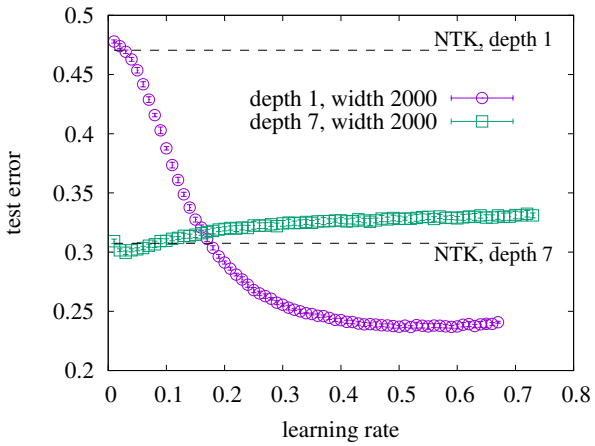


Figure 7: Learning rate dependence of the generalization performance. Nonlinear feedforward neural networks of different depths are trained on a simple task with varying learning rates. We see that, when the learning rate is vanishingly small so that the continuous-time approximation is good, the continuous neural tangent kernel (NTK) provides an accurate characterization of the result of training. However, as the learning rate becomes large, the learning deviates significantly and qualitatively from the NTK prediction, sometimes for the better, sometimes for the worse. Reproduced from Mori and Ueda (2020b). For other interesting experiments concerning large learning rate, see Lewkowycz et al. (2020).

A. Example of Failure of Continuous-Time Theory

See Figure 7 for an example on the generalization performance with different learning rates. For small learning rates, the continuous-time neural tangent kernel (NTK) theory successfully predicts the correct behavior. For a slightly larger λ, the prediction given by continuous theory deviates significantly from the experiments.

B. Effect of Overparametrization

One particular topic that is of interest in the recent deep learning literature is the role of overparametrization (Neyshabur et al., 2018b). Modern neural networks, defying the traditional way of thinking in statistical learning, often perform better when the number of parameters is larger than the number of data points. We comment that our formalism can also be extended straightforwardly to deal with this. In the overparametrized regime, many directions in the loss landscape are degenerate, and have zero curvature; this means that the Hessian matrix in a local minimum is positive semi-definite with many zero eigenvalues. In this situation, the difference between artificially added noise that is usually full-rank and a low-rank noise that is, e.g., proportional to the Hessian becomes important: on the one hand, when the Hessian is low rank, a full-rank noise causes an unconstrained Brownian motion in the null space, the model will thus diverge and one cannot expect to obtain good generalization here; on the other hand, a noise that is proportional to the Hessian only diffuses in the subspace spanned by the Hessian and will not diverge; this is exactly the result obtained in Hodgkinson and Mahoney (2020) using the formalism of iterated random functions. This implies that the generalization performance induced by minibatch sampling is better than that of an artificially injected Gaussian noise, which has been observed frequently in experiments (Hoffer et al., 2017; Zhu et al., 2019).

C. Additional Experiments for Non-Gaussian Noise

See Figure 8. We show that, for example, the theory agrees with the cases when the noise obeys the Student's t-distribution (heavy tail) and the χ^2 distribution (asymmetric); the setting is the same as in the 1d experiments in section 5. Also, this result remains valid even if $C = C(\Sigma)$ is dependent on the Σ itself.

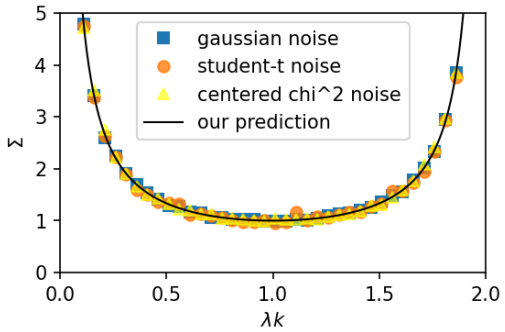
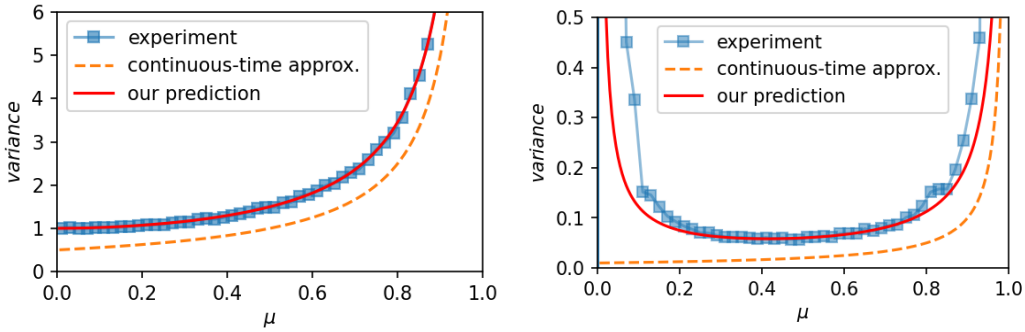


Figure 8: Comparison of the theoretical prediction with different kinds of noise. We choose the Student's t-distribution with $\nu = 4$ as an example of heavy-tail noise with a tail exponent 5, and a centered χ^2 distribution (by subtracting the mean from a standard χ^2 distribution with degree of freedom 3). The agreement is excellent, independent of the underlying noise distribution.

D. Additional Experiments for SGD with Momentum



(a) White noise (b) Minibatch noise

Figure 9: Comparison between the continuous-time theory and the discrete-time theory in the presence of momentum. (a) White noise with $\lambda k = 1$. In this case, the fluctuation does not diverge when $\mu < 1$. However, the error of the continuous-time approximation does not diminish even if μ gets large. (b) Minibatch noise with $\lambda k = 2$. Even in the presence of minibatch noise, the proposed theory agrees much better with the experiments.

In Figure 9(a), we plot the model fluctuation with white noise with $\lambda k = 1$; this is the case in which there is no divergence for $\mu < 1$. Here, we see that the continuous-time theory predicts an error that does not diminish even if μ is close to 1. In Figure 9(b), we show the experiments with minibatch noise for the same linear regression task adopted in section 5. The predicted discrete-time result agrees better than the continuous-time one. On the other hand, the agreement becomes worse as the fluctuation in w becomes large. This again suggests the limitation of the commonly used approximation of minibatch noise, i.e., $C \sim H(w) = K$.

E. Proofs and Additional Theoretical Considerations

E.1. Proofs in Section 4

E.1.1. Proof of Theorems 1, 2 and 3

Because Theorems 1 and 2 can be derived from Theorem 3 by assuming a scalar λ and $\mu = 0$ accordingly, we first prove Theorem 3.

