Appendix

Organization of Appendix. In Appendix A we present details of the numerical experiments and additional numerical experiments. We give more discussions of our work in Appendix B. Appendix C provides missing technical details of Section 3 while Appendix D provides those of Section 4.

A Details of Numerical Experiments and Additional Experiments

In A.1, we present the details of the numerical experiments. We compare the rank-1 linear networks with the diagonal linear networks empirically in A.2. Finally, in A.3, we conduct additional experiments to further verify the "alleviting" effect of the SGD sampling noise mentioned in Theorem 3. In A.4, we conduct experiments when the balanced initialization condition is not satisfied.

Data. We conduct over-parameterized regression with different linear networks. For the dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, we set $n = 40, d = 100$ and $x_i \sim \mathcal{N}(0, I)$. For $i \in \{1, \dots, n\}$, y_i is generated by $y_i = \theta^{*T} x_i$ where $\theta^* \in \mathbb{R}^d$, i.e., θ^* is the ground truth solution. We let 20 components of θ^* be informative.

A.1 Details of Numerical Experiments in Section 5

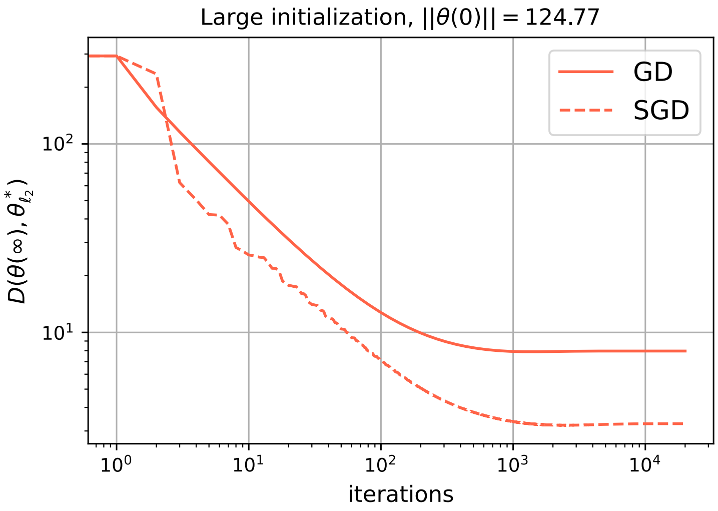


Figure 7: $D(\theta(t), \theta^*_{\ell_2})$ along training for rank-1 linear networks when the initialization is extremely large. SGD solution is closer to $\theta^*_{\ell_2}$ when compared to the GD solution for rank-1 linear networks.

We now present the details of numerical experiments conducted in Section 5.

ℓ_2 minimum norm solution. To get the ℓ_2 minimum norm solution $\theta^*_{\ell_2}$, we train a single layer linear network with zero initialization using GD for 20000 iterations, since GD will return an ℓ_2 minimum norm solution solution in this case according to Corollary 1.1 and [26].

Rank-1 linear networks. For the rank-1 linear network $f(x; u, v) = w_L^T W_{L-1} \cdots W_1 x$ where $W_k = u_k v_k^T \in \mathbb{R}^{d_k \times d_{k+1}}$, we let $L = 3$ and $\forall k \in \{1, \dots, L\} : d_k = 100$. The learning rate is 10^{-3} and the batch size is 4 if we run SGD. We construct different rank-1 linear networks as follows: for a randomly sampled $\tilde{\theta} \in \mathbb{R}^{100}$, we let the initialization $\theta_i(0)$ of the i-th rank-1 networks have the same direction as $\tilde{\theta}$ but with different scales. We then train each rank-1 linear network with GD and SGD, respectively, for 20000 iterations. In particular:

1. Fig. 5(a) presents the results of the distances between $\theta^*_{\ell_2}$ and GD and SGD solutions, respectively, of each trained rank-1 networks with different initialization scales.

2. In Fig. 5(b), we measure the distances between θ^* and GD and SGD solutions, respectively, for each trained rank-1 networks with different initialization scales.
3. Fig. 6(a) plots the distances between $\theta^*_{\ell_2}$ and the model parameters θ along training when the initialization scales are different for both GD and SGD. The numbers in the bracket denote $\|\theta(0)\|$.
4. Fig. 7 is about the distances between $\theta^*_{\ell_2}$ and the model parameters θ along training for both GD and SGD when $\|\theta(0)\|$ is extremely large.

Standard linear networks. For the standard linear network $f(x;W) = w_L^T W_{L-1} \cdots W_1 x$ where $W_k \in \mathbb{R}^{d_k \times d_{k+1}}$, we let $L = 4$ and $\forall k \in \{1, 2, 3\} : d_k = 100$. The learning rate is 10^{-3} and the batch size is 4 if we run SGD. Other settings are similar to that of rank-1 linear networks.

1. In Fig. 5(c), we plot the distances between $\theta^*_{\ell_2}$ and GD and SGD solutions, respectively. Similar to the case of rank-1 linear networks, for all initialization scales, $D(\theta(\infty), \theta^*_{\ell_2})$ is smaller if the network is trained with SGD when compared to GD.
2. Fig. 5(d) presents the results of the distances between θ^* and GD and SGD solutions.
3. Fig. 6(b) plots the distances between $\theta^*_{\ell_2}$ and the model parameters θ along training when the initialization scales $\|\theta(0)\|$ are different for both GD and SGD. The numbers in the bracket denote $\|\theta(0)\|$.

Non-linear networks. For the non-linear network $f(x;W) = w_L^T \sigma(W_{L-1} \cdots \sigma(W_1 x))$ where $W_k \in \mathbb{R}^{d_k \times d_{k+1}}$, we let $L = 4$ and $\forall k \in \{1, 2, 3\} : d_k = 100$. The learning rate is 10^{-3} and the batch size is 4 if we run SGD. We use the ReLU activation $\sigma(x) = \mathrm{ReLU}(x)$. We use the same dataset as in the experiments of rank-1 linear networks. Since the non-linear networks do not have the overall parameterization of θ as in the linear networks case, to measure the initialization scale, we first straight all weight matrices to vectors and stack them to get a single vector, then we calculate the ℓ_2 norm of this vector as the scale of the initialization of a non-linear network, i.e., we use $\sqrt{(\sum_k \|W_k(0)\|_F^2)}$ as the initialization scale where $\|\cdot\|_F$ is the Frobenius norm. Due to the same reason, we can not measure quantities such as $D(\theta, \theta^*_{\ell_2})$, therefore, we report the test error of the model in a newly sampled test set instead. Fig. 6(c) plots the test error of the model along training when the initialization scales are different for both GD and SGD. The numbers in the bracket denote initialization scales.

A.2 Additional Experiments of Comparison with Diagonal Nets

Results in Section 3.2 indicate that diagonal linear networks exhibit different implicit bias in comparison with rank-1 and standard linear networks, e.g., both rank-1 and standard linear networks prefer ℓ_2 minimum norm solution for GD when the initialization is nearly-zero while, on the contrary, diagonal linear networks prefer such solution when the initialization is sufficiently large. In this section, we empirically compare the implicit bias for rank-1 linear networks and diagonal linear networks to show this phenomenon.

In particular, we use the same settings as in the experiments for rank-1 linear networks in A.1 while only change the model to diagonal linear networks. As in previous works [3, 22], the re-parameterization of diagonal linear network is

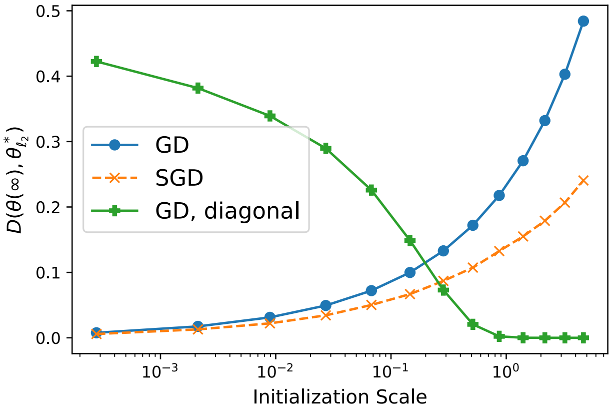
$$\theta = w_+ \odot w_+ - w_- \odot w_-,$$

where $w_+ \in \mathbb{R}^{100}, w_- \in \mathbb{R}^{100}$ and $\odot$ is the elementwise product. Let $\mathbf{e} = (1, \cdots, 1)^T \in \mathbb{R}^{100}$, we set the initialization as

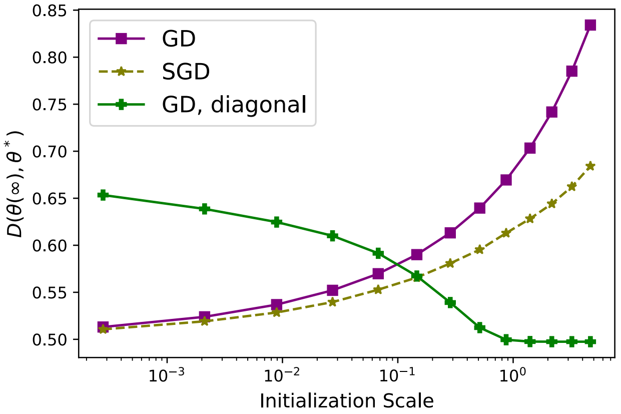
$$w_+(0) = C\mathbf{e}, \quad w_-(0) = C\mathbf{e},$$

where C is a positive constant measuring the initialization scale. For each diagonal linear network with different C, we run GD for 20000 iterations and calculate $D(\theta(\infty), \theta^*_{\ell_2})$ and $D(\theta(\infty), \theta^*)$. The results are plotted in Fig. 8, where, for convenience of comparing the implicit bias of GD for rank-1 linear networks with that of diagonal linear networks, we also plot the results of rank-1 linear networks of Fig. 5(a) and Fig. 5(b) in Fig. 8(a) and Fig. 8(b), respectively.

As shown in Fig. 8(a), as the initialization scale ($\|\theta(0)\|$ for rank-1 linear networks, C for diagonal linear networks) increases, $D(\theta(\infty), \theta^*_{\ell_2})$ decreases for rank-1 linear linear networks trained with



(a)



(b)

Figure 8: For different initialization scale ($\|\theta(0)\|$ for rank-1 linear networks and C for diagonal linear networks): **(a)** $D(\theta(\infty), \theta^*_{\ell_2})$ for rank-1 linear nets and diagonal linear networks (the green solid line). **(b)** $D(\theta(\infty), \theta^*)$ for rank-1 linear nets and diagonal linear nets (the green solid line).

both GD and SGD, while it increases for diagonal linear networks trained with GD. This indicates the drastic difference between the implicit bias of GD exhibited by diagonal linear networks and rank-1 linear networks (also standard linear networks).

A.3 Additional Experiments for the "Alleviating" Effect in Theorem 3

Recall the form of $V^{\mathcal{S}}(\theta, t)$ in Theorem 3

$$V^{\mathcal{S}}(\theta,t) = \frac{1}{\Omega_L}\|\theta\|^{\Omega_L} + \frac{\theta^T\theta(0)}{\|\theta(0)\|^{\lambda_L}} - \frac{2\lambda_L\eta\theta^T}{nb}\int_0^t \frac{\mathcal{L}\left(\theta(s)\right)\operatorname{tr}\left(P_\perp(\theta(s))X^TX\right)}{\|\theta(s)\|^{2-\lambda_L}}\theta(s)ds,$$

we let

$$p_\theta(t) = \frac{\theta^T(t)\theta(0)}{\|\theta(0)\|^{\lambda_L}}, \tag{13}$$

$$q_\theta(t) = \frac{2\lambda_L\eta\theta^T(t)}{nb}\int_0^t \frac{\mathcal{L}\left(\theta(s)\right)\operatorname{tr}\left(P_\perp(\theta(s))X^TX\right)}{\|\theta(s)\|^{2-\lambda_L}}\theta(s)ds. \tag{14}$$

To quantitatively measure the "alleviating" effect of the SGD sampling noise, we train 3 rank-1 linear networks with different initialization scales using SGD. The batch size is 4 and the learning rate

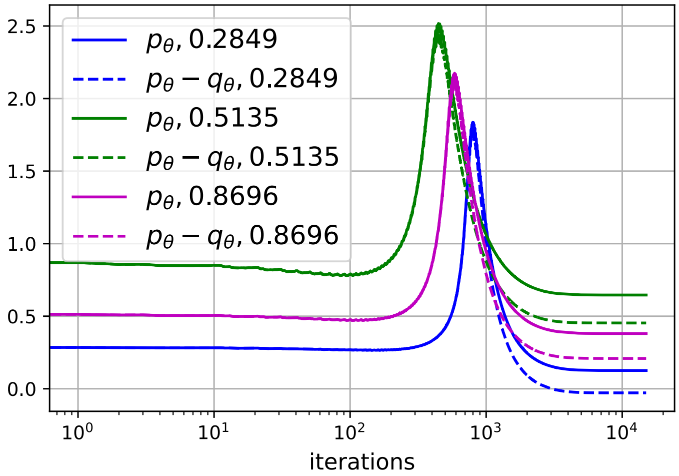


Figure 9: SGD sampling noise alleviates the dependence on the initialization. Numbers after the comma denote the initialization scales. The solid lines are for p_θ (Eq. (13)) and the dotted lines are for $p_\theta - q_\theta$ (Eq. (14)).

is 5×10^{-5}. For the rank-1 linear network $f(x; u, v) = w_L^T W_{L-1} \cdots W_1 x$ where $W_k = u_k v_k^T \in \mathbb{R}^{d_k \times d_{k+1}}$, we let $L = 3$ and $\forall k \in \{1, \ldots, L\} : d_k = 100$. We calculate both $p_\theta(t)$ and $q_\theta(t)$ along training, where $q_\theta(t)$ measures the alleviating effect of the SGD sampling noise and their difference $p_\theta(\infty) - q_\theta(\infty)$ is the alleviated initialization dependence of the SGD solution compared to GD solution.

