OPT2020: 12th Annual Workshop on Optimization for Machine Learning

Direction Matters: On the Implicit Regularization Effect of Stochastic Gradient Descent with Moderate Learning Rate

Jingfeng Wu UUUJF@JHU.EDU
Johns Hopkins University, USA

Difan Zou KNOWZOU@CS.UCLA.EDU
University of California, Los Angeles, USA

Vladimir Braverman VOVA@CS.JHU.EDU
Johns Hopkins University, USA

Quanquan Gu QGU@CS.UCLA.EDU
University of California, Los Angeles, USA

Abstract

Understanding the algorithmic regularization effect of *stochastic gradient descent* (SGD) is one of the key challenges in modern machine learning and deep learning theory. Most of the existing works, however, focus on *very small or even infinitesimal* learning rate regime, and fail to cover practical scenarios where the learning rate is *moderate and annealing*. In this paper, we make an initial attempt to characterize the particular regularization effect of SGD in the moderate learning rate regime by studying its behavior for optimizing an overparameterized linear regression problem. In this case, SGD and GD are known to converge to the unique minimum-norm solution; however, with the moderate and annealing learning rate, we show that they exhibit different *directional bias*: SGD converges along the large eigenvalue directions of the data matrix, while GD goes after the small eigenvalue directions. Furthermore, we show that such directional bias does matter when early stopping is adopted, where the SGD output achieves nearly optimal estimation error but the GD output is only suboptimal.

1. Introduction

Stochastic gradient descent (SGD) and its variants play a key role in training deep learning models. From the optimization perspective, SGD is favorable in many aspects, e.g., scalability for large-scale models [13], parallelizability with big training data [9], and rich theory for its convergence [7, 8]. From the learning perspective, more surprisingly, overparameterized deep nets trained by SGD usually generalize well, even in the absence of explicit regularizers [20, 34, 35]. This suggests that SGD favors certain "good" solutions among the numerous global optima of the overparameterized model. Such phenomenon is attributed to the *implicit regularization effect* of SGD. It remains one of the key theoretical challenges to characterize the algorithmic bias of SGD, especially with moderate and annealing learning rate as typically used in practice [13, 20].

In the *small learning rate regime*, the regularization effect of SGD is relatively well understood, thanks to the recent advances on the implicit bias of *gradient descent* (GD) [1, 3, 6, 10–12, 17–19, 23, 25–27, 31]. According to classical stochastic approximation theory [21], with a sufficiently small learning rate, the randomness in SGD is negligible (which scales with learning rate), and as a consequence SGD will behave highly similar to its deterministic counterpart, i.e., GD. Based on this

fact, the regularization effect of SGD with small learning rate can be understood through that of GD. Take linear models for example, GD has been shown to be biased towards max-margin/minimum-norm solutions depending on the problem setups [1, 11, 31]; correspondingly, follow-ups show that SGD with small learning rate has the same bias (up to certain small uncertainty governed by the learning rate) [2, 11, 28].

However, the regularization theory for SGD with small learning rate cannot explain the benefits of SGD in the *moderate learning rate regime*, where the initial learning rate is moderate and followed by annealing [16, 22, 24, 29]. In particular, empirical studies show that, in the moderate learning rate regime, (small batch) SGD generalizes much better than GD/large batch SGD [15, 20, 34, 36]. This observation implies that, instead of imitating the bias of GD as in the small learning rate regime, SGD in the moderate learning rate regime admits superior bias than GD — it requires a dedicated characterization for the implicit regularization effect of SGD with moderate learning rate.

In this paper, we reveal a particular regularization effect of SGD with moderate learning rate that involves *convergence direction*. In specific, we consider an overparameterized linear regression model learned by SGD/GD. In this setting, SGD and GD are known to converge to the unique minimum-norm solution [11] (see also Section 2.1). However, with a moderate and annealing learning rate, we show that SGD and GD favor different convergence directions: *SGD converges along the large eigenvalue directions of the data matrix*; in contrast, GD goes after the small eigenvalue directions. Furthermore, we show the particular directional bias of SGD with moderate learning rate benefits generalization when early stopping is used. This is because converging along the large eigenvalue directions (SGD) leads to solutions with nearly optimal estimation error, while converging along the small eigenvalue directions (GD) can only give suboptimal solutions. To our knowledge, these results initiate the regularization theory for SGD in the moderate learning rate regime, and complement existing results for the small learning rate.

2. Preliminary

Let $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ be a pair of d-dimensional feature vector and 1-dimensional label. We consider a linear regression problem with square loss defined as $\ell(x, y; w) := (w^\top x - y)^2$, where $w \in \mathbb{R}^d$ is the model parameter. Let $\mathcal{D}$ be the population distribution over (x, y), then the test loss is $L_{\mathcal{D}}(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(x, y; w)\right]$. Let $\mathcal{S} := \{(x_i, y_i)\}_{i=1}^n$ be a training set of n data points drawn i.i.d. from the population distribution $\mathcal{D}$. Then the training/empirical loss is defined as the average of the individual loss over all training data points, $L_{\mathcal{S}}(w) := \frac{1}{n}\sum_{i=1}^n \ell_i(w)$, where $\ell_i(w) := \ell(x_i, y_i; w) = (w^\top x_i - y_i)^2$. We use $\{\eta_k\}$ to denote a *learning rate scheme* (LR). Then *gradient descent* (GD) iteratively performs the following update:

$$w_{k+1} = w_k - \eta_k \nabla L_{\mathcal{S}}(w_k) = w_k - \frac{2\eta_k}{n}\sum_{i=1}^n x_i(x_i^\top w_k - y_i). \tag{GD}$$

Next we introduce *mini-batch stochastic gradient descent* (SGD). Let b be the batch size. For simplicity suppose $n = mb$ for an integer m (number of mini-batches). Then at each epoch, SGD first randomly partitions the training set into m disjoint mini-batches with size b, and then sequentially performs m updates using the stochastic gradients calculated over the m mini-batches. Specifically, at the k-th epoch, let the mini-batch index sets be $\mathcal{B}_1^k, \mathcal{B}_2^k, \ldots, \mathcal{B}_m^k$, where $|\mathcal{B}_j^k| = b$ and

$\bigcup_{j=1}^{m} \mathcal{B}_j^k = \{1, 2, \ldots, n\}$, then SGD takes m updates as follows

$$w_{k,j+1} = w_{k,j} - \frac{\eta_k}{b} \sum_{i \in \mathcal{B}_j^k} \nabla \ell_i(w_{k,j}) = w_{k,j} - \frac{2\eta_k}{b} \sum_{i \in \mathcal{B}_j^k} x_i(x_i^\top w_{k,j} - y_i), \quad j = 1, \ldots, m. \tag{SGD}$$

We also write $w_{k+1} = w_{k,m+1}$ and $w_k = w_{k,1}$ to be consistent with notations in (GD).

2.1. The minimum-norm bias

Before presenting our results on the directional bias, let us first recap the well-known minimum-norm bias for SGD/GD optimizing linear regression problem [4, 5, 11]. We rewrite the training loss as $L_{\mathcal{S}}(w) = \frac{1}{n} \left\| X^\top w - Y \right\|_2^2$, where $X = (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $Y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Then its global minima are given by $\mathcal{W}_* := \left\{ w \in \mathbb{R}^d : Pw = PY \right\}$, where P is the projection operator onto the data manifold, i.e., the column space of X. We focus on overparameterized cases where $\mathcal{W}_*$ contains multiple elements.

Notice that every gradient $\nabla \ell_i(w) = 2x_i(x_i^\top w - y_i)$ is spanned in the data manifold, thus (GD) and (SGD) can never move along the direction that is orthogonal to the data manifold. In other words, (GD) and (SGD) implicitly admit the following *hypothesis class*:

$$\mathcal{H}_{\mathcal{S}} = \left\{ w \in \mathbb{R}^d : P_\perp w = P_\perp w_0 \right\}, \tag{1}$$

where w_0 is the initialization and $P_\perp = I - P$ is the projection operator onto the orthogonal complement to the column space of X.