Proof. We assume that the stationary distributions of both $\mathbf{m}$ and $\mathbf{w}$ exist and $\lim_{t\to\infty}\mathbb{E}[\mathbf{w}_t\mathbf{w}_t^{\mathrm{T}}] := \Sigma$. The goal is to find Σ. We assume that $\mathbf{w}_0$ and $\mathbf{m}_0$ are sampled from the stationary distribution. This is valid as long as we are interesting in the asymptotic behavior of $\mathbf{w}_t$. By definition,

$$\begin{aligned}\Sigma := \mathbb{E}[\mathbf{w}_t\mathbf{w}_t^{\mathrm{T}}] &= \mathbb{E}[(\mathbf{w}_{t-1} - \mu\Lambda\mathbf{m}_{t-1} - \Lambda K\mathbf{w}_{t-1} - \Lambda\eta_{t-1})(\mathbf{w}_{t-1} - \mu\Lambda\mathbf{m}_{t-1} - \Lambda K\mathbf{w}_{t-1} - \Lambda\eta_{t-1})^{\mathrm{T}}]\\ &= \mathbb{E}[(I_D - \Lambda K)\mathbf{w}_{t-1}\mathbf{w}_{t-1}^{\mathrm{T}}(I_D - K\Lambda)] + \mu^2\Lambda M\Lambda + \Lambda C\Lambda - (A + A^{\mathrm{T}}),\end{aligned} \tag{28}$$

where $A := \mu(I_D - \Lambda K)\mathbb{E}\left[\mathbf{w}_{t-1}\mathbf{m}_{t-1}^{\mathrm{T}}\right]\Lambda$ and $M := \mathbb{E}\left[\mathbf{m}_{t-1}\mathbf{m}_{t-1}^{\mathrm{T}}\right]$. Notice that $\mathbf{w}_0$ is initialized according to the stationary distribution. Therefore, the distribution does not depend on t, namely $\mathbb{E}[\mathbf{w}_t\mathbf{w}_t^{\mathrm{T}}] = \mathbb{E}[\mathbf{w}_{t-1}\mathbf{w}_{t-1}^{\mathrm{T}}] = \Sigma$. For the covariance matrix of the momentum, similarly,

$$\begin{aligned}\Lambda M\Lambda &= \mathbb{E}[(\mathbf{w}_{t-1} - \mathbf{w}_{t-2})(\mathbf{w}_{t-1} - \mathbf{w}_{t-2})^{\mathrm{T}}]\\ &= 2\Sigma - \mathbb{E}[\mathbf{w}_{t-1}\mathbf{w}_{t-2}^{\mathrm{T}}] - \mathbb{E}[\mathbf{w}_{t-2}\mathbf{w}_{t-1}^{\mathrm{T}}].\end{aligned} \tag{29}$$

For the final two terms $A + A^{\mathrm{T}}$, we have

$$\begin{aligned}A &= \mu(I_D - \Lambda K)\mathbb{E}\left[\mathbf{w}_{t-1}\mathbf{m}_{t-1}^{\mathrm{T}}\right]\Lambda\\ &= \mu(I_D - \Lambda K)\mathbb{E}[\mathbf{w}_{t-1}(\mathbf{w}_{t-2} - \mathbf{w}_{t-1})^{\mathrm{T}}]\\ &= -\mu(I_D - \Lambda K)\Sigma + \mu(I_D - \Lambda K)\mathbb{E}[\mathbf{w}_{t-1}\mathbf{w}_{t-2}^{\mathrm{T}}],\end{aligned} \tag{30}$$

$$A^{\mathrm{T}} = -\mu\Sigma(I_D - K\Lambda) + \mu\mathbb{E}[\mathbf{w}_{t-2}\mathbf{w}_{t-1}^{\mathrm{T}}](I_D - K\Lambda). \tag{31}$$

Therefore, we are left to solve for $\mathbb{E}[\mathbf{w}_{t-1}\mathbf{w}_{t-2}^{\mathrm{T}}]$ and its transpose. Using the fact that the expectation values are time-independent for the stationary state, we obtain

$$\begin{aligned}\mathbb{E}[\mathbf{w}_{t-1}\mathbf{w}_{t-2}^{\mathrm{T}}] = \mathbb{E}[\mathbf{w}_t\mathbf{w}_{t-1}^{\mathrm{T}}] &= \mathbb{E}[(\mathbf{w}_{t-1} - \mu\Lambda\mathbf{m}_{t-1} - \Lambda K\mathbf{w}_{t-1} - \Lambda\eta_{t-1})\mathbf{w}_{t-1}^{\mathrm{T}}]\\ &= (I_D - \Lambda K)\Sigma - \mu\Lambda\mathbb{E}[\mathbf{m}_{t-1}\mathbf{w}_{t-1}^{\mathrm{T}}]\\ &= (I_D - \Lambda K)\Sigma + \mu\Sigma - \mu\mathbb{E}[\mathbf{w}_{t-2}\mathbf{w}_{t-1}^{\mathrm{T}}],\end{aligned} \tag{32}$$

$$\mathbb{E}[\mathbf{w}_{t-2}\mathbf{w}_{t-1}^{\mathrm{T}}] = \Sigma(I_D - K\Lambda) + \mu\Sigma - \mu\mathbb{E}[\mathbf{w}_{t-1}\mathbf{w}_{t-2}^{\mathrm{T}}]. \tag{33}$$

From the above two equations, we have

$$\mathbb{E}[\mathbf{w}_{t-1}\mathbf{w}_{t-2}^{\mathrm{T}}] = \frac{1}{1-\mu^2}\left[(I_D - \Lambda K)\Sigma + \mu\Sigma - \mu\Sigma(I_D - K\Lambda) - \mu^2\Sigma\right], \tag{34}$$

$$\mathbb{E}[\mathbf{w}_{t-2}\mathbf{w}_{t-1}^{\mathrm{T}}] = \frac{1}{1-\mu^2}\left[\Sigma(I_D - K\Lambda) + \mu\Sigma - \mu(I_D - \Lambda K)\Sigma - \mu^2\Sigma\right]. \tag{35}$$

Finally, substituting these results back into (28) yields

$$(1-\mu)(\Lambda K\Sigma + \Sigma K\Lambda) - \frac{1+\mu^2}{1-\mu^2}\Lambda K\Sigma K\Lambda + \frac{\mu}{1-\mu^2}(\Lambda K\Lambda K\Sigma + \Sigma K\Lambda K\Lambda) = \Lambda C\Lambda. \tag{36}$$

□

While Theorems 1 and 2 can be proven via the similar method as above, it is easier to derive them from Theorem 3. For Theorem 2, we assume a scalar learning rate λ.