As shown in Fig. 9, the effect coming from the SGD sampling noise, q_θ, is equivalent to make the dependence of V^S on the initialization closer to 0 (after about 1000 iterations, every dotted line is closer to the x-axis compared to the corresponding solid line with the same color), thus it controls the dependence of the SGD solution on the initialization. This phenomenon further verifies our claims.

A.3.1 Training loss for Fig. 4(b)

Fig. 4(b) indicates that the final direction of the integral term in Eq. (12) highly depends on the initialization $\theta(0)$ since the loss decays along training. To further support this argument, here we present the training loss when we perform the experiments of Fig. 4(b) in Fig. 10. It can be seen that, for a random initialization, the loss, the magnitude of the speed of α, has a high value at the start of the training, and decays very quickly, which explains why the direction of $\theta(0)$ is crucial to that of $\alpha(\infty)$.

A.4 Additional Experiments for Biased Initialization

In this section, we provide additional experiments to show that our conclusion still holds when removing the balanced initialization condition (Definition 1).

To make the initialization unbalanced, we add a small perturbation to the balanced initialization. Specifically, we define

$$\Delta = \frac{1}{2L-1} \sum_{k=1}^{L-1} \frac{|\|v_{k+1}\|^2 - \|u_k\|^2|}{\|u_k\|^2}$$

as the scale of the perturbation to the balanced initialization (larger Δ implies that the initialization is more unbalanced). All the other experiment details are kept unchanged as in Section 5. As shown in Fig. 11 and Fig. 12, we still observe similar phenomenons as in the case of the balanced initialization, e.g., SGD solutions are closer to the ℓ_2-norm minimization solution compared to GD, when a small perturbation is added to the balanced initialization. Thus the implicit bias is not unique to the balanced initialization. In particular:

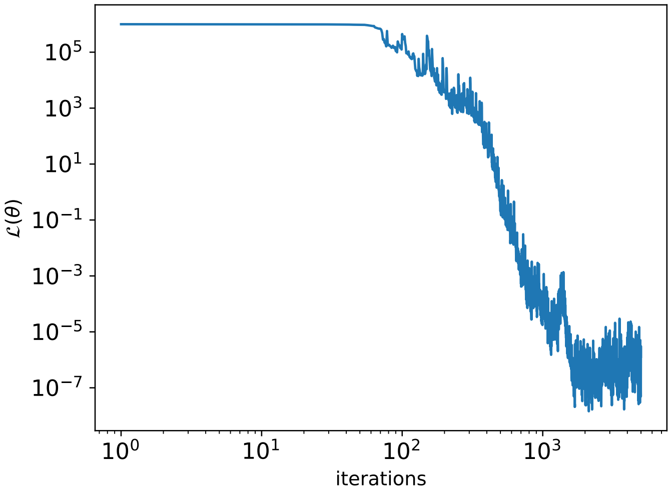


Figure 10: The empirical loss $\mathcal{L}(\theta)$ along training for Fig. 4(b).

- We report $D(\theta(\infty), \theta^*_{\ell_2})$ for both GD and SGD for different levels of perturbation Δ (denoted in the title of each figure) in Fig. 11. In the last figure, we fix the initialization scale and report $D((\theta(\infty), \theta^*_{\ell_2})$ of both GD and SGD for different Δ. It can be seen that, without the balanced initialization, GD and SGD still prefer ℓ_2-norm minimization solution $\theta^*_{\ell_2}$ for small initialization, while the SGD solution is closer to $\theta^*_{\ell_2}$ due to its initialization reduction effect.
- We report $D(\theta(t), \theta^*_{\ell_2})$ during optimization for both GD and SGD for different levels of perturbation Δ and the same scale of initialization ($\|\theta(0)\| = 0.8696$) in Fig. 12, which further clearly reveals that there are still similar phenomenons when $\Delta \neq 0$ as in the case when the initialization is balanced.

B More Discussions

Our work proposes the rank-1 linear network which is a plausible proxy of standard linear networks with some neurons fully connected with neurons in its last and next layers. By showing that the proposed rank-1 linear networks are standard linear networks with special initialization, our conclusions may be generalized to standard linear networks. In comparison, the diagonal linear network, a special kind of linear networks that receives a lot of attention recently, does not have fully connected neurons. Furthermore, we find that the implicit bias of both GD and SGD for diagonal linear networks are not consistent with ours. The diagonal linear networks also exhibit drastically different implicit bias of GD when compared to standard linear networks, while the conclusions for rank-1 linear networks are consistent with those of standard linear networks. We also reveal the key role of the over-parameterization in characterizing the implicit bias of SGD, namely that it will only be different with that of GD for over-parameterization model.

The inconsistency between the implicit bias of GD and SGD for diagonal linear networks and rank-1 linear networks leads us to suggest intriguing questions for future work such as *what about other architectures* and *is there any unified analytical approach for studying implicit bias of GD and SGD for different architectures?* And it is interesting to reveal whether the "alleviating" initialization effect of the SGD sampling noise is general accross different architectures.

We precisely characterize the implicit bias of both GD and SGD for rank-1 linear networks, where the dependence on the initialization and depth is explicit and clear. In this sense, we take a step forward in the direction of characterizing the implicit bias of optimization algorithms.

Finally, our analysis characterizes the implicit bias of SGD through analyzing the overall parametrization θ. This is different with another line of recent work [4, 10, 19, 27, 9] which focused on the

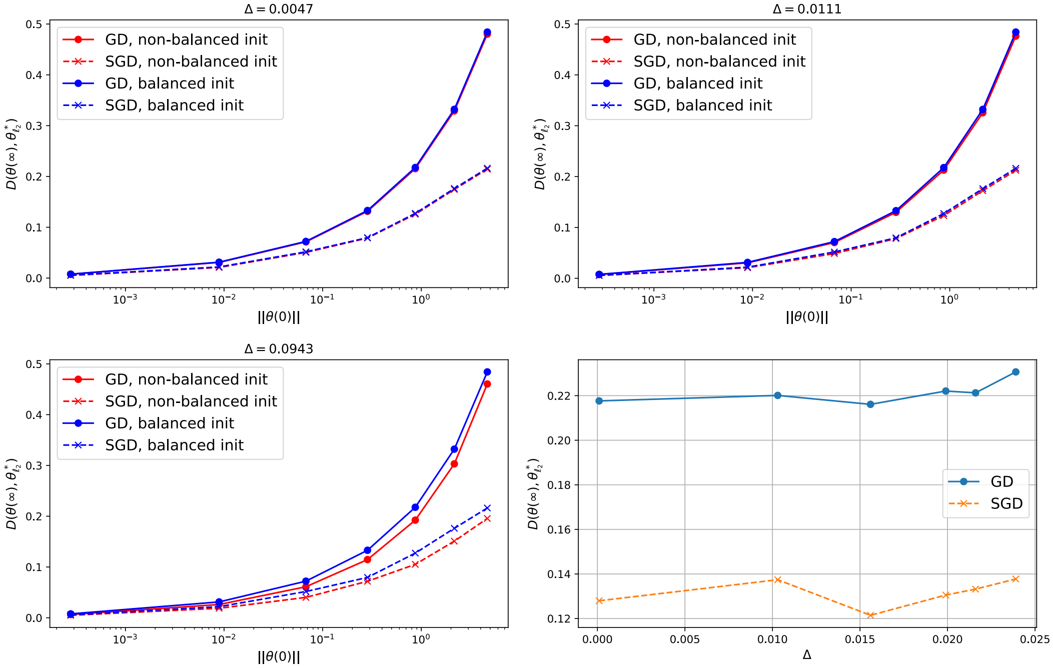


Figure 11: $D(\theta(\infty), \theta^*_{\ell_2})$ for different $\|\theta(0)\|$ when the initialization is unbalanced ($\Delta \neq 0$, larger Δ means the initialization is more unbalanced). We use solid lines for the results of GD and dashed lines for SGD. For results under the balanced initialization, we use blue lines; for the results when a small perturbation is added to the balanced initialization, i.e., $\Delta \neq 0$, we use red lines.

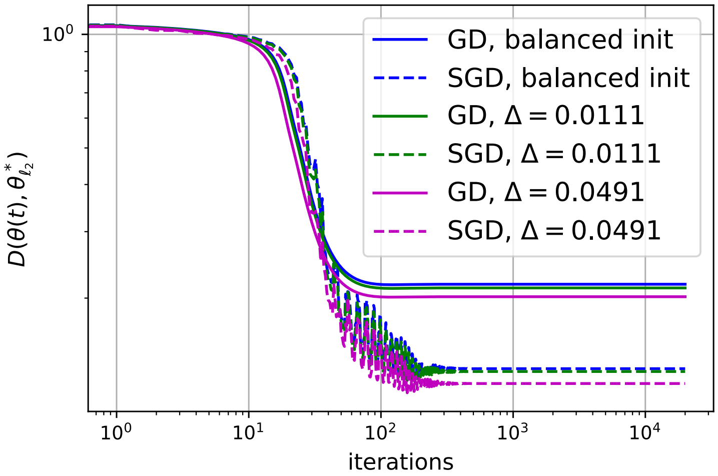


Figure 12: $D(\theta(t), \theta^*_{\ell_2})$ along training for rank-1 linear networks when the initialization is unbalanced ($\Delta \neq 0$).

flatness of the loss landscape by directly analyzing the independent model parameters, i.e., u and v for the rank-1 linear networks, which is also crucial to fully understand the learning dynamics.

It is worth to mention that removing the balanced initialization (Definition 1) is also a promising direction. As verified by the numerical experiments in Appendix A.4, similar phenomena exist for unbalanced initialization. From the theoretical aspect, balanced initialization enables us to derive the exact dynamics of the overall parameter θ, which is necessary to precisely characterize the implicit bias of GD/SGD. And it is difficult to discuss arbitrary initialization without the balanced initialization assumption. The effect of removing the balanced initialization is that the induced mirror flow potential should be composed of two parts: the original potential presented in Section 3 and a perturbation due to the imbalance of the initialization to it. This implies that the ℓ_2-norm solution is still returned for small initialization. On the other hand, the case for SGD is much more complicated: the Brownian motion term of the corresponding SDE will also be affected by the imbalance of the initialization, which in turn induces a much more complex time varying mirror flow potential. We believe that the exact theoretical characterization of the implicit bias of SGD without the balanced initialization is a valuable future direction.

There are also some limitations in the current work. For example, although the conclusions of numerical experiments conducted on non-linear networks resemble that of rank-1 and standard linear networks, we can not directly generalize the current theoretical analysis to non-linear neural networks, which normally do not have overall parametrization vectors as θ that is necessary for our analysis. Moreover, the exact characterization of the stochastic integral is also absent in the current work, while we expect that the integral term $\mathcal{L}(\theta)\text{tr}\left(P_\perp(\theta)X^TX\right)\frac{\theta}{\|\theta\|}$ in Eq. (12) might have close relation with the property of the training dynamics.

C Proofs for Section 3

In this section, we present the technical details of Section 3. In particular, Section C.1 discusses the balanced initialization, Section C.2 proves Theorem 1 for rank-1 linear networks and Section C.3 proves Theorem 2 for standard linear networks.

For a rank-1 linear network $f(x;u,v) = w_L^T W_{L-1}\cdots W_1 x$ where $W_k = u_k v_k^T$ for any $k \in \{1,\dots,L-1\}$, recalling the definition Eq. (3) and that the network $f(x;u,v)$ can be written as $f(x;u,v) = \rho_k v_{k+1}^T u_k \rho_{-k} v_1^T x$. For convenience, we let $\xi = w_L^T W_{L-1}\cdots W_2 u_1$.