Putting things together, for any global optimum $w \in \mathcal{W}_*$ (hence $Pw = PY$), we have

$$\|w - w_0\|_2^2 = \|Pw - Pw_0\|_2^2 + \|P_\perp w - P_\perp w_0\|_2^2 = \|PY - Pw_0\|_2^2 + \|P_\perp w - P_\perp w_0\|_2^2,$$

where the right hand side is minimized when $P_\perp w = P_\perp w_0$, i.e., $w \in \mathcal{H}_{\mathcal{S}}$, thus w is the solution found by SGD/GD in the non-degenerated cases (when the learning rate is set properly so that the algorithms can find a global optimum). In sum, SGD/GD is biased to find the global optimum that is *closest to the initialization*, which is referred as the "minimum-norm" bias in literature since the initialization is usually set to be zero.

3. Main Theory

In this section we present our main theoretical results. The proofs are deferred to Appendix B.

We specify the population distribution of $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ in the following manner: (1) we consider the feature vector given by $x = \zeta \cdot \xi$, where $\zeta \in \mathbb{R}^1$ is a magnitude random variable that takes value in $(0, 1]$, and $\xi \in \mathbb{R}^d$ is an angle random variable that follows a sphere uniform distribution, i.e., $\xi \sim \mathcal{U}(S^{d-1})$; (2) we consider a realizable setting where the label is given by $y = w_*^\top x$, i.e., there exists a true parameter $w_* \in \mathbb{R}^d$ that generates the label from the feature vector. Then the test loss is $L_{\mathcal{D}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[(w - w_*)^\top x x^\top (w - w_*) \right] = \mu \|w - w_*\|_2^2$, where $\mu = \mathbb{E}[\zeta]/d$. For an i.i.d. generated training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$, the training loss and the individual losses are

$$L_{\mathcal{S}}(w) = \frac{1}{n} (w - w_*)^\top X X^\top (w - w_*), \quad \ell_i(w) = (w - w_*)^\top x_i x_i^\top (w - w_*), \quad i = 1, \ldots, n,$$

where $X = (x_1, \ldots, x_n)$. We denote by P the projection operator onto the column space of X (the data manifold). For $i \in [n]$, we denote $\lambda_i := \|x_i\|_2^2$. By the previous data generalization process, we have $\lambda_i \in (0, 1]$. Without loss of generality, we assume $\{\lambda_i\}_{i\in[n]}$ are sorted in a descending order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. With the these preparations, we are ready to state our main theorems.

3.1. The directional bias of SGD

We first present Theorems 1 and 2 that characterize the different directional biases of SGD and GD in the moderate learning rate regime.

Theorem 1 (The directional bias of SGD with moderate LR, informal) *Suppose $d \geq \mathrm{poly}\,(n)$, $\lambda_1 > \lambda_2 + \mathrm{o}\,(1)$ and $\lambda_n > \mathrm{o}\,(1)$, where $\mathrm{o}\,(1)$ is a small positive constant depending on n/d, and the initialization is away from w_*. Consider (SGD) with the following moderate learning rate scheme*

$$\eta_k = \begin{cases} \eta \in \left(\frac{b}{\lambda_1 - \mathrm{o}(1)}, \frac{b}{\lambda_2 + \mathrm{o}(1)}\right), & k = 1, \ldots, k_1; \\ \eta' \in \left(0, \frac{b}{2\lambda_1}\right), & k = k_1 + 1, \ldots, k_2. \end{cases} \tag{2}$$

Then for $\epsilon = \mathrm{o}\,(1)$, there exist $k_1 > \mathcal{O}\left(\log(1/\epsilon)\right)$ and k_2, such that with high probability the output of SGD w^{sgd} satisfies

$$(1-\epsilon)\cdot\gamma_1 \leq \frac{\left(P(w^{\mathrm{sgd}} - w_*)\right)^\top \cdot XX^\top \cdot P\left(w^{\mathrm{sgd}} - w_*\right)}{\|P\left(w^{\mathrm{sgd}} - w_*\right)\|_2^2} \leq \gamma_1, \tag{3}$$

where γ_1 is the largest eigenvalue of the data matrix $XX^\top$.

Theorem 2 (The directional bias of GD with moderate or small LR, informal) *Suppose $d \geq \mathrm{poly}\,(n)$, $\lambda_{n-1} > \lambda_n + \mathrm{o}\,(1)$ and $\lambda_n > \mathrm{o}\,(1)$, where $\mathrm{o}\,(1)$ is a small positive constant depending on n/d, and the initialization is away from w_*. Consider (GD) with the following moderate or small learning rate scheme*

$$\eta_k \in \left(0, \frac{n}{2\lambda_1 + \mathrm{o}\,(1)}\right), \quad k = 1, \ldots, k_2. \tag{4}$$

Then for any $\epsilon > 0$, if $k_2 > \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$, then with high probability the output of GD w^{gd} satisfies

$$\gamma_n \leq \frac{\left(P(w^{\mathrm{gd}} - w_*)\right)^\top \cdot XX^\top \cdot P\left(w^{\mathrm{gd}} - w_*\right)}{\|P\left(w^{\mathrm{gd}} - w_*\right)\|_2^2} \leq (1+\epsilon)\cdot\gamma_n, \tag{5}$$

where γ_n is the smallest eigenvalue of the data matrix $XX^\top$ restricted in the column space of X.

Remark 1 *As the Rayleigh quotient (3) (resp. (5)) converges to its maximum (resp. minimum), the vector gets closer to the eigenvector of the largest (resp. smallest) eigenvalue [32]. Thus Theorems 1 and 2 suggest that, when projected onto the data manifold, SGD and GD converge to the optimum along the largest and smallest eigenvalue direction respectively. Here we are only interested in the projection onto the data manifold, since SGD/GD cannot move along the direction that is orthogonal to the data manifold as discussed in Section 2.1.*

Remark 2 *We legitimately assume $b < n/2 - \mathrm{o}(1)$ since it is not very meaningful to discuss SGD that uses more than (roughly) half of the training set as a mini-batch. Then the learning rate schedule in (2) intersects with that in (4), i.e.,* (4) covers both moderate and small learning rate schemes. *And their intersection determines a moderate learning rate scheme, where SGD converges along the large eigenvalue directions while GD goes after the small eigenvalue directions. This justifies the regularization effect of SGD with moderate learning rate.*

Remark 3 *Technically, in Theorem 1 one can set $b = n$ to include GD as a special case, so that GD also follows the large eigenvalue directions. However, the initial learning rate in (2) needs to be at least $\frac{n}{\lambda_1} \geq n$ ignoring the small order term, which is way too large to be even numerically stable in practical big data circumstances. The typical learning rate for GD falls into learning rate schedule in (4). Thus it follows the small eigenvalue directions according to Theorem 2.*

Note GD with small learning rate also converges along the small eigenvalue directions, since the learning rate schedule in (4) covers the small learning rate scheme. In complement, the following Theorem 3 shows that in the small learning rate regime, SGD is imitating GD and converges along the small eigenvalue directions as well. Theorems 1, 2 and 3 together show that, converging along the large eigenvalue directions is a distinct regularization effect that is unique to *SGD* with *moderate learning rate*.

Theorem 3 (The directional bias of SGD with small LR, informal) *Theorem 2 applies to* (SGD) *with the following small learning rate scheme*

$$\eta_k = \eta' \in \Big(0,\ \frac{b}{2\lambda_1 + \mathrm{o}(1)}\Big), \quad k = 1, \ldots, k_2. \tag{6}$$

3.2. Effects of the directional bias

Next we justify the benefit of the particular directional bias of SGD with moderate learning rate. Recall the hypothesis class $\mathcal{H}_S$ (Eq. (1)) for SGD and GD. Then for an algorithm with output w^{alg}, we have the following generalization error decomposition [30],

$$L_{\mathcal{D}}(w^{\mathrm{alg}}) - \inf_w L_{\mathcal{D}}(w) = \underbrace{L_{\mathcal{D}}(w^{\mathrm{alg}}) - \inf_{w' \in \mathcal{H}_S} L_{\mathcal{D}}(w')}_{\Delta(w^{\mathrm{alg}}),\ \text{estimation error}} + \underbrace{\inf_{w' \in \mathcal{H}_S} L_{\mathcal{D}}(w') - \inf_w L_{\mathcal{D}}(w)}_{\text{approximation error}}.$$

The *approximate error* is an intrinsic error determined by the hypothesis class, and is not improvable unless enlarging the hypothesis class. In contrast, the *estimation error* $\Delta(w^{\mathrm{alg}})$ is determined by the algorithm as well as its hyperparameters. Thus, in the following theorem, we use the estimation error to compare the generalization performance of the SGD and GD outputs in different learning rate regimes.