Proof. Let $\Lambda = \lambda I_D$. Then from Eq. (9), we have

$$(1-\mu)\lambda(K\Sigma+\Sigma K) - \frac{1+\mu^2}{1-\mu^2}\lambda^2 K\Sigma K + \frac{\mu}{1-\mu^2}\lambda^2(K^2\Sigma+\Sigma K^2) = \lambda^2 C. \tag{37}$$

□

Theorem 1 can be derived from Theorem 2 by setting $\mu = 0$.

Proof. Let λ be a scalar and $\mu = 0$. Then from Eq. (7), we have

$$\Sigma K + K\Sigma - \lambda K\Sigma K = \lambda C. \tag{38}$$

□

E.1.2. Proof of Corollary 1

We first prove a lemma about commutation relations.

Lemma 1. $[\Sigma, K] = 0$, *if and only if* $[C, K] = 0$.

Proof. 1. We first prove that if $[\Sigma, K] = 0$, then $[C, K] = 0$, which is straightforward. Equation. (7) implies that C is a analytical function of Σ and K, i.e., $C = C(K, \Sigma)$. The commutator is

$$[C, K] = [C(K, \Sigma), K]. \tag{39}$$

If $[\Sigma, K] = 0$, it directly follows that $[C, K] = 0$.

2. Now we prove the if $[C, K] = 0$, then $[\Sigma, K] = 0$, which is not so straightforward. We introduce simplified notations: $X := (1-\mu)I_D$ and $Y := I_D - \lambda K$. By iteration, we have

$$\begin{aligned}\mathbf{w}_t &= (X+Y)\mathbf{w}_{t-1} - X\mathbf{w}_{t-2} + \lambda\eta_{t-1} \\ &\cdots \\ &= g_{t-1}\mathbf{w}_1 - X g_{t-2}\mathbf{w}_0 + \lambda\sum_{i=0}^{t-1} g_i\eta_{t-1-i},\end{aligned} \tag{40}$$

where the coefficient matrices g_i satisfy the following recurrence relation

$$g_t = (X+Y)g_{t-1} - X g_{t-2}, \quad \text{for} \quad t \geq 2, \tag{41}$$

where the initial terms are given by

$$g_0 = I_D, \quad g_1 = X + Y. \tag{42}$$

It follows from the relation $\lim_{t\to\infty} g_t = 0$ that

$$\lim_{t\to\infty}\mathbf{w}_t = \lim_{t\to\infty}\lambda\sum_{i=0}^{t-1} g_i\eta_{t-1-i} \sim \mathcal{N}\left(0, \lambda^2 \lim_{t\to\infty}\sum_{i=0}^{t-1} g_i C g_i\right) := \mathcal{N}(0, \Sigma). \tag{43}$$

Because every g_i is a function of K, $[C, K] = 0$ is equivalent with $[\Sigma, K] = 0$.

□

With this lemma, we prove Corollary 1.

Proof. The matrix equation (7) satisfied by the parameter covariance matrix can be equivalently written in the form containing commutators as

$$\underbrace{(1-\mu)\lambda K\left(2I_D-\frac{\lambda}{1+\mu}K\right)\Sigma}_{\text{commuting contribution}}+\underbrace{(1-\mu)\lambda\left(I_D-\frac{\lambda}{1+\mu}K\right)[\Sigma,K]+\frac{\mu}{1-\mu^2}\lambda^2[K,[K,\Sigma]]}_{\text{non-commuting contribution}}=\lambda^2 C, \tag{44}$$

where the non-commuting contribution is finite if $[C,K]\neq 0$. Otherwise, if $[C,K]=0$, we have $[\Sigma,K]=0$ such that the non-commuting term vanishes and the model fluctuation is

$$\begin{aligned}\Sigma&=\left[\frac{\lambda K}{1+\mu}\left(2I_D-\frac{\lambda K}{1+\mu}\right)\right]^{-1}\frac{\lambda^2 C}{1-\mu^2}\\&:=[\tilde{\lambda}K(2I_D-\tilde{\lambda}K)]^{-1}\tilde{C},\end{aligned} \tag{45}$$

where we introduce the following rescaling:

$$\tilde{\lambda}:=\frac{\lambda}{1+\mu},\quad \tilde{C}:=\frac{1+\mu}{1-\mu}C. \tag{46}$$

□

Remark. *We notice that together with $[C,K]=0$, the form of the matrix equation satisfied by Σ is invariant under this rescaling:*

$$\tilde{\lambda}(K\Sigma+\Sigma K)-\tilde{\lambda}^2 K\Sigma K=\tilde{\lambda}^2\tilde{C}. \tag{47}$$

This suggests that the learning rate can be $1+\mu$ times larger.

E.1.3. Proof of Theorem 4 and Corollaries 2 and 3

We first prove Theorem 4.

Proof. According to Theorem 1, if the algorithm is updated under Gaussian noise with covariance matrix C, the stationary distribution of the model parameters $\mathbf{w}$ is $\mathcal{N}(0,\Sigma)$, where Σ satisfies Eq. (6). Due to Lemma 1, we have $[\Sigma,K]=0$ because $C=\sigma^2 I_D+aK$ commutes with K. Referring to Corollary 1, the model fluctuation is

$$\Sigma=\lambda(\sigma^2 I_D+aK)[K(2I_D-\lambda K)]^{-1}. \tag{48}$$

□

Corollaries 2 and 3 can be easily proven from Theorem 4.

Proof. If $\sigma^2=0$, then $\Sigma=a\lambda(2I_D-\lambda K)^{-1}$. If $a=0$, then $\Sigma=\sigma^2\lambda[K(2I_D-\lambda K)]^{-1}$. □

E.1.4. Proof of Theorem 5

Proof. In the presence of momentum, we multiply both sides of Eq. (44) by $R:=\left(2I_D-\frac{\lambda}{1+\mu}K\right)^{-1}$ to the left to obtain

$$(1-\mu)\lambda K\Sigma+RA_1+RA_2=\lambda^2 RC, \tag{49}$$

where $A_1:=(1-\mu)\lambda\left(I_D-\frac{\lambda}{1+\mu}K\right)[\Sigma,K]$ and $A_2:=\frac{\mu}{1-\mu^2}\lambda^2[K,[K,\Sigma]]$ are terms involving commutators. Taking the trace on both sides yields

$$(1-\mu)\lambda\mathrm{Tr}[K\Sigma]+\mathrm{Tr}[RA_1]+\mathrm{Tr}[RA_2]=\lambda^2\mathrm{Tr}\left[\left(2I_D-\frac{\lambda}{1+\mu}K\right)^{-1}C\right]. \tag{50}$$