C.1 Balanced initialization

For a rank-1 linear network Eq. (2), dynamics of gradient flow is given by

$$\frac{du_k}{dt} = -\frac{2}{n}\rho_k\rho_{-k}v_1^T X^T r v_{k+1}, \tag{15}$$

$$\frac{dv_{k+1}}{dt} = -\frac{2}{n}\rho_k\rho_{-k}v_1^T X^T r u_k. \tag{16}$$

Based on this set of dynamics, we first discuss he following useful lemma that characterizes the dynamics of norms of model parameters:

Lemma 1. *For $f(x;u,v)$ trained with gradient flow, we have*

$$\forall k \in \{1,\dots,L-1\} : \frac{d\|u_k\|^2}{dt} = \frac{d\|v_k\|^2}{dt} = \frac{d\|v_{k+1}\|^2}{dt}, \tag{17}$$

i.e., layer norms grow at the same rate. Furthermore, if $\forall k \in \{1,\dots,L-1\} : \|u_k(0)\| = \|v_{k+1}(0)\| = \|v_k(0)\|$, we have

$$\forall k \in \{2,\dots L-1\} : \frac{d\left\langle v_{k+1}, u_k\right\rangle^2}{dt} = \frac{d\left\langle v_k, u_{k-1}\right\rangle^2}{dt}. \tag{18}$$

Proof. Using Eq.(16), we have

$$\frac{1}{2}\frac{d\|u_k\|^2}{dt} = \left(\frac{du_k}{dt}\right)^T u_k = -\frac{2}{n}\rho_k\rho_{-k}v_1^T X^T r v_{k+1}^T u_k = -\frac{2}{n}\xi v_1^T X^T r, \tag{19}$$

$$\frac{1}{2}\frac{d\|v_{k+1}\|^2}{dt} = \left(\frac{du_k}{dt}\right)^T u_k = -\frac{2}{n}\rho_k\rho_{-k}v_1^T X^T r u_k^T v_{k+1} = -\frac{2}{n}\xi v_1^T X^T r. \tag{20}$$

Therefore, both $\frac{d\|u_k\|^2}{dt}$ and $\frac{d\|v_{k+1}\|^2}{dt}$ do not depend on k and are same then Eq. (17) follows.

We now discuss Eq. (18). Since we assume $\forall k \in \{1, \dots, L-1\} : \|u_k(0)\| = \|v_{k+1}(0)\| = \|v_k(0)\|$ and Eq. (17) implies that for any $t > 0$:

$$\|u_k(t)\|^2 - \|u_k(0)\|^2 = \|v_{k+1}(t)\|^2 - \|v_{k+1}(0)\|^2 = \|v_k(t)\|^2 - \|v_k(0)\|^2, \tag{21}$$

we have

$$\|u_k(t)\|^2 = \|v_{k+1}(t)\|^2 = \|v_k(t)\|^2 = \|u_1(t)\|^2 \tag{22}$$

To show Eq. (18), we note that

$$\begin{aligned}\frac{d\langle v_{k+1}, u_k\rangle}{dt} &= \left(\frac{dv_{k+1}}{dt}\right)^T u_k + v_{k+1}^T\frac{du_k}{dt} \\ &= -\frac{2}{n}\rho_k\rho_{-k}v_1^T X^T r(\|u_k\|^2 + \|v_{k+1}\|^2) \\ &= -\frac{2}{n}\xi v_1^T X^T r\frac{\|u_k\|^2 + \|v_{k+1}\|^2}{\langle v_{k+1}, u_k\rangle} = -\frac{4}{n}\xi v_1^T X^T r\frac{\|u_1\|^2}{\langle v_{k+1}, u_k\rangle},\end{aligned} \tag{23}$$

where we use Eq.(16) in the second equality and the third equality is because $\xi = \rho_k\rho_{-k}\langle v_{k+1}, u_k\rangle$. As a result, the above equation implies that

$$\frac{1}{2}\frac{d(\langle v_{k+1}, u_k\rangle)^2}{dt} = -\frac{4}{n}\xi v_1^T X^T r\|u_1\|^2, \tag{24}$$

which does not depend on k, and Eq. (18) follows. □

To simplify the analysis, in Theorem 1, we have required the balanced initialization across layers (Definition 1). Recall that the balanced initialization is defined as

Definition 1 (Balanced initialization for rank-1 linear networks). *Given an L-layer rank-1 linear network Eq. (2), for any $k \in \{1, \dots, L-1\}$, the balanced initialization means that*

$$\frac{\langle v_{k+1}(0), u_k(0)\rangle^2}{\|v_{k+1}(0)\|^2\|u_k(0)\|^2} = 1, \tag{25}$$

$$\|v_{k+1}(0)\| = \|u_k(0)\| = \|v_1(0)\|. \tag{26}$$

Eq. (25) states that u_k of the k-th layer is aligned with v_{k+1} of the $(k+1)$-th layer in direction while Eq. (26) means they have the same magnitudes as $v_1(0)$. The balanced initialization has been suggested by several previous works [3, 2, 6, 28] for standard linear networks defined as follows.

Definition 2 (Balanced initialization for standard linear networks)**.** *Given an L-layer standard linear network $f(x; W) = w_L^T W_{L-1}\cdots W_1 x$, for any $k \in \{1, \dots, L-1\}$, the balanced initialization means that*

$$W_{k+1}^T(0)W_{k+1}(0) = W_k(0)W_k^T(0)$$

for any $k \in \{1, \dots, L\}$.

This directly means that $W_{k+1}(0)$ and $W_k(0)$ share same singular values and W_{k+1}'s right singular vector aligns with the left singular vector of $W_k(0)$. In our case, such reasoning gives us

$$\frac{v_{k+1}(0)}{\|v_{k+1}(0)\|} = \frac{u_k(0)}{\|u_k(0)\|} \tag{27}$$

$$\|v_{k+1}(0)\|\|u_{k+1}(0)\| = \|v_k(0)\|\|u_k(0)\|, \tag{28}$$

where Eq. (27) is similar to Eq. (25), which shows that $u_k(0)$ aligns with $v_{k+1}(0)$, and we adapt the condition (28) to Eq. (26) for rank-1 linear net since v_k and u_k are the independent model parameters.

A nice property of GD is that the balanced property across layers will be maintained during training [6, 11, 2], i.e., $W_{k+1}^T(t)W_{k+1}(t) = W_k(t)W_k^T(t)$ for $t > 0$ and $k \in \{1, \dots, L\}$, which can be showed by taking derivative with respect to time on $W_{k+1}^T(t)W_{k+1}(t)$ and $W_k(t)W_k^T(t)$. For the rank-1 linear network case, according to Lemma 1 and Eq. (25), $\langle v_{k+1}(t), u_k(t)\rangle$ are the same for all $k \in \{1, \dots, L-1\}$, thus GD also maintains the balanced property for rank-1 deep linear networks. Although the balanced initialization conditions are slightly strict, it can be approximately accurate if the initialization scale is not large, which is rather common in practice. Under such initialization conditions, we are able to precisely characterize the implicit bias of GD and focus more on the effects coming from the overall initialization of model parameters, rather than the difference between layers.

C.2 Proof of Theorem 1

In this section, we prove Theorem 1. Basically the idea is to show the existence of a potential function $V(\theta)$ such that the model parameter θ follows a mirror descent with respect to $V(\theta)$:

$$\theta(\infty) = \arg\min_{\theta} V(\theta) \quad s.t.\ X\theta = y. \tag{29}$$

This method is called infinitesimal mirror descent (IMD) approach and can be found in, e.g., [3]. Note that, to apply this method, the dynamics of θ should satisfy certain condition, which might be strict. For example, given the linear model $f(x;\theta) = \theta^T x$ for $\theta \in \mathbb{R}^d$, both parameterization of θ with standard linear networks, i.e. $\theta^T = w_L^T W_{L-1} \cdots W_1$, and a much simpler one $\theta = cv$ for $c \in \mathbb{R}$ and $v \in \mathbb{R}^d$ do not satisfy the condition of applying the IMD approach, while the rank-1 linear networks satisfy the condition, which implies that the parameterization of rank-1 linear networks is different with a scalar times a vector. Furthermore, the above example also confirms our motivation of studying rank-1 linear networks as a proxy of standard linear networks, especially considering that the implicit bias of SGD for rank-1 linear networks is more amenable.

We first present a useful Lemma in [20]:

Lemma 2. *If H has rank 1 and G is invertible, then*

$$(G+H)^{-1} = G^{-1} - \frac{1}{1+g} G^{-1} H G^{-1}. \tag{30}$$

where $g = tr\left(HG^{-1}\right)$.

We now prove Theorem 1.

Proof. Recall that $r \in \mathbb{R}^n$ with $r_i = (f_i - y_i)$, we let

$$\Phi = \sum_{k=1}^{L-1} \phi_k, \tag{31}$$

$$\phi_k = \rho_k^2 \rho_{-k}^2 (\|u_k\|^2 + \|v_{k+1}\|^2). \tag{32}$$

The key step is to derive the dynamics of the overall parameter θ which can be done by noting that

$$\frac{d\theta}{dt} = v_1 \frac{d\xi}{dt} + \xi \frac{dv_1}{dt},$$

where, according to Lemma 1,

$$\begin{aligned}
\frac{d\xi}{dt} &= \sum_{k=1}^{L-1} \rho_k \rho_{-k} \frac{d\langle v_{k+1}, u_k\rangle}{dt} \\
&= -\frac{2}{n} r^T X v_1 \sum_{k=1}^{L-1} \rho_k^2 \rho_{-k}^2 (\|u_k\|^2 + \|v_{k+1}\|^2) \\
&= -\frac{2}{n} \Phi v_1^T X^T r && (33) \\
\frac{dv_1}{dt} &= -\frac{2}{n} \xi X^T r && (34) \\
\implies \frac{d\theta}{dt} &= -\frac{2}{n} \left(\xi^2 I + \Phi v_1 v_1^T\right) X^T r. && (35)
\end{aligned}$$

Eq. (35) can be rewritten as

$$\left(\xi^2 I + \Phi v_1 v_1^T\right)^{-1} \frac{d\theta}{dt} = -\frac{2}{n} X^T r. \tag{36}$$

According to Lemma 2, note that

$$\operatorname{tr}\left(\Phi \frac{v_1 v_1^T}{\xi^2}\right) = \frac{\Phi \|v_1\|^2}{\xi^2},$$

the inverse appeared in Eq. (36) is

$$(\xi^2 I + \Phi v_1 v_1^T)^{-1} = \frac{1}{\xi^2} I - \frac{\xi^{-2} v_1 v_1^T \Phi \xi^{-2}}{1 + \frac{\Phi \|v_1\|^2}{\xi^2}} = \frac{1}{\xi^2} I - \frac{\theta\theta^T}{\frac{\xi^6}{\Phi} + \xi^2 \|\theta\|^2}. \tag{37}$$

It is now left for us to express ξ^2 and Φ in terms of θ. In the following, we assume the balanced initialization in Theorem 1 and apply Lemma 1.

1. ξ^2. This can be done by noting that $\|\theta\|^2 = \xi^2 \|v_1\|^2$, where ξ^2 is given by

$$\xi^2 = \prod_{k=1}^{L-1} \langle v_{k+1}, u_k \rangle^2. \tag{38}$$

Note that $\langle v_{k+1}, u_k \rangle^2$ grow at the same rate for different k according to Lemma 1 and $\langle v_{k+1}, u_k \rangle^2$ are the same at initialization for different k due to our assumption, we have $\langle v_{k+1}, u_k \rangle^2 = \langle v_2, u_1 \rangle^2$ and $\xi^2 = \langle v_2, u_1 \rangle^{2(L-1)}$. Note that

$$\frac{1}{2} \frac{d \langle v_2, u_1 \rangle^2}{dt} = -\frac{4}{n} \xi v_1^T X^T r \|u_1\|^2 = -\frac{4}{n} \xi v_1^T X^T r \|v_1\|^2 = \frac{1}{2} \frac{d\|v_1\|^4}{dt}, \tag{39}$$

we have $\langle v_2, u_1 \rangle^2 - \langle v_2(0), u_1(0) \rangle^2 = \|v_1\|^4 - \|v_1(0)\|^4$. Since we have $\langle v_2, u_1 \rangle^2 = \|u_1\|^4 = \|v_1\|^4$ at initialization according to our assumption, ξ^2 can be finally written as

$$\xi^2 = \|v_1\|^{4(L-1)}. \tag{40}$$

As a result,

$$\|\theta\| = \|v_1\|^{2L-1} \implies \|v_1\| = \|\theta\|^{\frac{1}{2L-1}}, \quad \xi^2 = \|\theta\|^{\frac{4(L-1)}{2L-1}}. \tag{41}$$

2. ξ^6/Φ. By taking some simple algebra, we have

$$\Phi = \frac{2(L-1)\xi^2}{\|v_1\|^2} \implies \frac{\xi^6}{\Phi} = \frac{\xi^4 \|v_1\|^2}{2(L-1)}. \tag{42}$$