Theorem 4 (Effects of the directional bias, informal) *Let $\mathcal{W}_\alpha := \{w \in \mathcal{H}_S : L_S(w) = \alpha\}$ be an α-level set of the training loss $L_S(w)$. Let $\Delta^*_\alpha := \inf_{w \in \mathcal{W}_\alpha} \Delta(w)$ be the minimum estimation error within the α-level set $\mathcal{W}_\alpha$. Under the same conditions as Theorems 1 and 2, if $b < \frac{n}{2} - \mathrm{o}(1)$, then there exists k_1 and k_2 such that with high probability:*

- *The output of* (SGD) *with moderate LR* (2) *satisfies $\Delta(w^{\mathrm{sgd}}) < (1+\epsilon) \cdot \Delta^*_\alpha$, where α is the training loss of w^{sgd}, and $\epsilon = \mathrm{o}(1)$ is a small constant;*

- *The output of* (GD) *with moderate or small LR* (4) *satisfies* $\Delta(w^{\mathrm{gd}}) > M \cdot \Delta^*_\alpha$, *where* α *is the training loss of* w^{gd}, *and* $M = \gamma_1/\gamma_n - \mathrm{o}(1) > 1$ *is a constant;*
- *The output of* (SGD) *with small LR* (6) *satisfies* $\Delta(w^{\mathrm{sgd}}) > M \cdot \Delta^*_\alpha$, *where* α *is the training loss of* w^{sgd}, *and* $M = \gamma_1/\gamma_n - \mathrm{o}(1) > 1$ *is a constant.*

Remark 4 *In practice it is usually intractable and unnecessary to achieve the exact global minima of the training loss; instead we often* early stop *the algorithm once obtaining a small enough training loss, i.e., reaching an* α*-level set. In this case, Theorem 4 guarantees SGD with moderate learning is nearly optimal in terms of estimation error, but GD and SGD with small learning rate is only suboptimal.*

4. Conclusion

We characterize a distinct directional regularization effect of SGD with moderate learning rate, where SGD converges along the large eigenvalue directions of the data matrix. In contrast, neither GD nor SGD with small learning rate can achieve this effect. Moreover, we show this directional bias benefits generalization when early stopping is adopted.

References

[1] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1370–1378, 2019.

[2] Alnur Ali, Edgar Dobriban, and Ryan J Tibshirani. The implicit regularization of stochastic gradient flow for least squares. *arXiv preprint arXiv:2003.07802*, 2020.

[3] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019.

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

[5] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[6] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

[7] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[8] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*, 2019.

[9] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[10] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[11] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *ICML*, 2018.

[12] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[15] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[16] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2019.

[17] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

[18] Ziwei Ji and Matus Jan Telgarsky. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[19] Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136, 2020.

[20] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[21] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[22] Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training. *arXiv preprint arXiv:2002.10376*, 2020.

[23] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[24] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pages 11674–11685, 2019.

[25] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354, 2018.

[26] Edward Moroshko, Suriya Gunasekar, Blake Woodworth, Jason D Lee, Nathan Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *arXiv preprint arXiv:2007.06738*, 2020.

[27] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[28] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019.

[29] Preetum Nakkiran. Learning rate annealing can provably help generalization, even for convex problems. *arXiv preprint arXiv:2005.07360*, 2020.

[30] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[31] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[32] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

[33] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[34] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. *The 37th International Conference on Machine Learning*, 2020.

[35] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[36] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *The 36th International Conference on Machine Learning*, 2019.

Appendix A. Preliminary for Appendix

A.1. Additional notations

We adopt the notations and settings in main text. In addition we make the following notations.

For a vector $x \in \mathbb{R}^d$, denote its direction as $\bar{x} := \frac{x}{\|x\|_2}$. For simplicity assume the training data $\{x_1, \ldots, x_n\}$ are linear independent. For training data $x_i,\ i \in [n]$, we denote $\lambda_i = \|x_i\|_2^2$, then by construction we have $\lambda_i \in (0, 1]$. Without loss of generality let $\lambda_1 \geq \cdots \geq \lambda_n$. We define

$$X := (x_1, \ldots, x_n) \in \mathbb{R}^{d\times n},$$
$$X_{-1} := (x_2, x_3, \ldots, x_n) \in \mathbb{R}^{d\times(n-1)}.$$

Then based on the above definitions, we define the following two projection operators

$$P = X(X^\top X)^{-1}X^\top,$$
$$P_\perp = I - P.$$

Clearly for any $v \in \mathbb{R}^d$, Pv projects v onto subspace $\operatorname{span}\{x_1, \ldots, x_n\}$, while $P_\perp v$ projects v onto the orthogonal complement of $\operatorname{span}\{x_1, \ldots, x_n\}$. Furthermore we introduce two more projection operators

$$P_{-1} = X_{-1}(X_{-1}^\top X_{-1})^{-1}X_{-1}^\top,$$
$$P_1 = P - P_{-1} = I - P_\perp - P_{-1}.$$

For any $v \in \mathbb{R}^d$, $P_{-1}v$ projects v onto subspace $\operatorname{span}\{x_2, \ldots, x_n\}$, while $P_1 v$ projects v into the orthogonal complement of $\operatorname{span}\{x_2, \ldots, x_n\}$ with respect to $\operatorname{span}\{x_1, \ldots, x_n\}$. In the following, we often write the column space of P, which refers to $\{Pv : v \in \mathbb{R}^d\}$, similarly for P_{-1}, P_1 and $P_\perp$ as well. Clearly the column space of P is also $\operatorname{span}\{x_1, \ldots, x_n\}$; the column space of P_{-1} is also the data manifold $\operatorname{span}\{x_2, \ldots, x_n\}$. We highlight that the total space $\mathbb{R}^d$ can be decomposed as the direct sum of the column space of P_{-1}, P_1 and $P_\perp$, i.e., $I = P_{-1} + P_1 + P_\perp$. By definition, it is easy to verify that

$$P_\perp X = 0,$$
$$PX = X,$$
$$P_1 X = (P_1 x_1, 0, \ldots, 0),$$
$$P_{-1}X = (P_{-1}x_1, x_2, \ldots, x_n).$$

Then we define the following matrices which will be repeatedly used in the subsequent proof.

$$H := XX^\top,$$
$$H_{-1} := (P_{-1}X)(P_{-1}X)^\top,$$
$$H_1 := (P_1 X)(P_1 X)^\top,$$
$$H_c := (P_{-1}x_1)(P_1 x_1)^\top + (P_1 x_1)(P_{-1}x_1)^\top.$$

Based on the above definitions, it is easy to show that

$$\begin{aligned} H &= (P_1 X + P_{-1}X)(P_1 X + P_{-1}X)^\top \\ &= (P_{-1}X)(P_{-1}X)^\top + (P_1 X)(P_1 X)^\top + (P_1 X)(P_{-1}X)^\top + (P_{-1}X)(P_1 X)^\top \\ &= H_{-1} + H_1 + H_c. \end{aligned}$$

A.2. Lemmas

We present the following theorems and lemmas as preparation for our analysis.

Theorem *[Gershgorin circle theorem, restated for symmetric matrix] Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Let A_{ij} be the entry in the i-th row and the j-th column. Let*

$$R_i(A) := \sum_{j\neq i} |A_{ij}|\,,\ i = 1, \ldots, n.$$

Consider n Gershgorin discs

$$D_i(A) := \{z \in \mathbb{R}, |z - A_{ii}| \leq R_i(A)\}\,,\ i = 1, \ldots, n.$$

The eigenvalues of A are in the union of Gershgorin discs

$$G(A) := \bigcup_{i=1}^{n} D_i(A).$$

Furthermore, if the union of k of the n discs that comprise $G(A)$ forms a set $G_k(A)$ that is disjoint from the remaining $n-k$ discs, then $G_k(A)$ contains exactly k eigenvalues of A, counted according to their algebraic multiplicities.

