcal{W}_t^{bT}W_{t,2}^b{W_{t,2}^b}^T\mathcal{W}_t^b\middle|\mathcal{F}_0\right]\right) = \gamma_0 + \gamma_1\frac{1}{b} + \cdots + \gamma_q\frac{1}{b^q}, \tag{21}$$

where $\gamma_k = \sum_{i=1}^{m_k} \prod_{j=0}^{m_{ki}} \mathrm{tr}\left(M_{ij}^k\right), k \in [0:q]$. Here m_k, m_{ki} and $q \leqslant 6 \cdot 3^t$ are constants independent of b, and $\sum_{j=0}^{m_{ki}} \deg\left(M_{ij}^k\right) \leqslant 6 \cdot 3^t$. Note that $W_{0,1}^b, W_{0,2}^b$ are fixed, and we have $\gamma_k, k \in [0:q]$ are constants independent of b.

Similarly we observe that there exist constants $q' \leqslant 2 \cdot 3^{t+1}$ and $\gamma_k', k \in [0:q']$ such that

$$\left\| \mathbb{E}\left[{W_{t,2}^b}^T \mathcal{W}_t^b \middle| \mathcal{F}_0 \right] \right\|^2 = \gamma_0' + \gamma_1' \frac{1}{b} + \cdots + \gamma_q' \frac{1}{b^{q'}}. \tag{22}$$

By defining $\gamma_i = 0, i > q$ and $\gamma_i' = 0, i > q'$, and combining (21) and (22) we have

$$\begin{aligned}
\mathsf{var}\left(g_{t,1}^b \middle| \mathcal{F}_0\right) &= \frac{1}{b}\left((p+2)\mathrm{tr}\left(\mathbb{E}\left[\mathcal{W}_t^{bT} W_{t,2}^b {W_{t,2}^b}^T \mathcal{W}_t^b \middle| \mathcal{F}_0 \right] \right) - \left\| \mathbb{E}\left[{W_{t,2}^b}^T \mathcal{W}_t^b \middle| \mathcal{F}_0 \right] \right\|^2 \right) \\
&= \frac{p+2}{b}\left(\gamma_0 + \gamma_1 \frac{1}{b} + \cdots + \gamma_q \frac{1}{b^q} \right) - \frac{1}{b}\left(\gamma_0' + \gamma_1' \frac{1}{b} + \cdots + \gamma_q' \frac{1}{b^{q'}} \right) \\
&= \sum_{k=1}^{\max\{q,q'\}} \left((p+1)\gamma_k - \gamma_k' \right) \frac{1}{b^k}.
\end{aligned}$$

Note that γ_k's and γ_k''s are all constants independent of b, and $\max\{q, q'\} \leqslant 2 \cdot 3^{t+1}$. This completes the proof.

□

Proof of Theorem 5. We first show that in

$$\mathsf{var}\left(g_{t,i}^b \middle| \mathcal{F}_0\right) = \beta_1 \frac{1}{b} + \cdots + \beta_r \frac{1}{b^r}$$

we have $\beta_1 \geqslant 0$. If $r = 1$, the statement obviously holds. Let us assume that the statement does not hold for $r > 1$, i.e. $\beta_1 < 0$. Taking b large enough such that $\beta_1 b^{r-1} + \beta_2 b^{r-2} + \cdots + \beta_r < 0$ yields

$$\mathsf{var}\left(g_{t,i}^b \middle| \mathcal{F}_0\right) = \frac{1}{b^r}\left(\beta_1 b^{r-1} + \beta_2 b^{r-2} + \cdots + \beta_r \right) < 0,$$

which contradicts the fact that $\mathsf{var}\left(g_{t,i}^b \middle| \mathcal{F}_0\right) \geqslant 0$. Therefore, we have $\beta_1 \geqslant 0$.

Let b_0 be large enough such that for all $b \geqslant b_0$, we have $\beta_1 b^{r-1} + 2\beta_2 b^{r-2} + \cdots + r\beta_r \geqslant 0$. We denote $f(b) = \beta_1 \frac{1}{b} + \beta_2 \frac{1}{b^2} + \cdots + \beta_r \frac{1}{b^r} \geqslant 0$. For all $b > b_0$ we have

$$f'(b) = -\frac{1}{b^{r+1}}\left(\beta_1 b^{r-1} + 2\beta_2 b^{r-2} + \cdots + r\beta_r \right) \leqslant 0.$$

Therefore, for all $b > b_0$ we have $\left(\mathsf{var}\left(g_{t,i}^b \middle| \mathcal{F}_0\right)\right)' = -\frac{r}{b^{r+1}} f(b) + \frac{1}{b^r} f(b) \leqslant 0$, and thus $\mathsf{var}\left(g_{t,i}^b \middle| \mathcal{F}_0\right)$ is a decreasing function of b for all $b > b_0$.

□

B.3 Extension to Deep Linear Networks

The extension from two-layer linear network to deep linear network is straightforward. Here we only provide the ideas on how to translate the proof of two-layer network to d-layer network, but not the strict proof. For simplicity, we remove all superscripts b of matrices in this subsection.

Assume that the d-layer linear network is given by $f(x; w) = W_d W_{d-1} \cdots W_2 W_1 x$, where $W_i, i \in [d]$ is the parameter matrix on the i-th layer and $w = (W_1, \ldots, W_d)$. The population loss is defined as

$$\mathcal{L}(w) = \mathbb{E}_{x \sim \mathcal{N}(0, I_p)}\left[\frac{1}{2} \left\| W_d \cdots W_1 x - W_d^* \cdots W_1^* x \right\|^2 \right].$$

Similar to (1) and (2), we have

$$\begin{aligned}
g_{t,k} &= \frac{1}{b} \sum_{i=1}^{b} \nabla_{W_{t,k}} \left(\frac{1}{2} \left\| W_{t,d} \cdots W_{t,1} x_{t,i} - W_d^* \cdots W_1^* x_{t,i} \right\|^2 \right) \\
&= \frac{1}{b} \sum_{i=1}^{b} W_{t,k+1}^T \cdots W_{t,d}^T \left(W_d \cdots W_1 - W_d^* \cdots W_1^* \right) x_{t,i} x_{t,i}^T W_{t,1}^T \cdots W_{t,k-1}^T, \quad k \in [d].
\end{aligned}$$

We denote $\mathcal{W}_t = W_{t,d} \cdots W_{t,1} - W_d^* \cdots W_1^*$. The remaining are all the same as the proofs in Appendix B.2, except we should replace all appearance of $\{W_{t,2}, W_{t,1}\}$ to $\{W_{t,d}, W_{t,d-1}, \cdots, W_{t,1}\}$ and all $\{W_2^*, W_1^*\}$ to $\{W_d^*, W_{d-1}^*, \cdots, W_1^*\}$. We can do this because the stochastic gradient $g_{t,k}$ is still the sum of multiplicative terms of parameter matrices $\{x_{t,i}\}$ and constant matrices $\{W_{t,d}, \cdots, W_{t,1}, W_d^*, \cdots, W_1^*\}$ so the Lemmas in Appendix B.2 still apply.

In conclusion, we can again represent $\mathsf{var}\,(g_{t,k}|\mathcal{F}_0), k \in [d]$ as a polynomial of $\frac{1}{b}$ with finite degree and without the constant term. By the same approach in the proof of Theorem 5, we can show that the variance is a decreasing function of the mini-batch size b.