Now Eq. (36) becomes

$$\|\theta\|^{-\frac{2(L-1)}{2L-1}} \left(I - \frac{\theta\theta^T}{\frac{\|\theta\|^2}{2(L-1)} + \|\theta\|^2} \right) \frac{d\theta}{dt} = -\frac{2|\xi|}{n} X^T r. \tag{43}$$

These conditions are now sufficient for us to find the form of the potential $V(\theta)$. Suppose that $V(\theta)$ can be written as

$$V(\theta) = \hat{V}(\|\theta\|) + h^T \theta \tag{44}$$

for some vector h and satisfies the following relation:

$$\begin{aligned} \nabla_\theta^2 V(\theta) &= \nabla_\theta^2 \hat{V}(\theta) \\ &= \|\theta\|^{-\frac{2(L-1)}{2L-1}} \left(I - \frac{1}{1 + \frac{1}{2(L-1)}} \frac{\theta\theta^T}{\|\theta\|^2} \right), \end{aligned} \tag{45}$$

then Eq. (43) gives us

$$\frac{d}{dt} \left(\nabla_\theta V(\theta) \right) = -\frac{2}{n} X^T r \tag{46}$$

and the integration relation

$$\nabla_\theta V(\theta) - \nabla_\theta V(\theta)|_{\theta=\theta(0)} = \sum_{i=1}^{n} x_i \int \tilde{r}_i(\tau) d\tau \tag{47}$$

where we let $\tilde{r} = -2|\xi|r/n$. Requiring $\nabla_\theta V(\theta)|_{\theta=\theta(0)} = 0$ and denoting $\lambda_i = \int_0^\infty \tilde{r}_i(\tau)d\tau$ gives us the condition at $t = \infty$:

$$\nabla_\theta V(\theta)|_{\theta=\theta(\infty)} = \sum_{i=1}^{n} x_i \lambda_i. \tag{48}$$

Eq. (48) coincides with the KKT stationary condition of the optimization problem (29). Therefore, we can prove the theorem by deriving the explicit form of $V(\theta)$.

Solving $V(\theta)$. According to Eq. (44), we can derive the following relation:

$$\partial_\theta V(\theta) = \hat{V}' \frac{\theta}{\|\theta\|} + h^T \tag{49}$$

$$\begin{aligned} \partial_\theta^2 V(\theta) &= \frac{1}{\|\theta\|^2} \left[\left(\hat{V}'' \frac{\theta\theta^T}{\|\theta\|} + \hat{V}' I \right) \|\theta\| - \hat{V}' \frac{\theta\theta^T}{\|\theta\|} \right] \\ &= \frac{\hat{V}'}{\|\theta\|} \left[I - \left(1 - \|\theta\| \frac{\hat{V}''}{\hat{V}'} \right) \frac{\theta\theta^T}{\|\theta\|^2} \right]. \end{aligned} \tag{50}$$

Comparing this with Eq. (45), we conclude that

$$1 - \|\theta\| \frac{\hat{V}''}{\hat{V}'} = \frac{1}{1 + \frac{1}{2(L-1)}} \implies \frac{\hat{V}''}{\hat{V}'} = \frac{1}{\|\theta\|} \left(1 - \frac{1}{1 + \frac{1}{2(L-1)}} \right), \tag{51}$$

which, by noting that $\partial_{\|\theta\|} \ln \hat{V}'(\|\theta\|) = \frac{\hat{V}''(\|\theta\|)}{\hat{V}'(\|\theta\|)}$, can be solved as follows

$$\begin{aligned} \frac{\hat{V}''(\|\theta\|)}{\hat{V}'(\|\theta\|)} &= \frac{1}{\|\theta\|} \frac{1}{2L-1} \\ \implies \ln \hat{V}'(\|\theta\|) &= \frac{1}{2L-1} \ln \|\theta\| \\ \implies \hat{V}(\|\theta\|) &= \frac{2L-1}{2L} \|\theta\|^{\frac{2L}{2L-1}}. \end{aligned} \tag{52}$$

Furthermore, since

$$\frac{\hat{V}'}{\|\theta\|} = \|\theta\|^{-\frac{2(L-1)}{2L-1}},$$

Eq. (45) is automatically satisfied when $\hat{V}(\|\theta\|)$ has the form of Eq. (51). It is now left for us to get the form of h, which can be done by noting that

$$\begin{aligned} \partial_\theta V(\theta(0)) &= 0 \\ \implies \|\theta(0)\|^{\frac{1}{2L-1}} \frac{\theta(0)}{\|\theta(0)\|} &+ h = 0. \end{aligned}$$

Thus the final form of $V(\theta)$ is

$$V(\theta) = \frac{1}{\Omega_L} \|\theta\|^{\Omega_L} - \theta^T \frac{\theta(0)}{\|\theta(0)\|^{\lambda_L}}. \tag{53}$$

This completes the proof. □

C.2.1 Remove the assumption of convergence to the interpolation solution

The assumption that $X\theta(\infty) = y$ in Theorem 1 can be removed if the dimension of the span of X^T is larger than the number of samples n, i.e., when $\dim\left(\text{span}(X^T)\right) \geq n$.[1] This can be proved as follows.

[1] Since $\max(\dim(\text{span}(X^T))) = n$, this condition is in fact $\dim(\text{span}(X^T)) = n$.

Proposition 2. *For the over-parameterized regression of rank-1 linear networks Eq. (2) and the dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $n < d$, if*

$$dim\left(span(X^T)\right) \geq n,$$

then the gradient flow solution $\theta(\infty)$ satisfies that

$$X\theta(\infty) = y.$$

Proof. To prove this proposition, we study the dynamics of the loss function $\mathcal{L} = \frac{1}{n}\sum_i r_i^2$, where $r_i = \langle \theta, x_i \rangle - y_i$, that is given by

$$\begin{aligned}\frac{d\mathcal{L}}{dt} &= \frac{\partial \mathcal{L}}{\partial \theta}\frac{d\theta}{dt} \\ &= \frac{2}{n}r^T X\left[-\frac{2\xi^2}{n}\left(I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right)X^T r\right] \\ &= -\frac{4\xi^2}{n^2}\left[rX^T Xr + 2(L-1)\frac{r^T X\theta\theta^T X^T r}{\|\theta\|^2}\right] \\ &= -\frac{4\xi^2}{n^2}\left[\|X^T r\|^2 + 2(L-1)\frac{(\theta^T X^T r)^2}{\|\theta\|^2}\right] \\ &\leq 0,\end{aligned}$$

where we have the equality in the last line when $r = (0, \dots, 0)^T \in \mathbb{R}^d$, i.e., $X\theta = y$, or $X^T r = 0$, which is not possible since we have assumed that $\dim(\text{span}(X^T)) \geq n$. Therefore, $\mathcal{L}(\theta(t))$ keeps decreasing until $X\theta(t) = y$, i.e., until GD finds the interpolation solution. Noting that $\min_\theta \mathcal{L}(\theta) = 0$, we complete the proof. □

C.3 Proof of Theorem 2

In this section we prove Theorem 2. The techniques are similar to those in C.2, and we still need a time wrapping technique introduced in [3] to derive the form of $V_{std}(\theta)$, since the condition for applying the IMD approach is violated in this case. Recall that, for the standard linear networks $f(x; W) = w_L^T W_{L-1} \cdots W_1 x = \theta^T x$ where $W_k \in \mathbb{R}^{d_k \times d_{k+1}}$, our purpose is to find a potential function $V_{std}(\theta)$ such that the gradient flow solution $\theta(\infty)$ satisfies that

$$\theta(\infty) = \arg\min_\theta V_{std}(\theta), \quad s.t. X\theta = y. \tag{54}$$

Since we assume the balanced initialization (Definition 2), then according to [11, 6, 2], the norms of all layers grow at the same rate and are the same for any $t > 0$:

$$W_{k+1}^T(t)W_{k+1}(t) = W_k(t)W_k^T(t).$$

Furthermore, following the procedure of [2], we obtain that the dynamics of θ is

$$\frac{d\theta}{dt} = -\|\theta\|^{\frac{2(L-1)}{L}}\left[I + (L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right]X^T r. \tag{55}$$

We now present the proof.

Proof. Eq. (55) can be written as

$$\|\theta\|^{\frac{2(1-L)}{L}}\left[I + (L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right]^{-1}\frac{d\theta}{dt} = -X^T r,$$

where, according to Lemma 2, the inverse in above equation is given by

$$\begin{aligned}\left[I + (L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right]^{-1} &= I - \frac{1}{1 + \text{tr}\left(\frac{(L-1)\theta\theta^T}{\|\theta\|^2}\right)}(L-1)\frac{\theta\theta^T}{\|\theta\|^2} \\ &= I - \frac{L-1}{L}\frac{\theta\theta^T}{\|\theta\|^2}.\end{aligned} \tag{56}$$

As a result, we have

$$\|\theta\|^{\frac{2(1-L)}{L}}\left(I-\frac{L-1}{L}\frac{\theta\theta^T}{\|\theta\|^2}\right)\frac{d\theta}{dt}=-X^Tr. \tag{57}$$

In this case, we still assume that $V_{std}(\theta)$ has the following form

$$V_{std}(\theta)=\hat{V}_{std}(\|\theta\|)+\beta^T\theta \tag{58}$$

for some constant vector $\beta\in\mathbb{R}^d$. Then following a similar procedure as in C.2, we have that

$$\partial_\theta V_{std}(\theta)=\hat{V}'_{std}\frac{\theta}{\|\theta\|}+h^T \tag{59}$$

$$\partial_\theta^2 V_{std}(\theta)=\frac{\hat{V}'_{std}}{\|\theta\|}\left[I-\left(1-\|\theta\|\frac{\hat{V}''_{std}}{\hat{V}'_{std}}\right)\frac{\theta\theta^T}{\|\theta\|^2}\right]. \tag{60}$$

In the IMD approach, $V_{std}(\theta)$ should satisfy that

$$\frac{d}{dt}(\partial_\theta V_{std}(\theta))=-\frac{2}{n}X^Tr, \tag{61}$$

which requires that $\partial_\theta^2 V_{std}(\theta)=\|\theta\|^{\frac{2(L-1)}{L}}\left(I-\frac{L-1}{L}\frac{\theta\theta^T}{\|\theta\|^2}\right)\frac{d\theta}{dt}$. But this is not possible. Therefore, we multiply a time re-scale factor $g(\theta)$, as long as $g(\theta)$ is positive, to both sides of Eq. (57) and only require that $V_{std}(\theta)$ satisfies the above relation under the new time scale $\tau:\mathbb{R}\to\mathbb{R}$ such that $\tau'=g(\theta)$:

$$g(\theta)\|\theta\|^{\frac{2(1-L)}{L}}\left(I-\frac{L-1}{L}\frac{\theta\theta^T}{\|\theta\|^2}\right)\frac{d\theta}{dt}=-g(\theta)X^Tr. \tag{62}$$

Then the limit point at $t=\infty$ in Eq. (61) is also visited at the point $\tau=\int_0^\infty g(\theta(s))ds$ in Eq. (62). We now solve the explicit form of $V_{std}(\theta)$ that satisfies Eq. (62). By comparing Eq. (60) and the right hand side of Eq. (57), we obtain that the following relation should be satisfied:

$$\frac{\hat{V}'_{std}}{\|\theta\|}\left[I-\left(1-\|\theta\|\frac{\hat{V}''_{std}}{\hat{V}'_{std}}\right)\frac{\theta\theta^T}{\|\theta\|^2}\right]=g(\theta)\|\theta\|^{\frac{2(1-L)}{L}}\left(I-\frac{L-1}{L}\frac{\theta\theta^T}{\|\theta\|^2}\right). \tag{63}$$

This implies that we need:

- the terms in the bracket on both sides should match:

$$1-\|\theta\|\frac{\hat{V}''_{std}}{\hat{V}'_{std}}=\frac{L-1}{L}\implies\frac{1}{xL}=\frac{\hat{V}''_{std}}{\hat{V}'_{std}}$$

$$\implies\ln\hat{V}'_{std}=\frac{1}{L}\ln\|\theta\|+C\implies\hat{V}_{std}=\frac{C'L}{L+1}\|\theta\|^{\frac{1}{L}+1} \tag{64}$$

  for some constant C and C', where we can simply choose $C'=1$;

- the terms outside the bracket on both sides should also match:

$$\frac{\hat{V}'_{std}}{\|\theta\|}=g(\theta)\|\theta\|^{\frac{2(1-L)}{L}}\implies g(\theta)=\hat{V}'_{std}\|\theta\|^{\frac{2(L-1)}{L}-1}. \tag{65}$$

To obtain the form of the constant vector β, we note that $\partial_\theta V(\theta(0))=0$, which immediately gives us

$$\|\theta(0)\|^{\frac{1}{L}}\frac{\theta(0)}{\|\theta(0)\|}+\beta=0\implies\beta=-\theta(0)\|\theta(0)\|^{\frac{1}{L}-1}. \tag{66}$$

Combining all these terms, we have the final form of $V_{std}(\theta)$:

$$V_{std}(\theta)=\frac{L}{L+1}\|\theta\|^{\frac{1}{L}+1}-\theta(0)^T\theta\|\theta(0)\|^{\frac{1}{L}-1}. \tag{67}$$

As in C.2, denoting $\lambda_i=\int_0^\infty r_i(s)ds$ and noting that $\partial_\theta V_{std}(\theta(0))=0$ give us

$$\partial_\theta V_{std}=\sum_{i=1}^n x_i\lambda_i,$$

which is exactly the KKT stationary condition of the optimization problem (54). □

Convergence to the interpolation solution. Similar to C.2.1, we can also show the convergence to the interpolation solution when $\dim\big(\mathrm{span}(X^T)\big) \geq n$ by deriving the dynamics of $\mathcal{L}$:

$$\begin{aligned}\frac{d\mathcal{L}}{dt} &= \frac{\partial \mathcal{L}}{\partial \theta}\frac{d\theta}{dt} \\ &= -\frac{2\|\theta\|^{\frac{2(1-L)}{L}}}{n} r^T X \left[I + (L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right] X^T r \\ &= -\frac{2\|\theta\|^{\frac{2(1-L)}{L}}}{n}\left[\|r^T X\|^2 + (L-1)\left(\theta^T X^T r\right)^2\right].\end{aligned}$$

Since we assume that $\dim\big(\mathrm{span}(X^T)\big) \geq n$, $d\mathcal{L}/dt < 0$ until $X\theta(t) = y$, i.e., $\mathcal{L}$ keeps decreasing until GD finds the interpolation solution.