Proof See, e.g., Horn and Johnson [14], Chap 6.1, Theorem 6.1.1. ■

Theorem *[Hoffman-Wielandt theorem, restated for symmetric matrix] Let $A, E \in \mathbb{R}^{n\times n}$ be symmetric. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A, arranged in decreasing order. Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ be the eigenvalues of $A + E$, arranged in decreasing order. Then*

$$\sum_{i=1}^{n} \left|\hat{\lambda}_i - \lambda_i\right|^2 \leq \|E\|_F^2\,.$$

Proof See, e.g., Horn and Johnson [14], Chapter 6.3, Theorem 6.3.5 and Corollary 6.3.8. ■

Lemma 1 *Let $d \geq 4\log(2n^2/\delta)$ for some $\delta \in (0, 1)$. Then with probability at least $1 - \delta$, we have*

$$|\langle \bar{x}_i, \bar{x}_j\rangle| < \iota := \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{d}}\right), \quad i \neq j.$$

Proof See Section C.1. ■

By Lemma 1 we can assume $d \geq \text{poly}\,(n)$ such that $n\iota$ is sufficiently small depends on requirements.

The following two lemmas characterize the projected components of each training data onto the column space of P_1, P_{-1}, and $P_\perp$.

Lemma 2 *For $x_j \neq x_1$, we have*

- $P_{-1}x_j = x_j$;
- $P_1 x_j = 0$;
- $P_\perp x_j = 0$.

Proof These are by the construction of the projection operators. ■

Lemma 3 *Assume $\sqrt{n}\iota \le 1/4$. With probability at least $1-\delta$, we have*

- $0 \le \|P_{-1}\bar{x}_1\|_2 \le 2\sqrt{n}\iota$;
- $\sqrt{1-4n\iota^2} \le \|P_1\bar{x}_1\|_2 \le 1$;
- $P_\perp x_1 = 0$.

Proof See Section C.2. ■

The following four lemmas characterize the spectrum of the matrices H, H_{-1}, H_1 and H_c.

Lemma 4 *Let $\gamma_1, \dots, \gamma_n$ be the n non-zero eigenvalues of $H := XX^\top$ in decreasing order. then*

$$\lambda_n - n\iota \le \gamma_1, \dots, \gamma_n \le \lambda_1 + n\iota.$$

Furthermore, if there exist λ_r and λ_{r+1} such that $\lambda_r > \lambda_{r+1} + 2n\iota$, then

$$\lambda_n - n\iota \le \gamma_{r+1}, \dots, \gamma_n \le \lambda_{r+1} + n\iota < \lambda_r - n\iota \le \gamma_1, \dots, \gamma_r \le \lambda_1 + n\iota.$$

Proof See Section C.3. ■

Lemma 5 *Assume $\lambda_n \ge 3n\iota$. Consider the symmetric matrix $H_{-1} := P_{-1}X(P_{-1}X)^\top \in \mathbb{R}^{d\times d}$.*

- *0 is an eigenvalue of H_{-1} with algebraic multiplicity being $d-n+1$, and its corresponding eigenspace is the column space of $P_1 + P_\perp$.*
- *Restricted in the column space of P_{-1}, the $n-1$ eigenvalues of H_{-1} belong to*

$$(\lambda_n - n\iota,\ \ \lambda_2 + n\iota).$$

Proof See Section C.4. ■

Lemma 6 *Consider matrix $H_1 := P_1X(P_1X)^\top \in \mathbb{R}^{d\times d}$. We have H_1 has only one non-zero eigenvalue, which belongs to*

$$[\lambda_1\left(1-4n\iota^2\right),\ \ \lambda_1].$$

Moreover, the corresponding eigenspace is the column space of P_1, which is 1-dim.

Proof Clearly H_1 is rank-1 since the column space of P_1 is 1-dim. Thus it has only one non-zero eigenvalue, which is given by

$$\mathrm{tr}(H_1) = \sum_{i=1}^{n} \|P_1 x_i\|_2^2 = \|P_1 x_1\|_2^2 \in [\lambda_1 \left(1 - 4n\iota^2\right),\ \lambda_1],$$

where the last equality follows from Lemma 3. ■

Lemma 7 *Consider matrix* $H_c := (P_{-1}x_1)(P_1 x_1)^\top + (P_1 x_1)(P_{-1}x_1)^\top \in \mathbb{R}^{d\times d}$.

$$\|H_c\|_2 \le 2\lambda_1 \|P_{-1}\bar{x}_1\|_2 \le 4\sqrt{n}\iota.$$

Proof

$$\|H_c\|_2 \le 2\|P_{-1}x_1\|_2 \cdot \|P_1 x_1\|_2 \le 2\lambda_1 \|P_{-1}\bar{x}_1\|_2 \le 4\sqrt{n}\iota,$$

where the last equality follows from Lemma 3 and $\lambda_1 \le 1$.

■

Appendix B. Missing Proofs for the Theorems in Main Text

B.1. The directional bias of SGD with moderate learning rate

Reloading notations Let $\pi^k := \left\{\mathcal{B}_1^k, \dots, \mathcal{B}_m^k\right\}$ be a randomly chosen uniform m-partition of $[n]$, where $n = mb$. Then the SGD iterates at the k-th epoch can be formulated as:

$$w_{k,j+1} = w_{k,j} - \frac{2\eta_k}{b} \sum_{i\in\mathcal{B}_j^k} x_i x_i^\top (w_{k,j} - w_*), \quad j = 1, \dots, m,$$

where we assume that the learning rate is fixed within each epoch. Note here π^k is independently and randomly chosen at each epoch. For simplicity we often ignore the epoch-indicator k, and write the uniform partition as $\pi := \{\mathcal{B}_1, \dots, \mathcal{B}_m\}$. It is clear from context that π is random over epochs. For a mini-batch $\mathcal{B}_j \in \pi$, we denote $H(\mathcal{B}_j) := \sum_{i\in\mathcal{B}_j} x_i x_i^\top$.

Considering translating the variable by

$$v = w - w_*,$$

then we can reformulate the SGD update rule as

$$v_{k,j+1} = v_{k,j} - \frac{2\eta_k}{b} H(\mathcal{B}_j) v_{k,j} = \left(I - \frac{2\eta_k}{b} H(\mathcal{B}_j)\right) v_{k,j}, \quad j = 1, \dots, m. \tag{7}$$

Let

$$\begin{aligned}\mathcal{M}_\pi &:= \prod_{j=1}^{m} \left(I - \frac{2\eta_k}{b} H(\mathcal{B}_j)\right) \\ &:= \left(I - \frac{2\eta_k}{b} H(\mathcal{B}_m)\right) \cdot \left(I - \frac{2\eta_k}{b} H(\mathcal{B}_{m-1})\right) \cdots \left(I - \frac{2\eta_k}{b} H(\mathcal{B}_1)\right).\end{aligned}$$

Here the matrix production over a sequence of matrices $\{M_i \in \mathbb{R}^{d\times d}\}_{j=1}^{m}$ is defined from the left to right with descending index,

$$\prod_{j=1}^{m} M_j := M_m \times M_{m-1} \times \cdots \times M_1.$$

Let $v_{k+1} = v_{k,m+1}$ and $v_k = v_{k,1}$. Then we can further reformulate Eq. (7) and obtain the epoch-wise update of SGD as

$$v_{k+1} = \left(I - \frac{2\eta_k}{b} H(\mathcal{B}_m)\right) \cdot \left(I - \frac{2\eta_k}{b} H(\mathcal{B}_{m-1})\right) \cdots \left(I - \frac{2\eta_k}{b} H(\mathcal{B}_1)\right) \cdot v_k = \mathcal{M}_\pi v_k. \quad (8)$$

In light of the notion of v, the following lemma restates the related notations of loss functions, hypothesis class, level set, and estimation error defined in Section 2 and 3.

Lemma 8 (Reloading SGD notations) *Regarding repramaterization $v = w - w_*$, we can reload the following related notations:*

- *Empirical loss and population loss are*

$$L_{\mathcal{S}}(v) = \frac{1}{n}(P_1 v)^\top H_1 (P_1 v) + \frac{1}{n}(P_{-1} v)^\top H_{-1}(P_{-1} v) + \frac{1}{n}(Pv)^\top H_c (Pv),$$
$$L_{\mathcal{D}}(v) = \mu \left\|v\right\|_2^2.$$

- *The hypothesis class is*

$$\mathcal{H}_{\mathcal{S}} = \left\{v \in \mathbb{R}^d : P_\perp v = P_\perp v_0\right\}.$$

- *The α-level set is*

$$\mathcal{V} = \left\{v \in \mathcal{H}_{\mathcal{S}} : L_{\mathcal{S}}(v) = \alpha\right\}.$$

- *For $v \in \mathcal{H}_{\mathcal{S}}$, the estimation error is*

$$\Delta(v) = \mu \left\|Pv\right\|_2^2.$$

Moreover,

$$\Delta_* = \inf_{v\in\mathcal{V}} \Delta(v) = \frac{\mu n \alpha}{\gamma_1}.$$

Proof See Section C.5. ■

Based on the above definitions, the following lemma characterizes the one-step update of SGD.