Because the commuting terms A_1 and A_2 are anti-symmetric by definition and R is symmetric, the traces $\mathrm{Tr}[RA_1]$ and $\mathrm{Tr}[RA_2]$ vanish. Finally, we have

$$L_{\mathrm{train}} \coloneqq \frac{1}{2}\mathrm{Tr}[K\Sigma] = \frac{\lambda}{4(1-\mu)}\mathrm{Tr}\left[\left(I_D - \frac{\lambda}{2(1+\mu)}K\right)^{-1} C\right]. \tag{51}$$

□

E.2. Proofs in Section 6

E.2.1. Proof of Theorem 6

Proof. The goal of this approximate Bayesian inference task is to find the optimal learning rate which minimizes the KL divergence between the SGD stationary distribution $q(\mathbf{w})$ given in Theorem 1 and the posterior (15). The KL divergence is

$$\begin{aligned} D_{\mathrm{KL}}(q\|f) &= -\mathbb{E}_q(\ln f) + \mathbb{E}_q(\ln q) \\ &= \frac{1}{2}\left[N\mathrm{Tr}[K\Sigma] - \ln|NK| - \ln|\Sigma| - D\right], \end{aligned}$$

where $|\cdot|$ is the determinant and D is the dimension of the parameters $\mathbf{w}$.

Suppose that the noise covariance is $C = \frac{N-S}{NS}K$, which is an approximation of the noise induced by minibatch sampling (Hoffer et al., 2017). According to Theorem 4, the covariance of the model is

$$\Sigma = \lambda\frac{N-S}{NS}(2I_D - \lambda K)^{-1}. \tag{52}$$

Therefore, up to constant terms, the KL divergence is

$$D_{\mathrm{KL}} \overset{c}{=} \lambda\frac{N-S}{S}\mathrm{Tr}[(2I_D - \lambda K)^{-1}K] - D\ln\lambda + \ln|2I_D - \lambda K| - D. \tag{53}$$

Taking the derivative with respect to λ yields

$$\frac{\partial}{\partial\lambda}D_{\mathrm{KL}} = \frac{N-2S}{S}\mathrm{Tr}[(2I_D - \lambda K)^{-1}K] + \lambda\frac{N-S}{S}\mathrm{Tr}[(2I_D - \lambda K)^{-2}K^2] - \frac{D}{\lambda}. \tag{54}$$

The optimal λ is obtained by solving $\frac{\partial}{\partial\lambda}D_{\mathrm{KL}} = 0$, namely

$$\frac{N-2S}{S}\mathrm{Tr}[(2I_D - \lambda K)^{-1}K] + \lambda\frac{N-S}{S}\mathrm{Tr}[(2I_D - \lambda K)^{-2}K^2] = \frac{D}{\lambda}. \tag{55}$$

Equivalently, it can be written into Eq. (16), because K and $(2I_D - \lambda K)^{-1}$ are simultaneously diagonalizable. □

E.2.2. Proof of Theorem 7

We first derive the discrete-time version of the escaping efficiency (18) presented in Theorem 7.

Proof. Because the initial state is the exact minimum, namely $\mathbf{w}_0 = 0$, the parameters evolve at time t to

$$\mathbf{w}_t = \lambda\sum_{i=0}^{t-1}(I_D - \lambda K)^i\eta_{t-1-i}. \tag{56}$$

The loss for such parameters is

$$L(\mathbf{w}_t) = \frac{\lambda^2}{2}\sum_{i=0}^{t-1}\eta_{t-1-i}^{\mathrm{T}}K(I_D - \lambda K)^{2i}\eta_{t-1-i} + \mathrm{cross-terms}, \tag{57}$$

where the cross-terms involve not-equal-time contributions. The expectation value of the loss at time t is

$$\begin{aligned} E := \mathbb{E}[L(\mathbf{w}_t)] &= \frac{\lambda^2}{2}\sum_{i=0}^{t-1}\mathrm{Tr}\left[CK(I_D - \lambda K)^{2i}\right] \\ &= \frac{\lambda}{4}\mathrm{Tr}\left[\left(I_D - \frac{\lambda K}{2}\right)^{-1}\left[I_D - (I_D - \lambda K)^{2t}\right]C\right], \end{aligned} \tag{58}$$

where the cross-terms vanish due to the Gaussian property of the noise and in the second line we use the Neumann series that $\sum_{i=0}^{n}A^i = (I_D - A)^{-1}(I_D - A^{n+1})$.

□

E.2.3. Proof of Corollary 4

Proof. As a necessary condition, if each component inside the trace of E_d is greater than that of E_c, then the trace itself should be so as well. Specifically, we wish to show that

$$\left(1-\frac{\lambda k}{2}\right)^{-1}\left[1-(1-\lambda k)^{2t}\right] \geq 1-e^{-2\lambda k t}, \quad \forall\ 0<\lambda k<2 \text{ and } t\geq 0. \tag{59}$$

Equivalently, we wish to show that

$$\left(1-\frac{\lambda k}{2}\right)e^{-2\lambda k t} \geq (1-\lambda k)^{2t}-\frac{\lambda k}{2}. \tag{60}$$

Because $e^{-x}\geq 1-x$ for all $x\geq 0$, we have

$$\begin{aligned}\text{lhs} &:= \left(1-\frac{\lambda k}{2}\right)e^{-2\lambda k t}\\ &\geq \left(1-\frac{\lambda k}{2}\right)(1-\lambda k)^{2t} = (1-\lambda k)^{2t}-\frac{\lambda k}{2}(1-\lambda k)^{2t}\\ &\geq (1-\lambda k)^{2t}-\frac{\lambda k}{2} := \text{rhs}.\end{aligned} \tag{61}$$

□

E.2.4. Proof of Theorem 8

Proof. We first elaborate on the condition about the alignment assumption. As in Zhu et al. (2019), we denote the maximal eigenvalue and the corresponding eigenvector of C as c_1 and v_1, respectively. We have $u_1^{\mathrm{T}}Cu_1 \geq u_1^{\mathrm{T}}v_1c_1v_1^{\mathrm{T}}u_1 = c_1\langle u_1,v_1\rangle^2$. If the maximal eigenvalues of C and K are aligned in proportion, namely $c_1/\mathrm{Tr}[C]\geq a_1k_1/\mathrm{Tr}[K]$, and the angle between their eigenvectors is so small that $\langle u_1,v_1\rangle^2\geq a_2$, then we can conclude that $u_1^{\mathrm{T}}Cu_1\geq ak_1\frac{\mathrm{Tr}[C]}{\mathrm{Tr}[K]}$ with $a:=a_1a_2$.