C.4 Proof of Proposition 1

In this section, we prove Proposition 1 by analyzing the gradient of model parameters. It is helpful to recall that for a matrix A, we use A_{ij} to denote its i-th row j-th column element. For weight matrices, e.g., W_k, we use $W_{k;ij}$ to denote its i-th row j-th column element.

Proof. For a standard linear network that has the initialization of Proposition 1 and L is an odd number, i.e., for an integer p

$$\begin{aligned}\forall k = 2p+1 \in \{1,\dots,L\} : W_{k;ij}(0) = 0 \text{ if } i \neq c_k, \text{ where } c_k \in \{1,\dots d_{k+1}\} \\ \forall k = 2p \in \{1,\dots,L\} : W_{k;ij}(0) = 0 \text{ if } j \neq c_k,\end{aligned}$$

note that the $L-1$-th layer satisfies that $W_{L-1;ij} = W_{L-1;ij}\delta_{jc_{L-1}}$ where $\delta_{jl} = 1$ if $j = l$ otherwise $\delta_{jl} = 0$, we can write the networks at $t = 0$ as

$$f(x;W) = \sum_i \sum_j w_{L;i} W_{L-1;ij}\delta_{jc_{L-1}} \left(W_{L-2}\cdots W_1 x\right)_j.$$

Then the gradient w.r.t W_{L-1} at $t = 0$ is

$$\left(\nabla_{W_{L-1}}\mathcal{L}(\theta)\right)_{ij} = \frac{2}{n}\sum_{\mu=1}^{n} r_\mu w_{L;i}(W_{L-2}\cdots W_1 x)_j \delta_{jc_{L-1}} \tag{68}$$

$$\implies \left(\nabla_{W_{L-1}}\mathcal{L}(\theta)\right)_{ij} = 0 \text{ if } j \neq c_{L-1}. \tag{69}$$

This means that the parameter $W_{L-1;ij}$ will be updated only when $j = c_{L-1}$. As a result, the initialization shape for W_{L-1} will be maintained, i.e., only the non-zero column of W_{L-1} at $t = 0$ will be updated and all other elements of W_{L-1} will be zero for any $t > 0$ since the corresponding gradients vanish. Similarly, for W_{L-1}, we can write the network at $t = 0$ as:

$$f(x;W) = \sum_j \sum_l (w_L^T W_{L-1})_j \delta_{jc_{L-1}} W_{L-2;jl}(W_{L-3}\cdots W_1 x)_l,$$

then

$$\left(\nabla_{W_{L-2}}\mathcal{L}(\theta)\right)_{ij} = 0 \text{ if } i \neq c_{L-1}$$

and the initialization shape for W_{L-2} will also be maintained. Following a similar procedure, we conclude that all the initialization shapes will be maintained for any $t > 0$.

We now consider the diagonal initialization for weight matrices, namely that

$$W_{k;ij} = 0 \text{ if } i \neq j, \quad \forall k \in \{1,\dots,L-1\}.$$

For any $k \in \{1,\dots,L-1\}$, the gradient w.r.t W_k at $t = 0$ is

$$(\nabla_{W_k}\mathcal{L}(\theta))_{ij} = \frac{2}{n}\sum_{\mu=1}^{n} r_\mu (w_L^T \cdots W_{k+1})_i (W_{k-1}\cdots W_1 x)_j,$$

where, clearly, there is not any constraint on which elements of $(w_L^T \cdots W_{k+1})^T \in \mathbb{R}^d$ and $(W_{k-1}\cdots W_1 x) \in \mathbb{R}^d$ are zeros if we do not require that many diagonal elements of the initialization matrices are zeros, which clearly violates the diagonal initialization requirements. Thus we can not conclude that

$$(\nabla_{W_k}\mathcal{L}(\theta))_{ij} = 0 \text{ if } i \neq j,$$

i.e., the off-diagonal elements of W_k will also be updated to be non-zeros. Thus the diagonal initialization of weight matrices can not be maintained. The conclusions for other layers can be easily derived by following similar arguments. □

D Proofs for Section 4

In this section, we present the technical details for Section 4. In particular, we introduce our modelling details of the SGD dynamics in D.1, derive the SDE of the model parameter θ in D.2, and, finally, prove Theorem 3 in D.3.

D.1 SDE modelling details

Similar to the case of GD, to derive the implicit bias of SGD for rank-1 linear networks, we need to first derive the dynamics of the overall model parameter θ. For this purpose, the structure of the noise of SGD is crucial. For convenience, we first discuss the basic SGD where only one data is randomly sampled at each step, while the generalization to batch-size SGD is straightforward.

Structure of the noise. We present the details for u_k, and the case for v_{k+1} is similar. Recall that the empirical loss is $\mathcal{L} = \sum_i \ell_i$ / n, η is the learning rate and that the network can be written as $f(x; u, v) = \rho_k\rho_{-k}v_{k+1}^T u_k v_1^T x$, we start with the SGD update equation for u_k where we let j_t denote the index of the sampled data at the t-th step:

$$\begin{aligned} u_k(t+1) &= u_k(t) - \eta\frac{\partial \ell_{j_t}}{\partial u_k} = u_k(t) - \eta\frac{\partial \mathcal{L}}{\partial u_k} + \eta\left(\frac{\partial \mathcal{L}}{\partial u_k} - \frac{\partial \ell_{j_t}}{\partial u_k}\right) \\ &= u_k(t) - \frac{2\eta}{n}v_1^T X^T r\rho_k\rho_{-k}v_{k+1} \\ &\quad + 2\eta\left(\frac{1}{n}\sum_i r_i\rho_k\rho_{-k}v_1^T x_i v_{k+1} - r_{j_t}\rho_k\rho_{-k}v_1^T x_{j_t}v_{k+1}\right) \\ &= u_k(t) - \frac{2}{n}v_1^T X^T r\rho_k\rho_{-k}v_{k+1} + 2\eta\rho_k\rho_{-k}v_{k+1}v_1^T X^T Z_{j_t} \end{aligned} \tag{70}$$

where we let $\vec{e_{j_t}}$ be the basis vector in $\mathbb{R}^n$ such that the j_t-th element is 1 while all other elements are 0 and

$$Z_{j_t} = \mathrm{E}_{j_t}[r_{j_t}\vec{e_{j_t}}] - r_{j_t}\vec{e_{j_t}}, \quad \mathrm{cov}[Z_{j_t}] \sim \frac{\mathcal{L}}{n}\mathbf{I}_n. \tag{71}$$

As a result of this, the noise of SGD for u_k is now

$$\Sigma(u_k(t)) = \frac{4\eta^2\mathcal{L}}{n}(\rho_k\rho_{-k})^2 v_1^T X^T X v_1 v_{k+1}v_{k+1}^T. \tag{72}$$

Continuous Modelling of SGD. The continuous modelling techniques for SGD have been widely applied in recent works [1, 9, 23, 22] to study the dynamics of SGD. In our setting, the continuous counterpart of SGD is established as follows. First, the discrete SGD updating equations Eq. (8) can be equivalently written as

$$\begin{aligned} u_k(t+1) &= u_k(t) - \eta\nabla_{u_k}\mathcal{L}(\theta) + \eta\left[\nabla_{u_k}\mathcal{L}(\theta) - \nabla_{u_k}\ell_{i(t)}(\theta)\right], \\ v_{k+1}(t+1) &= v_{k+1}(t) - \eta\nabla_{v_{k+1}}\mathcal{L}(\theta) + \eta\left[\nabla_{v_{k+1}}\mathcal{L}(\theta) - \nabla_{v_{k+1}}\ell_{i(t)}(\theta)\right] \end{aligned}$$

for $k \in \{1, \ldots, L-1\}$ where both $\nabla_{u_k}\mathcal{L}(\theta) - \nabla_{u_k}\ell_{i(t)}(\theta)$ and $\nabla_{v_{k+1}}\mathcal{L}(\theta) - \nabla_{v_{k+1}}\ell_{i(t)}(\theta)$ are zero-mean noises with covariance matrices in $\mathbb{R}^{d_k \times d_k}$

$$\begin{aligned} \Sigma(u_k) &= \frac{4\mathcal{L}(\rho_k\rho_{-k})^2}{n}v_1^T X^T X v_1 v_{k+1}v_{k+1}^T, \\ \Sigma(v_{k+1}) &= \frac{4\mathcal{L}(\rho_k\rho_{-k})^2}{n}v_1^T X^T X v_1 u_k u_k^T, \end{aligned}$$

with ρ_k and ρ_{-k} defined in Eq. (3). Second, as we identify the noise covariance, letting $\eta \to 0$ [2], then we obtain the continuous counterpart of the discrete SGD that is a set of stochastic differential equations (SDE):

$$du_k = -\frac{2}{n} v_1^T X^T r \rho_k \rho_{-k} v_{k+1} dt + 2\sqrt{\frac{\eta \mathcal{L}}{n}} (\rho_k \rho_{-k}) v_{k+1} v_1^T X^T d\mathcal{W}_t \tag{73}$$

$$dv_{k+1} = -\frac{2}{n} v_1^T X^T r \rho_k \rho_{-k} u_k dt + 2\sqrt{\frac{\eta \mathcal{L}}{n}} (\rho_k \rho_{-k}) u_k v_1^T X^T d\mathcal{W}_t \tag{74}$$

where we let $r = (f(x_1; u, v) - y_1, \ldots, f(x_n; u, v) - y_n)^T \in \mathbb{R}^n$ be the residuals and $\mathcal{W}_t$ is a standard Brownian motion in $\mathbb{R}^n$. Similarly, recalling the definition of $\xi = w_L^T W_{L-1} \cdots W_2 u_1$, the SDE of v_1 is

$$dv_1 = -\frac{2}{n} \xi X^T r dt + 2\sqrt{\frac{\eta \mathcal{L}}{n}} \xi X^T d\mathcal{W}_t. \tag{75}$$

Generalization to batch-SGD. When b (a positive constant) data points $\mathcal{B}_t$ are sampled in each iteration of SGD, i.e., batch-SGD, we can change the SGD update equation as follows (taking u_k as an example)

$$u_k(t+1) = u_k(t) - \frac{\eta}{b} \sum_{j_t \in \mathcal{B}_t} \frac{\partial \ell_{j_t}}{\partial u_k}.$$

This only changes the noise $\Sigma(u_k(t))$ to a batch version

$$\Sigma_b(u_k(t)) = \frac{1}{b} \Sigma(u_k(t)),$$

which only affects the noise part of the SDE of u_k and leads it to become a batch version SDE:

$$du_k = -\frac{2}{n} v_1^T X^T r \rho_k \rho_{-k} v_{k+1} dt + 2\sqrt{\frac{\eta \mathcal{L}}{nb}} (\rho_k \rho_{-k}) v_{k+1} v_1^T X^T d\mathcal{W}_t.$$

This is equivalent to re-scale the learning rate η to

$$\eta_b = \frac{\eta}{b},$$

and leaving other parts unchanged. Thus the generalization to batch-SGD is straightforward—simply replacing all η with η_b.

D.2 The SDE of θ

In this section, we carefully derive the continuous dynamics of SGD for the parameterization of our rank-1 linear networks. We first discuss the balanced initialization condition.