Lemma 9 (One step SGD update) *Consider the j-th SGD update at the k-th epoch as given by Eq. (7). Set the learning rate be constant η during that epoch. Then for $j = 1, \ldots, m$ we have*

$$\begin{pmatrix} P_1 v_{k,j+1} \\ P_{-1} v_{k,j+1} \end{pmatrix} = \begin{pmatrix} I - \frac{2\eta}{b} P_1 H(\mathcal{B}_j) P_1 & -\frac{2\eta}{b} P_1 H(\mathcal{B}_j) P_{-1} \\ -\frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_1 & I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1} \end{pmatrix} \cdot \begin{pmatrix} P_1 v_{k,j} \\ P_{-1} v_{k,j} \end{pmatrix}$$

Moreover, if $1 \notin \mathcal{B}_j$, i.e., x_1 is not used in the j-the step, then

$$\begin{pmatrix} P_1 v_{k,j+1} \\ P_{-1} v_{k,j+1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1} \end{pmatrix} \cdot \begin{pmatrix} P_1 v_{k,j} \\ P_{-1} v_{k,j} \end{pmatrix}$$

Proof See Section C.6. ■

Eq. (8) indicates the key to analyze the convergence of v_{k+1} is to characterize the spectrum of the matrix $\mathcal{M}_\pi$. In particular the following lemma bounds the spectrum of $\mathcal{M}_\pi$ when projected onto the column space of P_{-1}.

Lemma 10 *Suppose $3n\iota < \lambda_n$. Suppose $0 < \eta < \frac{b}{\lambda_2 + 3n\iota}$. Let $\pi := \{\mathcal{B}_1, \dots, \mathcal{B}_m\}$ be a uniform m partition of index set $[n]$, where $n = mb$. Consider the following $d \times d$ matrix*

$$\mathcal{M}_{-1} := \prod_{j=1}^{m} \left(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1} \right) \in \mathbb{R}^{d \times d}.$$

Then for the spectrum of $\mathcal{M}_{-1}^\top \mathcal{M}_{-1}$ we have:

- *1 is an eigenvalue of $\mathcal{M}_{-1}^\top \mathcal{M}_{-1}$ with multiplicity being $d - n + 1$; moreover, the corresponding eigenspace is the column space of $P_1 + P_\perp$.*
- *Restricted in the column space of P_{-1}, the eigenvalues of $\mathcal{M}_{-1}^\top \mathcal{M}_{-1}$ are upper bounded by $(q_{-1}(\eta))^2 < 1$, where*

$$q_{-1}(\eta) := \max\left\{ \left| 1 - \frac{2\eta}{b}(\lambda_2 + n\iota) \right| + \frac{3n\eta\iota}{b},\ \left| 1 - \frac{2\eta}{b}(\lambda_n - n\iota) \right| + \frac{3n\eta\iota}{b} \right\} < 1.$$

Proof See Section C.7. ■

Consider the projections of v_k onto the column space of P_{-1} and P_1. For simplicity let

$$A_k := \|P_{-1} v_k\|_2, \quad B_k := \|P_1 v_k\|_2.$$

The following lemma controls the update of A_k and B_k.

Lemma 11 (Update rules for A_k and B_k) *Suppose $3n\iota < \lambda_n$. Suppose $0 < \eta < \frac{b}{\lambda_2 + 3n\iota}$. Consider the k-th epoch of SGD iterates given by Eq. (8). Set the learning rate in this epoch to be constant η. Denote*

$$\xi(\eta) := \frac{4\eta\sqrt{n}\iota}{b},$$
$$q_1(\eta) := \left| 1 - \frac{2\eta\lambda_1}{b} \|P_1 \bar{x}_1\|_2^2 \right|,$$
$$q_{-1}(\eta) := \max\left\{ \left| 1 - \frac{2\eta}{b}(\lambda_2 + n\iota) \right| + \frac{3n\eta\iota}{b},\ \left| 1 - \frac{2\eta}{b}(\lambda_n - n\iota) \right| + \frac{3n\eta\iota}{b} \right\} < 1.$$

Then we have the following:

- $A_{k+1} \le q_{-1}(\eta) \cdot A_k + \xi(\eta) \cdot B_k$.
- $B_{k+1} \le q_1(\eta) \cdot B_k + \xi(\eta) \cdot A_k$.
- $B_{k+1} \ge q_1(\eta) \cdot B_k - \xi(\eta) \cdot A_k$.

Proof See Section C.8. ■

Note we can rephrase the update rules for A_k and B_k as

$$\begin{pmatrix} A_{k+1} \\ B_{k+1} \end{pmatrix} \leq \begin{pmatrix} q_{-1}(\eta) & \xi(\eta) \\ \xi(\eta) & q_1(\eta) \end{pmatrix} \cdot \begin{pmatrix} A_k \\ B_k \end{pmatrix},$$

where "$\leq$" means "entry-wisely smaller than".

The following two lemmas characterize the long run behaviors of A_k and B_k with different learning rate.

Lemma 12 (The long run behavior of SGD with moderate LR) *Suppose $3n\iota < \lambda_n$, and $\lambda_2 + 4n\iota < \lambda_1$. Suppose v_0 is far away from 0. Consider the first k_1 epochs of SGD iterates given by Eq. (8). Set the learning rate during this stage to be constant, i.e., $\eta_k = \eta$ for $0 \leq k < k_1$. Suppose*

$$\frac{b}{\lambda_1 - 3\sqrt{n\iota}} < \eta < \frac{b}{\lambda_2 + 3n\iota}.$$

Then for $0 < \epsilon < 1$ and $0 < \beta < \beta_0 < B_0$ satisfying $\sqrt{n\iota} \leq \operatorname{poly}(\epsilon\beta)$, there exists $k_1 \geq \mathcal{O}\left(\log \frac{1}{\epsilon\beta}\right)$ such that

- $A_{k_1} \leq \epsilon \cdot \beta$.
- $B_{k_1} \leq \|Pv_0\|_2 \cdot \rho_1^{k_1} + \frac{\epsilon}{2} \cdot \beta = \operatorname{poly}\left(\frac{1}{\epsilon\beta}\right)$.
- *For all $k = 0, 1, \ldots, k_1$, $B_k > \beta_0$.*

Proof See Section C.9. ■

Lemma 13 (The long run behavior of SGD with small LR) *Suppose $3n\iota < \lambda_n$, and $\lambda_2 + 4n\iota < \lambda_1$. Suppose v_0 is far away from 0. Consider another $k_2 - k_1$ epochs of SGD iterates given by Eq. (8). Set the learning rate to be constant during the updates, i.e., $\eta_k = \eta'$ for $k_1 \leq k < k_2$. Suppose*

$$0 < \eta' < \frac{b}{2\lambda_1}.$$

Consider the ϵ and β given in Lemma 12. Then for $k \geq k_1$, we have

- $A_k \leq \epsilon \cdot \beta$.
- $B_k \leq \begin{cases} q \cdot B_{k-1}, & B_{k-1} > \beta, \\ \beta, & B_{k-1} < \beta. \end{cases}$ *where $q \in (0, 1)$ is a constant.*

Proof See Section C.10. ■

Theorem 5 (Theorem 1, formal version) *Suppose $3n\iota < \lambda_n$ and $\lambda_2 + 4n\iota < \lambda_1$. Suppose v_0 is away from 0. Consider the SGD iterates given by Eq. (8) with the following moderate learning rate scheme*

$$\eta_k = \begin{cases} \eta \in \left(\frac{b}{\lambda_1 - 3\sqrt{n\iota}}, \frac{b}{\lambda_2 + 3n\iota}\right), & k = 1, \dots, k_1; \\ \eta' \in \left(0, \frac{b}{2\lambda_1}\right), & k = k_1 + 1, \dots, k_2. \end{cases}$$