We then derive the efficiency ratio (20). For a single step, it is the same as the continuous-time one (Zhu et al., 2019). Decomposing $\mathrm{Tr}[KC]$, we have

$$\mathrm{Tr}[KC]=\sum_{i=1}^{D}k_iu_i^{\mathrm{T}}Cu_i\geq k_1u_1^{\mathrm{T}}Cu_1\geq ak_1^2\frac{\mathrm{Tr}[C]}{\mathrm{Tr}[K]}. \tag{62}$$

For the isotropic equivalence of the noise, we have

$$\mathrm{Tr}[K\bar{C}]=\frac{\mathrm{Tr}[C]}{D}\mathrm{Tr}[K]. \tag{63}$$

Therefore, we obtain

$$\frac{\mathrm{Tr}[KC]}{\mathrm{Tr}[K\bar{C}]}\geq aD\frac{k_1^2}{(\mathrm{Tr}[K])^2}\geq aD\frac{k_1^2}{\left[lk_1+(D-l)D^{-d}k_1\right]^2}\approx aD\frac{1}{\left[l+(D-l)D^{-d}\right]^2}=\mathcal{O}(aD^{2d-1}). \tag{64}$$

Next, for a long-time, the alignment argument should be slightly modified. While the order of eigenvalues of K is the same as that of $(2I_D-\lambda K)^{-1}$ and they share the same set of eigenvectors, the only thing that should be modified in the argument is that the maximal eigenvalues of C and $(2I_D-\lambda K)^{-1}$ are aligned in proportion such that

$$\frac{c_1}{\mathrm{Tr}[C]}\geq a_3\frac{(2-\lambda k_1)^{-1}}{\mathrm{Tr}\left[(2I_D-\lambda K)^{-1}\right]}, \tag{65}$$

where a_3 is different from a_1 in general. Then the final ratio should contain $a':=a_3a_2$, instead of $a=a_1a_2$. The remaining derivation is the same as above. □

E.2.5. Proof of Theorem 9

Proof. First we propose a new approximation on $P(w \in V_a)$. The width of the well a is approximated by $2\sqrt{\frac{\Delta L}{k}}$, where $\Delta L := L(b) - L(a)$ is the height of the potential barrier and b is the position of the barrier top as shown in Figure 6(a). The probability inside well a is approximated by a finite-range Gaussian integral as

$$\begin{aligned} P(w \in V_a) &\approx \int_{-\sqrt{\frac{\Delta L}{k}}}^{\sqrt{\frac{\Delta L}{k}}} P(w) dw \\ &= P(a) \sqrt{\frac{2\pi C}{\lambda k(2-\lambda k)}} \operatorname{erf}\left(\sqrt{\frac{\lambda(2-\lambda k)\Delta L}{C}} \right), \end{aligned} \tag{66}$$

where $\operatorname{erf}(z)$ is the error function. This probability is strictly smaller than 1, which is consistent with our expectations.

The probability current J can be rewritten as

$$\nabla \left[\exp\left(\frac{L(w) - L(l)}{T} \right) P_c(w) \right] = -J\mathcal{D}^{-1} \exp\left(\frac{L(w) - L(l)}{T} \right), \tag{67}$$

where l is a midpoint on the most probable escape path between a and b such that $k(\mathbf{w}) \approx k_a$ in the path $a \to l$ and $k(\mathbf{w}) \approx k_b$ in $l \to b$. In a stationary state, the probability current J is a constant and it can be obtained by integrating both sides of the above equation from a to b:

$$\text{lhs} = -\exp\left(\frac{L(a) - L(l)}{T_a} \right) P_c(a), \tag{68}$$

and

$$\begin{aligned} \text{rhs} &= -J \int_a^b \mathcal{D}^{-1} \exp\left(\frac{L(w) - L(l)}{T} \right) dw \\ &\approx -J\mathcal{D}_b^{-1} \int_{-\infty}^{\infty} \exp\left(\frac{L(b) - L(l) + \frac{1}{2}(w-b)^{\mathrm{T}} k_b (w-b)}{T_b} \right) dw \\ &= -J\mathcal{D}_b^{-1} \exp\left(\frac{L(b) - L(l)}{T_b} \right) \sqrt{\frac{2\pi T_b}{|k_b|}}, \end{aligned} \tag{69}$$

where we have approximated the integrand on the right-hand side (rhs) because it is peaked around the point b and $\mathcal{D}_b = T_b$. When the noise covariance is $C = \frac{1}{S} k_a$, the two "temperatures" are given by $T_a = \frac{\lambda}{2S} k_a$ and $T_b = \frac{\lambda}{2S} |k_b|$.

We propose two corrections to the approximation of the current: (1) we replace the continuous-time distribution $P_c(w)$ by the discrete-time one $P(w) = P(a) \exp\left(-\frac{1}{2} w^{\mathrm{T}} \Sigma^{-1} w\right)$; (2) the effective "temperature" at point a is enlarged because the fluctuation is larger. From the distribution, the "temperature" should be $T_a = \frac{\lambda}{2S} \frac{k_a}{1 - \lambda k_a/2}$. Specifically, the current is now approximated as

$$J \approx P(a) \exp\left(-\frac{1}{2} w^{\mathrm{T}} \Sigma^{-1} w \right) \exp\left(\frac{L(a) - L(l)}{T_a} - \frac{L(b) - L(l)}{T_b} \right) \sqrt{\frac{|k_b|}{2\pi T_b}}. \tag{70}$$

Substituting everything into the definition (22) yields the approximated Kramers rate:

$$\gamma \approx \frac{1}{2\pi} |k_b| \sqrt{\frac{2}{2 - \lambda k_a}} \operatorname{erf}\left(\sqrt{\frac{S(2-\lambda k_a)\Delta L}{\lambda k_a}} \right) \exp\left[-\frac{2S\Delta L}{\lambda} \left(\frac{l(1 - \lambda k_a/2)}{k_a} + \frac{1-l}{|k_b|} \right) \right], \tag{71}$$

□

Remark. *We emphasize that our corrections are not precise because the current is a dynamical quantity. To precisely characterize the Kramers rate, it may be necessary to develop a discrete-time version of the Fokker-Planck equation (21). Hence, our corrections do not guarantee the accuracy of the coefficients in the expressions.*

E.2.6. More on Approximation Error

In this subsection we derive the matrix equations satisfied by the stationary distribution of a class of SGD with a more general form of momentum called Quasi-Hyperbolic Momentum (QHM) (Ma and Yarats, 2018; Gitman et al., 2019). The update rule is given by

$$\begin{cases} \mathbf{g}_t = K\mathbf{w}_{t-1} + \eta_{t-1}; \\ \mathbf{m}_t = \mu\mathbf{m}_{t-1} + (1-\mu)\mathbf{g}_t; \\ \mathbf{w}_t = \mathbf{w}_{t-1} - \lambda\left[(1-\nu)g_t + \nu\mathbf{m}_t\right], \end{cases} \tag{72}$$

where the additional parameter $\nu \in [0,1]$ interpolates between the usual SGD (5) without momentum ($\nu = 0$) and a normalized version of SGD with momentum (5) ($\nu = 1$). The covariance of the model parameters is given in the following theorem.