Balanced initialization. Similar to the case for GD (Definition 2), we also assume the balanced initialization Eq. (26) across layers. Although the dynamics of SGD is different with that of GD, it still applies the gradient to update the parameters at every step that will maintain the balanced property thus the dynamics of SGD will also maintain the balanced property, i.e.,

$$\frac{\langle v_{k+1}(t), u_k(t) \rangle^2}{\|v_{k+1}(t)\|^2 \|u_k(t)\|^2} = 1$$

and

$$\|v_{k+1}(t)\| = \|u_k(t)\| = \|v_1(t)\|$$

for $t > 0$ and $k \in \{1, \ldots, L-1\}$, during the training of rank-1 linear networks.

With the equations for u_k and v_{k+1}, we now derive the SDE of θ summarized in the following lemma.

[2] More details of this modelling technique can be found in [14].

Lemma 3. *For an L-layer rank-1 linear network Eq. (2), if we assume balanced initialization, then the stochastic gradient flow of θ is*

$$d\theta = -\frac{2\xi^2}{n}H(\theta)X^T r dt + 2\xi^2\sqrt{\frac{\eta\mathcal{L}}{n}}H(\theta)X^T d\mathcal{W}_t + \frac{8\eta\mathcal{L}(L-1)}{n\|\theta\|^{\frac{2}{2L-1}}}\left[I + \frac{2L-3}{2}\frac{\theta\theta^T}{\|\theta\|^2}\right]X^T X\theta dt$$

where $H(\theta) = I + 2(L-1)\theta\theta^T/\|\theta\|^2$.

Proof. According to the Ito's Lemma, we have

$$d\theta = d(\xi v_1) = d\xi v_1 + \xi dv_1 + d\xi dv_1. \tag{76}$$

Thus to obtain the SDE of θ, we need to analyze every term of the above equation. We first give $d\xi$.

The form of $d\xi$. Let $\omega_k = v_{k+1}^T u_k$ and $\psi_k = \|v_{k+1}\|^2 + \|u_k\|^2$, we first characterize the SDE of ω_k. According to the Ito's calculus, we obtain that

$$d\omega_k = d(v_{k+1}^T u_k) = \underbrace{u_k^T dv_{k+1} + v_{k+1}^T du_k}_{\clubsuit} + \underbrace{du_k^T dv_{k+1}}_{\diamondsuit}, \tag{77}$$

where, by applying Eq. (73) and Eq. (74) and noting that $(d\mathcal{W}_t)^2 = dt$,

$$\clubsuit = -\frac{2}{n}v_1^T X^T r\rho_k\rho_{-k}\psi_k dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}}\rho_k\rho_{-k}\psi_k v_1^T X^T d\mathcal{W}_t$$
$$\diamondsuit = \frac{4\eta\mathcal{L}}{n}(\rho_k\rho_{-k})^2\omega_k v_1^T X^T X v_1 dt.$$

Combining the above two terms gives us the SDE of ω_k

$$d\omega_k = \left[-\frac{2}{n}v_1^T X^T r\rho_k\rho_{-k}\psi_k + \frac{4\eta(\rho_k\rho_{-k})^2\mathcal{L}}{n}\omega_k v_1^T X^T X v_1\right]dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}}\rho_k\rho_{-k}\psi_k v_1^T X^T d\mathcal{W}_t.$$

Since $\xi = d(\prod_{k=1}^{L-1}\omega_k)$, its SDE can be done by repeatedly applying the Ito's Lemma and the SDE of ω_k:

$$\begin{aligned} d\xi &= d(\prod_{k=1}^{L-1}\omega_k) \\ &= \underbrace{\sum_{k=1}^{L-1}\frac{\xi}{\omega_k}d\omega_k}_{\spadesuit} + \frac{1}{2}\sum_{k',k=1,k\neq k'}^{L-1}\underbrace{\frac{\xi}{\omega_k\omega_{k'}}d\omega_k d\omega_{k'}}_{\heartsuit}. \end{aligned} \tag{78}$$

For convenience, we first define several helper notations:

$$\Phi_1 = \sum_{k=1}^{L-1}\phi_k = \sum_{k=1}^{L-1}\frac{\psi_k}{\omega_k^2},$$
$$\Phi_2 = \sum_{k=1}^{L-1}\frac{1}{\omega_k^2},$$
$$\Phi_3 = \frac{1}{2}\sum_{k,k'=1,k\neq k'}^{L-1}\frac{\psi_k\psi_{k'}}{\omega_k^2\omega_{k'}^2}.$$

Now plugging the form of $d\omega_k$ into $\spadesuit$ gives us the first term of $d\xi$:

$$\begin{aligned}\spadesuit &= -\frac{2}{n} v_1^T X^T r \xi^2 \left(\sum_{k=1}^{L-1} \frac{\psi_k}{\omega_k^2}\right) dt + \frac{4\eta\mathcal{L}\xi^3 v_1^T X^T X v_1}{n} \left(\sum_{k=1}^{L-1} \frac{1}{\omega_k^2}\right) dt \\ &\quad + 2\sqrt{\frac{\eta\mathcal{L}}{n}} \xi^2 \left(\sum_{k=1}^{L-1} \frac{\psi_k}{\omega_k^2}\right) \left(v_1^T X^T d\mathcal{W}_t\right) \\ &= -\frac{2}{n} v_1^T X^T r \xi^2 \Phi_1 dt + \frac{4\eta\mathcal{L}\xi^3 v_1^T X^T X v_1}{n} \Phi_2 dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}} \xi^2 \Phi_1 \left(v_1^T X^T d\mathcal{W}_t\right), \end{aligned} \tag{79}$$

and applying again $(d\mathcal{W}_t)^2 = dt$ and the form of $d\omega_k$ gives us each term of the second sum of $d\xi$:

$$\heartsuit = \frac{\xi}{\omega_k \omega_{k'}} \frac{4\eta\xi^2\mathcal{L}}{n\omega_k\omega_{k'}} \psi_k \psi_{k'} v_1^T X^T X v_1 dt.$$

Summing all $\heartsuit$ and $\spadesuit$, we obtain the SDE of ξ:

$$\begin{aligned} d\xi &= \left[-\frac{2}{n} v_1^T X^T r \xi^2 \Phi_1 + \frac{4\eta\mathcal{L}\xi^3 v_1^T X^T X v_1}{n} \Phi_2 + \frac{4\eta\xi^3\mathcal{L} v_1^T X^T X v_1}{n} \Phi_3\right] dt \\ &\quad + 2\sqrt{\frac{\eta\mathcal{L}}{n}} \xi^2 \Phi_1 \left(v_1^T X^T d\mathcal{W}_t\right) \\ &= \left[-\frac{2}{n} v_1^T X^T r \xi^2 \Phi_1 + \frac{4\eta\mathcal{L}\xi^3 v_1^T X^T X v_1}{n} (\Phi_2 + \Phi_3)\right] dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}} \xi^2 \Phi_1 \left(v_1^T X^T d\mathcal{W}_t\right). \end{aligned} \tag{80}$$

The SDE of v_1 is much simpler. To get this, we start with the SGD update equation for v_1:

$$\begin{aligned} v_1(t+1) &= v_1(t) - \eta \frac{\partial \mathcal{L}}{\partial v_1} + \eta \left(\frac{\partial \mathcal{L}}{\partial v_1} - \frac{\partial \ell_{j_t}}{\partial v_1}\right) \\ &= v_1(t) - \frac{2\eta}{n} \xi X^T r + 2\eta \left(\frac{1}{n} \sum_i \xi x_i r_i - r_{j_t} \xi x_{j_t}\right) \\ &= v_1(t) - \frac{2\eta}{n} \xi X^T r + 2\eta \xi X^T Z_{j_t}, \end{aligned} \tag{81}$$

which implies that the noise covariance in this case is

$$\Sigma(v_1(t)) = \frac{4\eta^2\xi^2\mathcal{L}}{n} X^T X.$$

Then using a similar approach as that of u_k, we get the SDE of v_1

$$dv_1 = -\frac{2}{n} \xi X^T r dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}} \xi X^T d\mathcal{W}_t. \tag{82}$$

Now it is sufficient for us to derive the form of $d\theta$.

The form of $d\theta$. Combined with the SDE of ξ, we now have

$$d\theta = d(\xi v_1) = \underbrace{\xi dv_1 + v_1 d\xi}_{\clubsuit} + \underbrace{d\xi dv_1}_{\spadesuit}. \tag{83}$$

For the $\clubsuit$ term, as we already have the form of dv_1 in Eq. (82) and $d\xi$ in Eq. (80), we simply plug them into $\clubsuit$ and obtain that:

$$\begin{aligned} \clubsuit &= -\frac{2}{n} \xi^2 X^T r dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}} \xi^2 \left(X^T d\mathcal{W}_t\right) + v_1 \left[-\frac{2}{n} v_1^T X^T r \xi^2 \Phi_1 + \frac{4\eta L \xi^3 v_1^T X^T X v_1}{n} (\Phi_2 + \Phi_3)\right] dt \\ &\quad + 2\sqrt{\frac{\eta\mathcal{L}}{n}} \Phi_1 \xi^2 v_1 v_1^T X^T d\mathcal{W}_t \\ &= -\frac{2}{n} \left(\xi^2 I + \xi^2 \Phi_1 v_1 v_1^T\right) X^T r dt + \frac{4\eta\mathcal{L}\xi^3 v_1 v_1^T X^T X v_1}{n} (\Phi_2 + \Phi_3) dt \\ &\quad + 2\sqrt{\frac{\eta\mathcal{L}}{n}} \left(\xi^2 I + \xi^2 \Phi_1 v_1 v_1^T\right) X^T d\mathcal{W}_t. \end{aligned} \tag{84}$$

For the ♠ term, we only need to consider the $d\mathcal{W}_t$ terms of Eq. (82) and (80):

$$\spadesuit = 4\frac{\eta\mathcal{L}\xi^3}{n}\Phi_1 X^T X v_1 dt.$$

Combining the above two equations gives us the final SDE of $d\theta$:

$$d\theta = -\frac{2}{n}\left(\xi^2 I + \xi^2\Phi_1 v_1 v_1^T\right) X^T r dt + \frac{4\eta\mathcal{L}}{n}\left[\xi^3\Phi_1 I + \xi^3(\Phi_2 + \Phi_3)v_1 v_1^T\right] X^T X v_1 dt \\ + 2\sqrt{\frac{\eta\mathcal{L}}{n}}\left(\xi^2 I + \xi^2\Phi_1 v_1 v_1^T\right) X^T d\mathcal{W}_t. \tag{85}$$

On the other hand, recall that our assumptions regarding the initial conditions of $f(x; u, v)$ in the Lemma (Definition 1) and following similar techniques as in the case for GD, we have that for any $t > 0$:

1. $\forall k : \|u_k(t)\|^2 = \|v_k(t)\|^2 = \|v_1(t)\|^2$
2. $\forall k : \omega_k^2 = \langle u_k(t), v_{k+1}(t)\rangle^2 = \|u_k(t)\|^4 = \|v_1(t)\|^4$.