Then for $0 < \epsilon < 1$ such that $\sqrt{n\iota} \le \text{poly}(\epsilon)$, there exist $k_1 > \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ and k_2 such that

$$(1 - \epsilon) \cdot \gamma_1 \le \frac{v_{k_2}^\top H v_{k_2}}{\|P v_{k_2}\|_2^2} \le \gamma_1.$$

Proof We choose k_1 and k_2 as in Lemma 12 and Lemma 13 with β set as a small constant, then we are guaranteed to have

$$A_{k_2} \le \epsilon \cdot \beta \le \epsilon \cdot B_{k_2},$$

from where we have

$$\frac{\|P_1 v_{k_2}\|_2^2}{\|P v_{k_2}\|_2^2} = \frac{B_{k_2}^2}{A_{k_2}^2 + B_{k_2}^2} \ge \frac{1}{1 + \epsilon^2} \ge 1 - \epsilon^2.$$

Then we have

$$\begin{aligned} \frac{v_{k_2}^\top H v_{k_2}}{\|P v_{k_2}\|_2^2} &= \frac{(P_1 v_{k_2})^\top H_1 (P_1 v_{k_2})}{\|P v_{k_2}\|_2^2} + \frac{(P_{-1} v_{k_2})^\top H_{-1} (P_{-1} v_{k_2})}{\|P v_{k_2}\|_2^2} + \frac{(P v_{k_2})^\top H_c (P v_{k_2})}{\|P v_{k_2}\|_2^2} \\ &\ge \lambda_1 \left(1 - 4n\iota^2\right) \cdot \frac{\|P_1 v_{k_2}\|_2^2}{\|P v_{k_2}\|_2^2} + 0 - 4\sqrt{n\iota} \\ &\ge \lambda_1 (1 - 4n\iota^2) \cdot (1 - \epsilon^2) - 4\sqrt{n\iota} \\ &\ge (\gamma_1 - n\iota)(1 - 4n\iota^2) \cdot (1 - \epsilon^2) - 4\sqrt{n\iota} \qquad (\text{since } \gamma_1 \le \lambda_1 + n\iota \text{ by Lemma 4}) \\ &= \gamma_1 (1 - 4n\iota^2)(1 - \epsilon^2) - n\iota (1 - 4n\iota^2)(1 - \epsilon^2) - 4\sqrt{n\iota} \\ &\ge \gamma_1 (1 - 0.5\epsilon) - 0.5\gamma_1 \epsilon \qquad (\text{since } \sqrt{n\iota} \le \text{poly}(\epsilon)) \\ &= \gamma_1 (1 - \epsilon). \end{aligned}$$

■

Theorem 6 (Theorem 4 first part, formal version) *Suppose $3n\iota < \lambda_n$ and $\lambda_2 + 4n\iota < \lambda_1$. Suppose v_0 is away from 0. Consider the SGD iterates given by Eq. (8) with the following moderate learning rate schedule*

$$\eta_k = \begin{cases} \eta \in \left(\frac{b}{\lambda_1 - 3\sqrt{n\iota}}, \frac{b}{\lambda_2 + 3n\iota}\right), & k = 1, \dots, k_1; \\ \eta' \in \left(0, \frac{b}{2\lambda_1}\right), & k = k_1 + 1, \dots, k_2. \end{cases}$$

Then for $0 < \epsilon < 1$ satisfying $\sqrt{n\iota} \le \text{poly}(\epsilon)$, there exist k_1 and k_2 such that SGD outputs an ϵ-optimal solution.

Proof We set

$$\beta = \sqrt{\frac{n\alpha}{\gamma_1}}, \tag{9}$$

$$\beta_0 = \sqrt{\frac{n\alpha}{\gamma_n}} > \beta, \tag{10}$$

and apply Lemma 12 to choose a k_1 such that

$$\|P_{-1}v_{k_1}\|_2 \le \epsilon \cdot \beta = \epsilon \cdot \sqrt{\frac{n\alpha}{\gamma_1}}; \tag{11}$$

$$\|P_1 v_k\|_2 \ge \beta_0 = \sqrt{\frac{n\alpha}{\gamma_n}}, \quad \forall\, 0 \le k \le k_1. \tag{12}$$

Thus for all $0 \le k \le k_1$,

$$\begin{aligned} L_{\mathcal{S}}(v_k) &= \frac{1}{n}(Pv_k)^\top XX^\top (Pv_k) \\ &\ge \frac{\gamma_n}{n}\|Pv_k\|_2^2 \qquad (\gamma_n \text{ is the smallest eigenvalue of } XX^\top \text{ in the column space of } P) \\ &\ge \frac{\gamma_n}{n}\|P_1 v_k\|_2^2 \\ &> \alpha, \qquad (\text{by Eq. (12)}) \end{aligned}$$

which implies SGD cannot reach the α-level set during the iteration of first stage, i.e., SGD does not terminate in this stage.

We thus consider the second stage. From Lemma 13 we know $\|P_1 v_{kn}\|_2$ will keep decreasing before being smaller than β, and $\|P_{-1}v_k\|_2$ stays small during this period, i.e., SGD fits $P_1 v$ while in the same time does not mess up $P_{-1}v$. Mathematically speaking, there exists k_2 and α such that

$$A_{k_2} := \|P_{-1}v_{k_2}\|_2 \le \epsilon \cdot \beta = \epsilon \cdot \sqrt{\frac{n\alpha}{\gamma_1}},$$
$$L_{\mathcal{S}}(v_{k_2}) = \alpha,$$

which implies SGD terminates at the k_2-th epoch. Then by Lemmas 3 and 8, we have

$$\begin{aligned} n\alpha &= nL_{\mathcal{S}}(v_{k_2}) \\ &= (P_1 v_{k_2})^\top H_1 (P_1 v_{k_2}) + (P_{-1}v_{k_2})^\top H_2 (P_{-1}v_{k_2}) + (Pv_{k_2})^\top H_c (Pv_{k_2}) \\ &\ge (P_1 v_{k_2 n})^\top H_1 (P_1 v_{k_2 n}) - \|P_{-1}\bar{x}_1\|_2^2 \cdot \|Pv_{k_2 n}\|_2^2 \\ &\ge (\lambda_1 - n\iota)B_{k_2}^2 - 4n\iota^2(A_{k_2}^2 + B_{k_2}^2) \\ &\ge (\gamma_1 - 3n\iota)B_{k_2}^2 - 4n\iota^2 A_{k_2}^2, \end{aligned}$$

which yields

$$B_{k_2}^2 \le \frac{n\alpha + 4n\iota^2 A_{k_2}^2}{\gamma_1 - 3n\iota} \le \left(1 + \frac{\epsilon}{2}\right) \cdot \frac{n\alpha}{\gamma_1}.$$

Then we can bound the estimation error as

$$\begin{aligned}\Delta(v_{k_2}) &= \mu \left\| P v_{k_2} \right\|_2^2 \\ &= \mu(B_{k_2}^2 + A_{k_2}^2) \\ &\le \left(1 + \frac{\epsilon}{2}\right) \cdot \frac{\mu n \alpha}{\gamma_1} + \epsilon^2 \cdot \frac{\mu n \alpha}{\gamma_1} \\ &\le (1+\epsilon) \cdot \frac{\mu n \alpha}{\gamma_1} \\ &= (1+\epsilon) \cdot \Delta_*,\end{aligned}$$

where we use the fact that $\Delta_* = \mu n \alpha / \gamma_1$ from Lemma 8. Hence SGD is ϵ-near optimal. ■

B.2. The directional bias of GD with moderate or small learning rate

Reloading notations Denote the eigenvalue decomposition of $XX^\top$ as

$$XX^\top = G \Gamma G^\top, \quad \Gamma := \operatorname{diag}\left(\gamma_1, \ldots, \gamma_n, 0, \ldots, 0\right), \quad G = \left(g_1, \ldots, g_n, \ldots, g_d\right),$$

where $G \in \mathbb{R}^{d \times d}$ is orthonormal, and $\gamma_1, \ldots, \gamma_n$ are given by Lemma 4.