Theorem 11. (*Model parameters covariance matrix of QHM*) *Let the algorithm be updated according to Eqs. (72). Then the covariance matrix Σ of the model parameters satisfies the following set of matrix equations:*

$$\begin{cases} \mu(1+\mu)(X+X^{\mathrm{T}}) + \lambda\left[1-\mu\nu(2+\mu)\right](XK+KX^{\mathrm{T}}) + a\Sigma + b(K\Sigma+\Sigma K) + cK\Sigma K = d\lambda^2 C; \\ X = \mu\alpha K^2\Sigma + \lambda[-1+\mu(1+\mu-\mu\nu)]K\Sigma K - \lambda\mu(1-\nu)\alpha K(KQ+QK)K + (1-\mu^2)Q; \\ Q - AQA = \Sigma, \end{cases} \tag{73}$$

where

$$\begin{aligned} &a := -2\mu(1+\mu), \quad b := \lambda\mu^2(1-\nu), \quad c := \lambda^2\left[1-\mu^2-2\mu\nu(1-\mu)\right], \quad d := 1+\mu[\mu-2\nu-2\mu\nu(1-\nu)], \\ &Q := \sum_{i=0}^{\infty} A^i\Sigma A^i, \quad A := \mu[I_D - \lambda(1-\nu)K], \quad \alpha := \lambda[1-\nu+\nu(1-\mu)]. \end{aligned} \tag{74}$$

Remark. *By setting $\nu = 0$ or $\nu = 1$, the previous unnormalized result (6) or (7) can be recovered with reparametrization of $\lambda \to \lambda/(1-\mu)$ (Gitman et al., 2019). Therefore, this result is the most general one in this work.*

Proof. The proof of this theorem is essentially similar to that in Appendix E.1.1 for SGD with momentum, but more complicated. By definition, we have

$$\begin{aligned} \mathbb{E}[\mathbf{w}_t\mathbf{w}_t^{\mathrm{T}}] := \Sigma &= \mathbb{E}\left[(I_D-\alpha K)\mathbf{w}_{t-1}\mathbf{w}_{t-1}^{\mathrm{T}}(I_D-\alpha K)\right] + \lambda^2\nu^2\mu^2\mathbb{E}[\mathbf{m}_{t-1}\mathbf{m}_{t-1}^{\mathrm{T}}] + \lambda^2\alpha^2 C \\ &\quad - \lambda\nu\mu\mathbb{E}\left[(I_D-\alpha K)\mathbf{w}_{t-1}\mathbf{m}_{t-1}^{\mathrm{T}} + \mathbf{m}_{t-1}\mathbf{w}_{t-1}^{\mathrm{T}}(I_D-\alpha K)\right] \\ &= (I_D-\alpha K)\Sigma(I_D-\alpha K) + \lambda^2\nu^2\mu^2 M + \lambda^2\alpha^2 C - (G+G^{\mathrm{T}}), \end{aligned} \tag{75}$$

where $G := \lambda\nu\mu(I_D-\alpha K)\mathbb{E}[\mathbf{w}_t\mathbf{m}_t^{\mathrm{T}}]$, $M := \mathbb{E}[\mathbf{m}_t\mathbf{m}_t^{\mathrm{T}}]$ and $\alpha := \lambda[1-\nu+\nu(1-\mu)]$. For momentum, the update rule gives

$$\lambda\nu\mathbf{m}_t = -\mathbf{w}_t + [I_D - \lambda(1-\nu)K]\mathbf{w}_{t-1} - \lambda(1-\nu)\eta_{t-1}. \tag{76}$$

Therefore, we have

$$\lambda^2\nu^2 M = 2\Sigma + \lambda^2(1-\nu)^2 K\Sigma K + \lambda^2(1-\nu)^2 C - \lambda(1-\nu)(\Sigma K + K\Sigma) - (X+X^{\mathrm{T}}) + \lambda(1-\nu)(XK+KX^{\mathrm{T}}), \tag{77}$$

where $X := \mathbb{E}[\mathbf{w}_t\mathbf{w}_{t-1}^{\mathrm{T}}]$. Similarly, this X satisfies

$$\begin{aligned} X &= (I_D-\alpha K)\Sigma - \lambda\nu\mu\mathbb{E}[\mathbf{m}_{t-1}\mathbf{w}_{t-1}^{\mathrm{T}}] \\ &= (I_D-\alpha K)\Sigma + \mu\Sigma - \mu[I_D - \lambda(1-\nu)K]X^{\mathrm{T}}, \end{aligned} \tag{78}$$

$$X^{\mathrm{T}} = \Sigma(I_D-\alpha K) + \mu\Sigma - \mu X[I_D - \lambda(1-\nu)K]. \tag{79}$$

The relations between G and X are

$$G = -\mu(I_D-\alpha K)\Sigma + \mu(I_D-\alpha K)X[I_D-\lambda(1-\nu)K], \tag{80}$$

$$G^{\mathrm{T}} = -\mu\Sigma(I_D-\alpha K) + \mu[I_D-\lambda(1-\nu)K]X^{\mathrm{T}}(I_D-\alpha K). \tag{81}$$

Although no simple expression of X can be obtained, it is possible to provide a set of equations satisfied by Σ. Substituting everything back into Eq. (75) yields a matrix equation involving Σ and X:

$$\mu(1+\mu)(X+X^{\mathrm{T}})+\lambda\left[1-\mu\nu(2+\mu)\right](XK+KX^{\mathrm{T}})+a\Sigma+b(K\Sigma+\Sigma K)+cK\Sigma K=d\lambda^2 C, \tag{82}$$

where $a:=-2\mu(1+\mu)$, $b:=\lambda\mu^2(1-\nu)$, $c:=\lambda^2\left[1-\mu^2-2\mu\nu(1-\mu)\right]$, $d:=1+\mu[\mu-2\nu-2\mu\nu(1-\nu)]$.