Plugging these terms back to the definitions of Φ_1, Φ_2 and Φ_3, we obtain that

$$\psi_k = 2\|v_1\|^2 \text{ and } \xi^2 = \|\theta\|^{\frac{4(L-1)}{2L-1}} \tag{86}$$

as in the case for GD and

$$\Phi_1 = \frac{2(L-1)}{\|v_1\|^2} = \frac{2(L-1)\xi^2}{\|\theta\|^2}, \tag{87}$$

$$\Phi_2 = \frac{L-1}{\|v_1\|^4} = \frac{\xi^4(L-1)}{\|\theta\|^4}, \tag{88}$$

$$\Phi_3 = \frac{2(L-1)(L-2)}{\|v_1\|^4} = \frac{2\xi^4(L-1)(L-2)}{\|\theta\|^4}, \tag{89}$$

where we use that $\|\theta\| = \|\xi v_1\| = |\xi|\|v_1\|$. Therefore, the final form of $d\theta$ is now

$$d\theta = -\frac{2\xi^2}{n}H(\theta)X^T r dt + 2\xi^2\sqrt{\frac{\eta\mathcal{L}}{n}}H(\theta)X^T d\mathcal{W}_t \\ + \frac{8\eta\mathcal{L}(L-1)}{n\|\theta\|^{\frac{2}{2L-1}}}\left[I + \frac{2L-3}{2}\frac{\theta\theta^T}{\|\theta\|^2}\right] X^T X\theta dt$$

where

$$H(\theta) = I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}.$$

□

D.3 Proof for Theorem 3

In this section, we determine the form of $V^{\mathcal{S}}(\theta)$ such that θ follows a stochastic mirror flow

$$d\partial_\theta V^{\mathcal{S}}(\theta, t) = -\frac{\partial\mathcal{L}}{\partial\theta}dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}}(X^T d\mathcal{W}_t), \tag{90}$$

which then proves the claims of Theorem 3. For this purpose, we start with manipulating the SDE of θ derived in Lemma 3. Note that Eq. (85) can be written as

$$\left(\xi^2 I + \xi^2\Phi_1 v_1 v_1^T\right)^{-1} d\theta - \frac{4\eta\mathcal{L}}{n}\mathcal{P}dt = -\frac{2}{n}X^T r dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}}X^T d\mathcal{W}_t, \tag{91}$$

where

$$\mathcal{P} = \left(\xi^2 I + \xi^2\Phi_1 v_1 v_1^T\right)^{-1}\left[\xi^3\Phi_1 I + \xi^3(\Phi_2 + \Phi_3)v_1 v_1^T\right] X^T X v_1.$$

To solve the inverse appeared in the above equation, we apply Lemma 2 and noting that

$$\operatorname{tr}\left(\xi^2\Phi_1 v_1 v_1^T \xi^{-2} I\right) = \Phi_1\|v_1\|^2,$$

then

$$\begin{aligned}
\left(\xi^2 I + \xi^2\Phi_1 v_1 v_1^T\right)^{-1} &= \frac{1}{\xi^2} I - \frac{1}{1+\Phi_1\|v_1\|^2}\frac{\xi^2\Phi_1 v_1 v_1^T}{\xi^4} \\
&= \frac{1}{\xi^2}\left(I - \frac{\Phi_1 v_1 v_1^T}{1+\Phi_1\|v_1\|^2}\right) \\
&= \frac{1}{\xi^2}\left(I - \frac{1}{1+\frac{\xi^2}{\Phi_1\|\theta\|^2}}\frac{\theta\theta^T}{\|\theta\|^2}\right),
\end{aligned}$$

where we use that $\theta = \xi v_1$ in the last equality, which enables us to simplify $\mathcal{P}$:

$$\begin{aligned}
\mathcal{P} &= \left(\xi^2 I + \xi^2\Phi_1 v_1 v_1^T\right)^{-1}\left[\xi^3\Phi_1 I + \xi^3(\Phi_2+\Phi_3) v_1 v_1^T\right] X^T X v_1 \\
&= \Phi_1\left(I - \frac{1}{1+\frac{\xi^2}{\Phi_1\|\theta\|^2}}\frac{\theta\theta^T}{\|\theta\|^2}\right)\left[I + \frac{\Phi_2+\Phi_3}{\xi^2\Phi_1}\theta\theta^T\right] X^T X\theta \\
&= \Phi_1\left[I - \left(\frac{1}{\|\theta\|^2+\frac{\xi^2}{\Phi_1}} - \frac{\Phi_2+\Phi_3}{\xi^2\Phi_1} + \frac{\|\theta\|^2}{\|\theta\|^2+\frac{\xi^2}{\Phi_1}}\frac{\Phi_2+\Phi_3}{\xi^2\Phi_1}\right)\theta\theta^T\right] X^T X\theta \\
&= \Phi_1\left[I - \frac{\Phi_1^2-\Phi_2-\Phi_3}{\Phi_1^2\|\theta\|^2+\xi^2\Phi_1}\theta\theta^T\right] X^T X\theta.
\end{aligned}$$

Thus Eq. (91) for the overall SDE of θ now becomes

$$\frac{1}{\xi^2}\left(I - \frac{1}{1+\frac{\xi^2}{\Phi_1\|\theta\|^2}}\frac{\theta\theta^T}{\|\theta\|^2}\right)d\theta - \frac{4\eta\mathcal{L}}{n}\mathcal{P}dt = -\frac{\partial\mathcal{L}}{\partial\theta}dt + 2\sqrt{\frac{\eta\mathcal{L}}{n}}(X^T d\mathcal{W}_t). \tag{92}$$

The balanced initialization gives us Eq. (87), (88), and (89), and recall that

$$\lambda_L = \frac{2(L-1)}{2L-1},$$

thus we can further rewrite

$$\begin{aligned}
\frac{1}{\xi^2}\left(I - \frac{1}{1+\frac{\xi^2}{\Phi_1\|\theta\|^2}}\frac{\theta\theta^T}{\|\theta\|^2}\right) &= \frac{1}{\xi^2}\left(I - \frac{1}{1+\frac{1}{2(L-1)}}\frac{\theta\theta^T}{\|\theta\|^2}\right) \\
&= \frac{1}{\xi^2}\left(I - \lambda_L\frac{\theta\theta^T}{\|\theta\|^2}\right)
\end{aligned} \tag{93}$$

and

$$\begin{aligned}
\mathcal{P} &= \frac{1}{\xi^2}\left(I - \lambda_L\frac{\theta\theta^T}{\|\theta\|^2}\right)\left[\xi^2\Phi_1 I + (\Phi_2+\Phi_3)\theta\theta^T\right] X^T X\theta \\
&= \frac{1}{\xi^2}\left[\xi^2\Phi_1 I + \left((\Phi_2+\Phi_3)(1-\lambda_L) - \frac{\lambda_L\xi^2\Phi_1}{\|\theta\|^2}\right)\theta\theta^T\right] X^T X\theta \\
&= \frac{1}{\xi^2}\Big[\frac{2(L-1)\xi^4}{\|\theta\|^2} I + ((L-1+4(L-1)(L-2))(1-\lambda_L) - 2(L-1)\lambda_L)\frac{\xi^4}{\|\theta\|^4}\theta\theta^T\Big] X^T X\theta \\
&= \frac{\xi^2}{\|\theta\|^2}\left[2(L-1)I - \frac{3(L-1)}{2L-1}\frac{\theta\theta^T}{\|\theta\|^2}\right] X^T X\theta \\
&= \frac{2(L-1)\xi^2}{\|\theta\|^2}\left(I - \frac{\theta\theta^T}{2\|\theta\|^2}\right) X^T X\theta.
\end{aligned} \tag{94}$$

Thus, Eq. (91) now[3] can be rewritten as

$$\frac{1}{\xi^2}\left(I-\lambda_L\frac{\theta\theta^T}{\|\theta\|^2}\right)d\theta-\frac{8\eta\mathcal{L}}{n}\frac{(L-1)\xi^2}{\|\theta\|^2}\left(I-\frac{1}{2}\frac{\theta\theta^T}{\|\theta\|^2}\right)X^TX\theta dt$$
$$=-\frac{\partial\mathcal{L}}{\partial\theta}dt+2\sqrt{\frac{\eta\mathcal{L}}{n}}(X^Td\mathcal{W}_t). \tag{95}$$

Finding the form of $V^{\mathcal{S}}(\theta)$. We now proceed to find the form of $V^{\mathcal{S}}(\theta)$. For convenience, we apply a time re-scaling technique such that $d\tau=|\xi|dt$ ($d\mathcal{W}_\tau=\sqrt{|\xi|}d\mathcal{W}_t$ according to [12]) to the above equation. For convenience, we still use t to represent the time after re-scaling. Then the above equation becomes

$$\frac{1}{\|\theta\|^{\frac{2(L-1)}{2L-1}}}\left(I-\lambda_L\frac{\theta\theta^T}{\|\theta\|^2}\right)d\theta-\frac{8\eta\mathcal{L}}{n}\frac{(L-1)\xi^2}{\|\theta\|^2}\left(I-\frac{1}{2}\frac{\theta\theta^T}{\|\theta\|^2}\right)X^TX\theta dt$$
$$=-\frac{\partial\mathcal{L}}{\partial\theta}dt+2\sqrt{\frac{\eta|\xi|\mathcal{L}}{n}}(X^Td\mathcal{W}_t). \tag{96}$$

Recall that $\theta(0)\in\mathbb{R}^d$ is the initialization of θ, since $V^{\mathcal{S}}(\theta)$ for SGD should have similar form as that for GD, we borrow from the GD results and first define a constant vector $\gamma\in\mathbb{R}^d$

$$\gamma=-\|\theta(0)\|^{-\frac{2L-2}{2L-1}}\theta(0).$$

Suppose now that $V^{\mathcal{S}}$ has the following form:

$$V^{\mathcal{S}}(\theta,t)=\frac{2L-1}{2L}\|\theta\|^{\frac{2L}{2L-1}}+\gamma^T\theta+g(t)^T\theta\text{ for }g\in\mathbb{R}^d, \tag{97}$$

by similar techniques as in the case of GD, we obtain the first and second derivatives of $V^{\mathcal{S}}(\theta)$ w.r.t θ:

$$\partial_\theta V^{\mathcal{S}}(\theta,t)=\|\theta\|^{-\frac{2(L-1)}{2L-1}}\theta+\gamma+g(t),$$
$$\partial_\theta^2 V^{\mathcal{S}}(\theta,t)=\frac{1}{\|\theta\|^{\frac{2(L-1)}{2L-1}}}\left[I-\lambda_L\frac{\theta\theta^T}{\|\theta\|^2}\right].$$

There exists a function of θ, $G(\theta)$ whose exact form can be derived but not necessary for us, corresponding to our $V^{\mathcal{S}}(\theta,t)$ defined above, such that $d\partial_\theta V^{\mathcal{S}}(\theta,t)$ can be written as

$$d\partial_\theta V^{\mathcal{S}}(\theta)=\partial_\theta\left(\partial_\theta V^{\mathcal{S}}(\theta,t)\right)d\theta+\partial_t\left(\partial_\theta V^{\mathcal{S}}(\theta,t)\right)dt+\frac{\partial^2}{\partial\theta\partial\theta}\partial_\theta V^{\mathcal{S}}(\theta,t)(d\theta)^2$$
$$=\partial_\theta\left(\partial_\theta V^{\mathcal{S}}(\theta,t)\right)d\theta+\left[\partial_t\left(\partial_\theta V^{\mathcal{S}}(\theta,t)\right)+G(\theta)\frac{\partial^2}{\partial\theta\partial\theta}\partial_\theta V^{\mathcal{S}}(\theta,t)\right]dt$$

Now suppose that we can choose a particular $g(t)$, thus a particular $G(\theta)$, such that the following relation is satisfied:

$$\frac{\partial}{\partial t}\partial_\theta V^{\mathcal{S}}(\theta,t)+G(\theta)\frac{\partial^2}{\partial\theta\partial\theta}\partial_\theta V^{\mathcal{S}}(\theta,t)$$
$$=-\frac{8\eta\mathcal{L}}{n}\frac{(L-1)\xi^2}{\|\theta\|^2}\left(I-\frac{1}{2}\frac{\theta\theta^T}{\|\theta\|^2}\right)X^TX\theta, \tag{98}$$

then we can rewrite the SDE of $\partial_\theta V^{\mathcal{S}}(\theta,t)$ as follows:

$$d\partial_\theta V^{\mathcal{S}}(\theta)=\partial_\theta\left(\partial_\theta V^{\mathcal{S}}(\theta,t)\right)d\theta+\partial_t\left(\partial_\theta V^{\mathcal{S}}(\theta,t)\right)dt+G(\theta)\frac{\partial^2}{\partial\theta\partial\theta}\partial_\theta V^{\mathcal{S}}(\theta,t)dt$$
$$=\frac{1}{\|\theta\|^{\frac{2(L-1)}{2L-1}}}\left(I-\lambda_L\frac{\theta\theta^T}{\|\theta\|^2}\right)d\theta-\frac{8\eta\mathcal{L}}{n}\frac{(L-1)\xi^2}{\|\theta\|^2}\left(I-\frac{1}{2}\frac{\theta\theta^T}{\|\theta\|^2}\right)X^TX\theta dt \tag{99}$$

[3] Note that when $L=1$ we do not have the second dt term in the LHS of the equation, this term is brought by adding layers to the model.

which will give us the desired “mirror flow” equation:

$$d\partial_\theta V^{\mathcal{S}}(\theta) = -\nabla_\theta \mathcal{L} dt + 2\sqrt{\frac{\eta|\xi|\mathcal{L}}{n}} X^T d\mathcal{W}_t.$$