Clearly, $\operatorname{span}\{g_1, \ldots, g_n\} = \operatorname{span}\{x_1, \ldots, x_n\}$. Let

$$G_\parallel = \left(g_1, \ldots, g_n\right), \quad G_\perp = \left(g_{n+1}, \ldots, g_d\right),$$

then

$$P = G_\parallel G_\parallel^\top, \quad P_\perp = G_\perp G_\perp^\top.$$

Recall the GD iterates at the k-th epoch:

$$w_{k+1} = w_k - \frac{2\eta_k}{n} X X^\top \left(w_k - w_*\right).$$

Considering translating then rotating the variable as,

$$u = G^\top (w - w_*),$$

then we can reformulate the GD iterates as

$$u_{k+1} = u_k - \frac{2\eta_k}{n} \Gamma u_k = \left(I - \frac{2\eta_k}{n} \Gamma\right) u_k. \tag{13}$$

We present the following lemma to reload the related notations regarding the parameterization $u = G^\top (w - w_*)$.

Lemma 14 (Reloading GD notations) *Regarding reparametrization $u = G^\top (w - w_*)$, we can reload the following related notations:*

- *Empirical loss and population loss are*

$$L_{\mathcal{S}}(u) = \frac{1}{n} \sum_{i=1}^{n} \gamma_i \left(u^{(i)}\right)^2, \quad L_{\mathcal{D}}(u) = \mu \left\| u \right\|_2^2.$$

- *The hypothesis class is*

$$\mathcal{H}_{\mathcal{S}} = \left\{u \in \mathbb{R}^d : u^{(i)} = u_0^{(i)}, \text{ for } i = n+1, \ldots, d\right\}.$$

- *The α-level set is*

$$\mathcal{U} = \left\{u \in \mathcal{H}_u : L_{\mathcal{S}}(u) = \alpha\right\}.$$

- *For $u \in \mathcal{H}_{\mathcal{S}}$, the estimation error is*

$$\Delta(u) = \mu \sum_{i=1}^{n} \left(u^{(i)}\right)^2.$$

Moreover,

$$\Delta_* = \frac{\mu n \alpha}{\gamma_1}.$$

Proof See Section C.11. ■

The following lemma sovles GD iterates in Eq. (13).

Lemma 15 *For $t = 0, \ldots, T$,*

$$u_k^{(i)} = \begin{cases} \prod_{t=0}^{k-1} \left(1 - \frac{2\eta_t \gamma_i}{n}\right) \cdot u_0^{(i)}, & 1 \le i \le n; \\ u_0^{(i)}, & n+1 \le i \le d. \end{cases}$$

Proof This is by directly solving Eq. (13) where Γ is diagonal. ■

Theorem 7 (Theorem 2, formal version) *Suppose $\lambda_n + 2n\iota < \lambda_{n-1}$. Suppose u_0 is away from 0. Consider the GD iterates given by Eq. (13) with learning rate scheme*

$$\eta_k \in \left(0, \frac{n}{2\lambda_1 + 2n\iota}\right).$$

Then for $\epsilon \in (0,1)$, if $k \ge \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$, then we have

$$\gamma_n \le \frac{u_k^\top \Gamma u_k}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2} \le (1+\epsilon) \cdot \gamma_n.$$

Proof For $i = 1, \ldots, n$, denote $q_i(\eta) = 1 - \frac{2\gamma_i}{n} \cdot \eta$, where $\eta \in \left(0, \frac{n}{2\lambda_1 + 2n\iota}\right)$. Then we have $0 < q_i(\eta) < 1$ since

$$\eta < \frac{n}{2(\lambda_1 + n\iota)} < \frac{n}{2\gamma_1} \le \frac{n}{2\gamma_i},$$

where the second inequality follows from $\gamma_1 < \lambda_1 + n\iota$ by Lemma 4. Furthermore, since $\lambda_n + n\iota < \lambda_{n-1} - n\iota$, Lemma 4 gives us

$$0 < \gamma_n < \gamma_{n-1} \le \cdots \le \gamma_1 < 1, \tag{14}$$

which implies

$$1 > q_n(\eta) > q_{n-1}(\eta) \geq \cdots \geq q_1(\eta) > 0. \tag{15}$$

Moreover,

$$f(\eta) := \frac{q_{n-1}(\eta)}{q_n(\eta)} = \frac{1 - \frac{2\gamma_{n-1}}{n}\eta}{1 - \frac{2\gamma_n}{n}\eta}$$

is increasing, let $q = \max_{\eta < \frac{n}{2\lambda_1 + 2n\iota}} f(\eta)$, then $q < 1$ by our assumption on the learning rate.

From Lemma 15 we have

$$u_k^{(i)} = \prod_{t=0}^{k-1} q_i(\eta_t) \cdot u_0^{(i)}, \quad i = 1, \ldots, n. \tag{16}$$

By the assumption that

$$k > \frac{1}{2} \cdot \frac{\log \frac{\gamma_n \epsilon (u_0^{(n)})^2}{\gamma_1 n \sum_{i=1}^n \left(u_0^{(i)}\right)^2}}{\log q} = \mathcal{O}\left(\frac{1}{\epsilon}\right), \tag{17}$$

we have

$$\begin{aligned}
\frac{\sum_{i=1}^n \left(u_k^{(i)}\right)^2}{(u_k^{(n)})^2} &= 1 + \sum_{i=1}^{n-1} \frac{(u_k^{(i)})^2}{(u_k^{(n)})^2} \\
&= 1 + \sum_{i=1}^{n-1} \frac{\prod_{t=0}^{k-1} q_i(\eta_t)^2 \cdot (u_0^{(i)})^2}{\prod_{t=0}^{k-1} q_n(\eta_t)^2 \cdot (u_0^{(n)})^2} \qquad \text{(by Eq. (16))} \\
&\leq 1 + \frac{\sum_{i=1}^n \left(u_0^{(i)}\right)^2}{(u_0^{(n)})^2} \cdot \sum_{i=1}^{n-1} \prod_{t=0}^{k-1} \frac{q_i(\eta_t)^2}{q_n(\eta_t)^2} \\
&\leq 1 + \frac{\sum_{i=1}^n \left(u_0^{(i)}\right)^2}{(u_0^{(n)})^2} \cdot n \cdot \prod_{t=0}^{k-1} \frac{q_{n-1}(\eta_t)^2}{q_n(\eta_t)^2} \qquad \text{(by Eq. (15))} \\
&\leq 1 + \frac{\sum_{i=1}^n \left(u_0^{(i)}\right)^2}{(u_0^{(n)})^2} \cdot n \cdot q^{2k} \\
&\leq 1 + \frac{\gamma_n}{\gamma_1}\epsilon, \qquad \text{(by Eq. (17))}
\end{aligned}$$

which further yields

$$1 \geq \frac{(u_k^{(n)})^2}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2} \geq \frac{1}{1 + \frac{\gamma_n}{\gamma_1}\epsilon} \geq 1 - \frac{\gamma_n}{\gamma_1}\epsilon. \tag{18}$$

By the above inequalities we have

$$\begin{aligned}
\frac{u_k^\top \Gamma u_k}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2} &= \sum_{i=1}^n \frac{(u_k^{(i)})^2}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2} \cdot \gamma_i \\
&= \frac{(u_k^{(n)})^2}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2} \cdot \gamma_n + \sum_{i=1}^{n-1} \frac{(u_k^{(i)})^2}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2} \cdot \gamma_i \\
&\le \gamma_n + \sum_{i=1}^{n-1} \frac{(u_k^{(i)})^2}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2} \cdot \gamma_1 \qquad \text{(by Eq. (14))} \\
&= \gamma_n + \left(1 - \frac{(u_k^{(n)})^2}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2}\right) \cdot \gamma_1 \\
&\le \gamma_n + \frac{\gamma_n}{\gamma_1}\epsilon \cdot \gamma_1 \qquad \text{(by Eq. (18))} \\
&= \gamma_n \cdot (1+\epsilon).
\end{aligned}$$

Finally we note that $\frac{u_k^\top \Gamma u_k}{\sum_{i=1}^n \left(u_k^{(i)}\right)^2} \ge \gamma_n$ since γ_n is the smallest in $\{\gamma_i\}_{i=1}^n$. ■

Theorem 8 (Theorem 4 second part, formal version) *Suppose $\lambda_n + 2n\iota < \lambda_{n-1}$. Suppose u_0 is away from 0. Consider the GD iterates given by Eq. (13) with learning rate scheme*

$$\eta_k \in \left(0,\ \frac{n}{2\lambda_1 + 2n\iota}\right).$$

Then for $\epsilon \in (0,1)$, if $k \ge \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$,, then GD outputs an M-suboptimal solution, where $M = \frac{\gamma_1}{\gamma_n}(1-\epsilon) > 1$ is a constant.