Then we try to express X in terms of Σ. Notice that X and X^{T} satisfy a set of equations with the following form:

$$\begin{cases} X+AX^{\mathrm{T}}=B, \\ X^{\mathrm{T}}+XA=B^{\mathrm{T}}, \end{cases} \tag{83}$$

where $A:=\mu[I_D-\lambda(1-\nu)K]$ and $B:=(1+\mu)\Sigma-\alpha K\Sigma$. From them we have

$$X-AXA=B-AB^{\mathrm{T}}:=D. \tag{84}$$

Therefore, by iteration, we have

$$X=D+AXA=D+A(D+AXA)A=D+ADA+A^2XA^2=\cdots=\sum_{i=0}^{\infty}A^iDA^i \tag{85}$$

$$=\mu\alpha K^2\Sigma+\lambda[-1+\mu(1+\mu-\mu\nu)]K\Sigma K-\lambda\mu(1-\nu)\alpha K(KQ+QK)K+(1-\mu^2)Q, \tag{86}$$

where we define $Q:=\sum_{i=0}^{\infty}A^i\Sigma A^i$.

Finally, it can be shown by expanding everything that Q satisfies

$$(I_D-A)Q(I_D+A)+(I_D+A)Q(I_D-A)=2\Sigma. \tag{87}$$

After simplification, we have

$$Q-AQA=\Sigma. \tag{88}$$

□

From Eqs. (73), the approximation error for QHM can be calculated.

Corollary 7. *The training error for QHM is*

$$L_{\mathrm{train}}=\frac{\lambda^2}{2}\mathrm{Tr}[h(K)^{-1}KC], \tag{89}$$

where

$$h(K):=\frac{1}{d}\left\{aI_D+2bK+cK^2+[\mu(1+\mu)f(K)+\lambda[1-\mu\nu(2+\mu)]g(K)](I_D-A^2)^{-1}\right\}, \tag{90}$$

$$f(K):=2(1-\mu^2)K+\lambda\left[-2+\mu[3+\mu(2-3\nu)]\right]K^2, \tag{91}$$

$$g(K):=2(1-\mu^2)I_D+2\lambda[-1+\mu(2+\mu-2\mu\nu)]K-2\lambda\mu(1-\nu)\alpha K^2. \tag{92}$$

Remark. *We emphasize that our result (89) is exact, whereas the result in Gitman et al. (2019) is obtained with a low-order approximation.*

Proof. By using the similar technique in Appendix E.1.4, Eq. (73) results in

$$h(K)K\Sigma+R=\lambda^2KC, \tag{93}$$

where R denotes the terms involving commutative factors such as $[\Sigma,K]$, etc, and

$$h(K):=\frac{1}{d}\left\{aI_D+2bK+cK^2+[\mu(1+\mu)f(K)+\lambda[1-\mu\nu(2+\mu)]g(K)](I_D-A^2)^{-1}\right\}, \tag{94}$$

$$f(K):=2(1-\mu^2)K+\lambda\left[-2+\mu[3+\mu(2-3\nu)]\right]K^2, \tag{95}$$

$$g(K):=2(1-\mu^2)I_D+2\lambda[-1+\mu(2+\mu-2\mu\nu)]K-2\lambda\mu(1-\nu)\alpha K^2. \tag{96}$$

By definition, the approximation error is

$$L_{\rm train} = \frac{1}{2}{\rm Tr}[K\Sigma] = \frac{\lambda^2}{2}{\rm Tr}[h(K)^{-1}KC]. \tag{97}$$

□

E.2.7. Parameter Fluctuations of Second-Order Optimization Methods

In this subsection, we deal with the covariance of the stationary distribution obtained by second-order optimization methods. We first deal with the stationary distribution of Damped Newton's Method (DNM), which is the oldest and most important second-order optimization method, first invented by Newton (Nesterov et al., 2018). It is of interest to investigate how the second-order methods behave asymptotically in a stochastic setting.

Theorem 12. (*Model fluctuation of DNM*) *Let the learning rate matrix be a matrix:* $\Lambda := \lambda K^{-1}$. *Then,*

$$\Sigma = \frac{1+\mu}{1-\mu}\frac{\lambda}{2(1+\mu)-\lambda}K^{-1}CK^{-1}. \tag{98}$$

Proof. Due to Theorem 3, while $\Lambda := \lambda K^{-1}$, Eq. (9) gives

$$\lambda\frac{1-\mu}{1+\mu}\left[2(1+\mu)-\lambda\right]\Sigma = \lambda^2 K^{-1}CK^{-1}. \tag{99}$$

Therefore, we have

$$\Sigma = \frac{1+\mu}{1-\mu}\frac{\lambda}{2(1+\mu)-\lambda}K^{-1}CK^{-1}. \tag{100}$$

□

Corollary. Suppose that the noise is due to minibatch sampling with the noise covariance being $C = \frac{N-S}{NS}K$. The model fluctuation is

$$\Sigma = \frac{1+\mu}{1-\mu}\frac{\lambda}{2(1+\mu)-\lambda}\frac{N-S}{NS}K^{-1}. \tag{101}$$

Proof. Substituting $C = \frac{N-S}{NS}K$ into Eq. (98) yields

$$\Sigma = \frac{1+\mu}{1-\mu}\frac{\lambda}{2(1+\mu)-\lambda}\frac{N-S}{NS}K^{-1}. \tag{102}$$

□

From Theorem 12, the approximation error for DNM can be calculated.

Corollary 8. *The approximation error for DNM is*

$$L_{\rm train} = \begin{cases} \frac{1}{2}{\rm Tr}[K\Sigma_{\rm general}] = \frac{1+\mu}{1-\mu}\frac{\lambda}{4(1+\mu)-2\lambda}{\rm Tr}[K^{-1}C]; \\ \frac{1}{2}{\rm Tr}\left[K\Sigma_{\rm minibatch}\right] = \frac{1+\mu}{1-\mu}\frac{D\lambda}{4(1+\mu)-2\lambda}\frac{N-S}{NS}. \end{cases} \tag{103}$$

Proof. The proof is simple by substituting Σ into the definition $L_{\rm train} = \frac{1}{2}{\rm Tr}[K\Sigma]$. □

Next, we consider the natural gradient descent (NGD) algorithm. In traditional statistics, the efficiency of any statistical estimator is upper bounded by the Cramér-Rao's inequality (CR bound) (Rao, 1992). An estimator that achieves the equality in the CR bound is said to be *Fisher-efficient*. A Fisher-efficient method is the fastest possible method to estimate a given statistical quantity. When the gradient descent is used, it is shown (Amari, 1998; Amari and Nagaoka, 2007) that if one defines the learning rate as a matrix, $\Lambda := \lambda J(\mathbf{w})^{-1}$, where $J(\mathbf{w}) := \mathbb{E}[\nabla L(\nabla L)^{\rm T}]$ is the Fisher information, then this optimization algorithm becomes Fisher-efficient in the limit of $t \to \infty$. This algorithm is called the *natural gradient descent*

because the Fisher information is the "natural" metric for measuring the distance in the probability space. The NGD algorithm has therefore attracted great attention both theoretically and empirically (Pascanu and Bengio, 2013; Amari, 1998). However, previous literature often takes the continuous-time limit and nothing is known about NGD in the discrete-time regime. We apply our formalism to derive the covariance of the stationary distribution of NGD in the discrete-time regime. To the best of our knowledge, this is the first work to treat the discrete-time NGD and to derive its asymptotic model fluctuations.