It is now suffice for us to find the particular $g(t)$ that makes Eq. (98) satisfied. For convenience, recall that for a vector a we use a_μ to denote its μ-th component, we define the following helper notations:

$$k = 2\xi^2\sqrt{\frac{\eta\mathcal{L}}{n}} \in \mathbb{R}, \tag{100}$$

$$B = \left[I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right] X^T \in \mathbb{R}^{d\times n}, \tag{101}$$

$$d\theta_\mu = O_\mu dt + \sum_i B_{\mu i} d\mathcal{W}_{t,i}, \tag{102}$$

$$D_\mu(\theta, t) = \left(\frac{\partial V^{\mathcal{S}}(\theta,t)}{\partial\theta}\right)_\mu = \|\theta\|^{-\lambda_L}\theta_\mu + \gamma_\mu + g_\mu(t), \quad D(\theta,t) \in \mathbb{R}^d, \tag{103}$$

$$\delta_{\rho\sigma} = 1 \text{ if } \rho = \sigma \text{ otherwise } \delta_{\rho\sigma} = 0, \tag{104}$$

where the exact form of O can be obtained from Lemma 3 but is not necessary for us. In the following, we use D to represent $D(\theta, t)$ for convenience. For our purpose of choosing the particular $g(t)$, $dD(\theta,t) = d\partial_\theta V^{\mathcal{S}}(\theta,t)$ should match the R.H.S of Eq. (99). To make this relation clear, according to the Ito’s Lemma, we obtain that

$$dD = \partial_t D dt + \partial_\theta D d\theta + \underbrace{\frac{1}{2}\sum_{\rho,\sigma}\frac{\partial^2 D}{\partial\theta_\rho\partial\theta_\sigma} d\theta_\rho d\theta_\sigma}_{\spadesuit},$$

where the last term is crucial. From the SDE of θ (Lemma 3), we have that the $d\theta_\rho d\theta_\sigma$ appeared in $\spadesuit$ can be written as

$$d\theta_\rho d\theta_\sigma = k^2 \sum_{j,i} B_{j\sigma}B_{\rho i} d\mathcal{W}_{t,i} d\mathcal{W}_{t,j} = k^2 \sum_i B_{i\sigma}B_{\rho i} dt.$$

On the other hand, according to

$$\partial_\theta D = \partial^2_\theta V^{\mathcal{S}}(\theta,t) = \frac{1}{\|\theta\|^{\lambda_L}}\left(I - \lambda_L \frac{\theta\theta^T}{\|\theta\|^2}\right),$$

we have that, with explicit subscripts of vectors,

$$\begin{aligned}
\frac{\partial^2 D_\mu}{\partial\theta_\rho\partial\theta_\sigma} &= \frac{\partial}{\partial\theta_\sigma}\left[\|\theta\|^{-\lambda_L}\left(\delta_{\mu\rho} - \lambda_L\frac{\theta_\rho\theta_\mu}{\|\theta\|^2}\right)\right] \\
&= -\lambda_L\|\theta\|^{-\lambda_L-2}\delta_{\mu\rho}\theta_\sigma - \lambda_L\frac{\partial}{\partial\theta_\sigma}\left(\|\theta\|^{-\lambda_L-2}\theta_\rho\theta_\mu\right) \\
&= -\lambda_L\|\theta\|^{-\lambda_L-2}\delta_{\mu\rho}\theta_\sigma \\
&\qquad - \lambda_L\left[-(\lambda_L+2)\|\theta\|^{-\lambda_L-4}\theta_\sigma\theta_\mu\theta_\rho + \|\theta\|^{-\lambda_L-2}(\delta_{\rho\sigma}\theta_\mu + \delta_{\mu\sigma}\theta_\rho)\right] \\
&= -\frac{\lambda_L}{\|\theta\|^{\lambda_L+2}}(\delta_{\mu\rho}\theta_\sigma + \delta_{\mu\sigma}\theta_\rho) - \frac{\lambda_L}{\|\theta\|^{\lambda_L+2}}\theta_\mu\left(\delta_{\rho\sigma} - (\lambda_L+2)\frac{\theta_\rho\theta_\sigma}{\|\theta\|^2}\right).
\end{aligned} \tag{105}$$

Note that the above expression implies that $\frac{\partial^2 D}{\partial\theta\partial\theta}$ is in fact a rank-3 tensor. Combined with the expression of $d\theta_\rho d\theta_\sigma$ derived above, we have that

$$\spadesuit = -\frac{k^2\lambda_L}{2\|\theta\|^{\lambda_L+2}}\sum_{\rho,\sigma}\sum_i\left[\underbrace{\delta_{\mu\rho}\theta_\sigma + \delta_{\mu\sigma}\theta_\rho}_{H^\rho_{\mu\sigma}} + \theta^\mu\underbrace{\left(\delta_{\rho\sigma} - (\lambda_L+2)\frac{\theta_\rho\theta_\sigma}{\|\theta\|^2}\right)}_{P_{\rho\sigma}}\right] B_{\sigma i}B_{i\rho}dt \tag{106}$$

where $H \in \mathbb{R}^{d\times d\times d}$ is a rank-3 tensor. To get the exact form of $\spadesuit$, we need to have the exact forms of

1. $\sum_{\rho,\sigma,i} H^{\rho}_{\mu\sigma} B_{\sigma i} B_{i\rho}$. Note that $HBB \in \mathbb{R}^d$ and there are two different terms, both of which will induce a same vector. In particular,

$$\sum_{\rho,\sigma,i} \delta_{\rho\mu}\theta_\sigma B_{\sigma i}B_{i\rho} = \sum_{\rho\sigma,i} B_{\mu i}B_{i\sigma}\theta_\sigma = (BB^T\theta)_\mu$$
$$\sum_{\rho,\sigma,i} \delta_{\mu\sigma}\theta_\rho B_{\sigma i}B_{i\rho} = ((\theta^T BB^T)^T)_\mu = (BB^T\theta)_\mu.$$

Thus we have

$$\sum_{\rho,\sigma,i} H^{\rho}_{\mu\sigma} B_{\sigma i} B_{i\rho} = 2(B^T B\theta)_\mu,$$

where

$$\begin{aligned}
BB^T\theta &= \left(I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right) X^T X \left(I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right)\theta \\
&= \left(I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right) X^T X \left(\theta + 2(L-1)\theta\right) \\
&= (2L-1)\left[I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right] X^T X\theta. \qquad (107)
\end{aligned}$$

2. $\sum_{\rho,\sigma,i} P^{\sigma}_{\rho} B^{i}_{\sigma} B^{\rho}_{i}$. Using the matrix notation, we can easily find that

$$\sum_{\rho,\sigma,i} P^{\sigma}_{\rho} B^{i}_{\sigma} B^{\rho}_{i} = \text{tr}\left(B^T P B\right),$$

where, recall the definition of B in Eq. (101),

$$\begin{aligned}
B^T P B &= X\left(I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right)\left(I - (\lambda_L + 2)\frac{\theta\theta^T}{\|\theta\|^2}\right)\left(I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right) X^T \\
&= X\left[I + \left(2(L-1) - (\lambda_L + 2) - 2(L-1)(\lambda_L + 2)\right)\frac{\theta\theta^T}{\|\theta\|^2}\right] \\
&\quad \times \left[I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right] X^T \\
&= X\left[I - 2(2L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right]\left[I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right] X^T \\
&= X\left[I - (8L^2 - 10L + 4)\frac{\theta\theta^T}{\|\theta\|^2}\right] X^T. \qquad (108)
\end{aligned}$$

Taking the trace of the above equation gives us the second term of $\spadesuit$:

$$\text{tr}\left(B^T P B\right) = \text{tr}\left(X^T X\right) - (8L^2 - 10L + 4)\frac{\theta^T}{\|\theta\|^2} X^T X\theta.$$

Now removing the μ subscript of the derived $\sum_{\rho,\sigma,i} H^{\rho}_{\sigma,\mu} B_{\sigma i} B_{i\rho}$ and recovering the matrix notation, we have the final form of $\spadesuit$ by combining it with the result of tr $\left(B^T P B\right)$

$$\begin{aligned}
\spadesuit &= -\frac{k^2\lambda_L}{2\|\theta\|^{\lambda_L+2}}\left[2(2L-1)\left(I + 2(L-1)\frac{\theta\theta^T}{\|\theta\|^2}\right) X^T X\theta\right] \\
&\quad - \frac{k^2\lambda_L}{2\|\theta\|^{\lambda_L+2}}\left[\theta \text{tr}\left(X^T X\right) - (8L^2 - 10L + 4)\frac{\theta\theta^T}{\|\theta\|^2} X^T X\theta\right] \\
&= \frac{k^2}{2}\left[-\frac{\lambda_L}{\|\theta\|^{\lambda_L+2}} \text{tr}\left(X^T X\right)\theta - \frac{4(L-1)}{\|\theta\|^{\lambda_L+2}}\left(I - \frac{L}{2L-1}\frac{\theta\theta^T}{\|\theta^2\|}\right) X^T X\theta\right] dt. \qquad (109)
\end{aligned}$$

Thus the SDE of D now becomes

$$\begin{aligned}
dD &= \partial_\theta D d\theta \\
&\quad + \left[\partial_t D - \frac{k^2\lambda_L}{2\|\theta\|^{\lambda_L+2}} \text{tr}\left(X^T X\right)\theta - \frac{2k^2(L-1)}{\|\theta\|^{\lambda_L+2}}\left(I - \frac{L}{2L-1}\frac{\theta\theta^T}{\|\theta^2\|}\right) X^T X\theta\right] dt
\end{aligned}$$

To find the particular $g(t)$ such that Eq. (98) is satisfied, we need to require that

$$\partial_t D - \frac{k^2 \lambda_L}{2\|\theta\|^{\lambda_L+2}} \mathrm{tr}\left(X^T X\right)\theta - \frac{2k^2(L-1)}{\|\theta\|^{\lambda_L+2}}\left(I - \frac{L}{2L-1}\frac{\theta\theta^T}{\|\theta^2\|}\right) X^T X\theta$$
$$= -\frac{8\eta\mathcal{L}}{n}\frac{(L-1)\xi^2}{\|\theta\|^2}\left(I - \frac{1}{2}\frac{\theta\theta^T}{\|\theta\|^2}\right) X^T X\theta.$$

Recall that $D = \|\theta\|^{-\lambda_L}\theta + \gamma + g(t)$, we have

$$\partial_t D = g'(t),$$

which, noting that $k^2 = 4\xi^4\eta\mathcal{L}/n$, further gives us that the following relation should be satisfied

$$g'(t) - \frac{2\xi^4\eta\mathcal{L}}{n\|\theta\|^{\lambda_L+2}}\left[\lambda_L \mathrm{tr}\left(X^T X\right)\theta + 4(L-1)\left(I - \frac{L}{2L-1}\frac{\theta\theta^T}{\|\theta^2\|}\right) X^T X\theta\right]$$
$$= -\frac{8\eta\mathcal{L}}{n}\frac{(L-1)\xi^2}{\|\theta\|^2}\left(I - \frac{1}{2}\frac{\theta\theta^T}{\|\theta\|^2}\right) X^T X\theta.$$

As a result of this, we can, noting that $\xi^2 = \|\theta\|^{2\lambda_L}$ from $\|\theta\|^2 = \xi^2\|v_1\|^2$, give the required $g(t)$ that makes Eq. (98) satisfied by solving the following equation:

$$\begin{aligned} g'(t) &= \frac{2\lambda_L\eta \mathrm{tr}\left(X^T X\right)\mathcal{L}\|\theta\|^{\lambda_L-2}}{n}\theta \\ &\quad + \frac{8\eta(L-1)\mathcal{L}\|\theta\|^{\lambda_L-2}}{n}\left[I - I - \left(\frac{L}{2L-1} - \frac{1}{2}\right)\right] X^T X\theta \\ &= \frac{2\lambda_L\eta\mathcal{L}\|\theta\|^{\lambda_L-2}}{n}\mathrm{tr}\left(\left(I - \frac{\theta\theta^T}{\|\theta\|^2}\right) X^T X\right)\theta. \end{aligned} \tag{110}$$

Now let the orthogonal projection operator of θ be $P_\perp(\theta) = I - \frac{\theta\theta^T}{\|\theta\|^2}$, then we can solve $g(t)$ as

$$g(t) = \frac{2\lambda_L\eta}{n}\int_0^t \mathcal{L}(\theta)\|\theta\|^{\lambda_L-2}\theta \mathrm{tr}\left(P_\perp(\theta)X^T X\right) ds. \tag{111}$$

With this particular $g(t)$, we can then give $V^{\mathcal{S}}(\theta, t)$

$$V^{\mathcal{S}}(\theta, t) = \frac{1}{\Omega_L}\|\theta\|^{\Omega_L} + \theta^T\left[\frac{2\lambda_L\eta}{n}\int_0^t \mathcal{L}(\theta)\|\theta\|^{\lambda_L-2}\theta \mathrm{tr}\left(P_\perp(\theta)X^T X\right) ds - \frac{\theta(0)}{\|\theta(0)\|^{\lambda_L}}\right] \tag{112}$$

that satisfies the relation

$$d\partial_\theta V^{\mathcal{S}}(\theta) = -\nabla_\theta\mathcal{L}dt + 2\sqrt{\frac{\eta\mathcal{L}\xi}{n}}X^T d\mathcal{W}_t.$$

Moreover, a particular interesting case is that, as $L \to \infty$,

$$\lim_{L\to\infty} V(\theta, t) = \|\theta\| + \theta^T\left[\frac{2\eta}{n}\int_0^T \mathcal{L}(\theta)\left\|P_\perp(\theta)X^T\right\|_F^2 \frac{\theta}{\|\theta\|}ds - \frac{\theta(0)}{\|\theta(0)\|}\right]. \tag{113}$$