Theorem 13. (*Model covaraince matrix of NGD*) *Let the learning rate matrix be* $\Lambda := \lambda J(\mathbf{w})^{-1}$*, where* $J(\mathbf{w}) = \mathbb{E}[K\mathbf{w}\mathbf{w}^{\mathrm{T}}K] = K\Sigma K$ *is the Fisher information. Then the model parameter covariance matrix satisfies the following quadratic matrix equation*

$$(K\Sigma)^2 - \frac{\lambda}{2(1+\mu)}K\Sigma - \frac{\lambda}{2(1-\mu)}CK^{-1} = 0. \tag{104}$$

Proof. By setting $\Lambda = \lambda(K\Sigma K)^{-1}$ in Eq. (9), we have

$$2(1-\mu)K^{-1} - \frac{1-\mu}{1+\mu}\lambda K^{-1}\Sigma^{-1}K^{-1} = \lambda K^{-1}\Sigma^{-1}K^{-1}CK^{-1}\Sigma^{-1}K^{-1}. \tag{105}$$

Multiplying by $K\Sigma K$ to the left and $K\Sigma$ to the right yields

$$(K\Sigma)^2 - \frac{\lambda}{2(1+\mu)}K\Sigma - \frac{\lambda}{2(1-\mu)}CK^{-1} = 0. \tag{106}$$

□

This matrix equation can be solved while C does not depend on Σ explicitly (Higham and Kim, 2001).

Corollary 9. *Suppose that the noise covariance* C *is a constant matrix that does not depend on* Σ *explicitly. Then the solution to Eq.* (104) *is*

$$\Sigma = \frac{1}{2}K^{-1}\left[Q + \frac{\lambda}{2(1+\mu)}I_D\right], \tag{107}$$

where $Q := \left[\frac{\lambda^2}{4(1+\mu)^2}I_D + \frac{2\lambda}{1-\mu}CK^{-1}\right]^{\frac{1}{2}}$.

Remark. *This result does not seem quite satisfactory, especially because it does not seem to reduce to any meaningful distribution. This means that, when the noise is arbitrary and not related to the use of minibatch sampling, one is not recommended to use NGD*[4].

Proof. By referring to the conclusion in Higham and Kim (2001) that the solution to a quadratic matrix equation of the form $AX^2 + BX + C = 0$ with $A = I_D$ and $[B, C] = 0$ is $X = -\frac{1}{2}B + \frac{1}{2}(B^2 - 4C)^{1/2}$, Eq. (104) can be solved explicitly:

$$\Sigma = \frac{1}{2}K^{-1}\left[Q + \frac{\lambda}{2(1+\mu)}I_D\right], \tag{108}$$

where $Q := \left[\frac{\lambda^2}{4(1+\mu)^2}I_D + \frac{2\lambda}{1-\mu}CK^{-1}\right]^{\frac{1}{2}}$. □

Now we consider the case where the noise is induced by minibatch sampling. Instead of using the conventional Hessian approximation that $C \approx K$, we here consider a better approximation that $C \approx \frac{N-S}{NS}\mathbb{E}[\nabla L \nabla L^T] = \frac{N-S}{NS}K\Sigma K$. The model fluctuation can be calculated.

Corollary. Let the NGD algorithm be updated with noise covariance being $C = \frac{N-S}{NS}K\Sigma K$. Then,

$$\Sigma = \lambda\frac{(1+\mu)\frac{N-S}{NS} + 1 - \mu}{2(1-\mu^2)}K^{-1}. \tag{109}$$

[4] Recall that the NGD is derived in an online learning setting, where the noise is by definition proportional to the minibatch noise with $N \to \infty$ and minibatch size 1 (Amari, 1998).

Proof. Substituting $C = \frac{N-S}{NS} K\Sigma K$ into Eq. (104) yields

$$(K\Sigma)^2 - \lambda \frac{(1+\mu)\frac{N-S}{NS} + 1 - \mu}{2(1-\mu^2)} K\Sigma = 0. \tag{110}$$

Because $K\Sigma$ is positive definite, we have

$$\Sigma = \lambda \frac{(1+\mu)\frac{N-S}{NS} + 1 - \mu}{2(1-\mu^2)} K^{-1}. \tag{111}$$

□

From Theorem 13, the approximation error can be calculated.

Corollary 10. *The approximation error for NGD is*

$$L_{\text{train}} = \begin{cases} \frac{1}{2}\text{Tr}[K\Sigma_{\text{general}}] = \frac{1}{4}\text{Tr}\left[Q + \frac{\lambda}{2(1+\mu)} I_D\right]; \\ \frac{1}{2}\text{Tr}\left[K\Sigma_{\text{minibatch}}\right] = \lambda \frac{(1+\mu)\frac{N-S}{NS}+1-\mu}{4(1-\mu^2)} D. \end{cases} \tag{112}$$

Proof. The proof is simple by substituting Σ into the definition $L_{\text{train}} = \frac{1}{2}\text{Tr}[K\Sigma]$. □

E.2.8. Proof of Theorem 10

Proof. Using the non-diagonal approximation, the preconditioning matrix at asymptotic time is

$$\begin{aligned} \Lambda &= \lambda \mathbb{E}[\mathbf{g}\mathbf{g}^{\mathrm{T}}]^{-\frac{1}{2}} \\ &= \lambda \mathbb{E}\left[(K\mathbf{w} + \eta)(K\mathbf{w} + \eta)^{\mathrm{T}}\right]^{-\frac{1}{2}} \\ &= \lambda (K\Sigma K + C)^{-\frac{1}{2}} \\ &= \frac{\lambda}{\sqrt{1+c}} (K\Sigma K)^{-\frac{1}{2}}. \end{aligned} \tag{113}$$

Substituting it into Eq. (9), we have

$$\Lambda K\Sigma + \Sigma K\Lambda - \Lambda K\Sigma K\Lambda = c\Lambda K\Sigma K\Lambda, \tag{114}$$

which can be rewritten as

$$\Lambda^{-1}K^{-1} + K^{-1}\Lambda^{-1} = (1+c) I_D. \tag{115}$$

It can be solved to give that

$$\Sigma = \frac{\lambda^2(1+c)}{4} I_D. \tag{116}$$

□

Remark. *The approximation error can be obtained easily as*

$$L_{\text{train}} = \frac{1}{2}\text{Tr}[K\Sigma] = \frac{\lambda^2(1+c)}{8}\text{Tr}[K]. \tag{117}$$