Proof Consider an α-level set where

$$\alpha = L_{\mathcal{S}}(u_k) = \frac{1}{n} u_k^\top \Gamma u_k. \tag{19}$$

From Lemma 15 we know $L_{\mathcal{S}}(u_k)$ is monotonic decreasing thus GD cannot terminate before the k-epoch, i.e., the output of GD is u_k.

Thus

$$\begin{aligned}\frac{\Delta(u)}{\Delta_*} &= \gamma_1 \frac{\sum_{i=1}^n \left(u_k^{(i)}\right)^2}{n\alpha} && \text{(by Lemma 14)}\\ &= \gamma_1 \cdot \frac{\sum_{i=1}^n \left(u_k^{(i)}\right)^2}{u_k^\top \Gamma u_k} && \text{(by Eq. (19))}\\ &\geq \gamma_1 \cdot \frac{1}{(1+\epsilon)\gamma_n} && \text{(by Theorem 7)}\\ &\geq \frac{\gamma_1}{\gamma_n}(1-\epsilon)\\ &=: M,\end{aligned}$$

where we have $M > 1$ by letting $\epsilon < 1 - \frac{\gamma_n}{\gamma_1}$. ■

B.3. The directional bias of SGD with small learning rate

We analyze SGD with small learning rate by repeating the arguments in previous two sections.

Let us denote $X_{-n} := (x_1, x_2, \ldots, x_{n-1})$ and

$$P_{-n} = X_{-n}(X_{-n}^\top X_{-n})^{-1} X_{-n}^\top$$
$$P_n = P - P_{-n}.$$

That is, P_{-n} is the projection onto the column space of X_{-n} and P_n is the projection onto the orthogonal complement of the column space of X_{-n} with respect to the column space of X.

Let us reload

$$H := XX^\top,$$
$$H_{-n} := (P_{-n}X)(P_{-n}X)^\top,$$
$$H_n := (P_nX)(P_nX)^\top,$$
$$H_c := (P_{-n}x_n)(P_nx_n)^\top + (P_nx_n)(P_{-n}x_n)^\top.$$

Then

$$H = H_{-n} + H_n + H_c.$$

Following a routine check we can reload the following lemmas.

Lemma 16 (Variant of Lemma 10) *Suppose $3n\iota < \lambda_n$. Suppose $0 < \eta < \frac{b}{\lambda_1 + 3n\iota}$. Let $\pi := \{\mathcal{B}_1, \ldots, \mathcal{B}_m\}$ be a uniform m partition of index set $[n]$, where $n = mb$. Consider the following $d \times d$ matrix*

$$\mathcal{M}_{-n} := \prod_{j=1}^m \left(I - \frac{2\eta}{b} P_{-n} H(\mathcal{B}_j) P_{-n}\right) \in \mathbb{R}^{d\times d}.$$

Then for the spectrum of $\mathcal{M}_{-n}^\top \mathcal{M}_{-n}$ we have:

- 1 *is an eigenvalue of* $\mathcal{M}_{-n}^{\top}\mathcal{M}_{-n}$ *with multiplicity being* $d-n+1$*; moreover, the corresponding eigenspace is the column space of* $P_n + P_{\perp}$.
- *Restricted in the column space of* P_{-n}*, the eigenvalues of* $\mathcal{M}_{-n}^{\top}\mathcal{M}_{-n}$ *are upper bounded by* $\left(q_{-n}(\eta)\right)^2 < 1$*, where*

$$q_{-n}(\eta) := \max\left\{\left|1-\frac{2\eta}{b}(\lambda_1+n\iota)\right| + \frac{3n\eta\iota}{b},\ \left|1-\frac{2\eta}{b}(\lambda_{n-1}-n\iota)\right| + \frac{3n\eta\iota}{b}\right\} < 1.$$

Proof This is by a routine check of the proof of Lemma 10. ■

Consider the projections of v_k onto the column space of P_{-n} and P_n. For simplicity we reload the following notations

$$A_k := \left\|P_{-n}v_k\right\|_2, \quad B_k := \left\|P_n v_k\right\|_2.$$

The following lemma controls the update of A_k and B_k.

Lemma 17 (Variant of Lemma 11) *Suppose* $3n\iota < \lambda_n$*. Suppose* $0 < \eta < \frac{b}{\lambda_1+3n\iota}$*. Consider the* k*-th epoch of SGD iterates given by Eq. (8). Set the learning rate in this epoch to be constant* η*. Denote*

$$\xi(\eta) := \frac{4\eta\sqrt{n\iota}}{b},$$
$$q_n(\eta) := \left|1-\frac{2\eta\lambda_n}{b}\left\|P_n\bar{x}_n\right\|_2^2\right|,$$
$$q_{-n}(\eta) := \max\left\{\left|1-\frac{2\eta}{b}(\lambda_1+n\iota)\right| + \frac{3n\eta\iota}{b},\ \left|1-\frac{2\eta}{b}(\lambda_{n-1}-n\iota)\right| + \frac{3n\eta\iota}{b}\right\} < 1.$$

Then we have the following:

- $A_{k+1} \le q_{-n}(\eta)\cdot A_k + \xi(\eta)\cdot B_k$.
- $B_{k+1} \le q_n(\eta)\cdot B_k + \xi(\eta)\cdot A_k$.
- $B_{k+1} \ge q_n(\eta)\cdot B_k - \xi(\eta)\cdot A_k$.

Proof This is by a routine check of the proof of Lemma 11. ■

Lemma 18 (Variant of Lemma 13) *Suppose* $3n\iota < \lambda_n$ *and* $\lambda_n + 4n\iota < \lambda_{n-1}$*. Consider the SGD iterates given by Eq. (8) with the following small learning rate scheme*

$$\eta_k = \eta' \in \left(0,\ \frac{b}{2\lambda_1+2n\iota}\right), \quad k = 1,\dots,k_2.$$

Then for $0<\epsilon<1$ *satisfying* $\sqrt{n\iota} \le \mathrm{poly}\left(\epsilon\right)$*, if* $k_2 \ge \mathcal{O}\left(\log\frac{1}{\epsilon}\right)$*, then* $\frac{A_{k_2}}{B_{k_2}} \le \epsilon$.

Proof From the assumption we have $\eta' < \frac{b}{2(\lambda_1+n\iota)}$ and $\eta' < \frac{b}{2\lambda_n}$, thus

$$\begin{aligned}
\xi &:= \xi(\eta') = \frac{4\eta'\sqrt{n}\iota}{b}, \\
q_n &:= q_n(\eta') = \left|1 - \frac{2\eta'\lambda_n}{b}\left\|P_n\bar{x}_n\right\|_2^2\right| \\
&= 1 - \frac{2\eta'\lambda_n}{b}\left\|P_n\bar{x}_n\right\|_2^2 \\
&\le 1 - \frac{2\lambda_n(1-4n\iota^2)}{b}\eta', \qquad (\text{since } \|P_n\bar{x}_n\|_2^2 \ge 1 - 4n\iota^2 \text{ by reloading Lemma 3 }) \\
&< 1 \\
q_{-n} &:= q_{-n}(\eta') = \max\left\{\left|1 - \frac{2\eta'}{b}(\lambda_1 + n\iota)\right| + \frac{3n\eta'\iota}{b},\ \left|1 - \frac{2\eta'}{b}(\lambda_{n-1} - n\iota)\right| + \frac{3n\eta'\iota}{b}\right\} \\
&= \max\left\{1 - \frac{2\eta'}{b}(\lambda_1 + n\iota) + \frac{3n\eta'\iota}{b},\ 1 - \frac{2\eta'}{b}(\lambda_{n-1} - n\iota) + \frac{3n\eta'\iota}{b}\right\} \\
&= 1 - \frac{2\eta'}{b}(\lambda_{n-1} - n\iota) + \frac{3n\eta'\iota}{b} \\
&= 1 - \frac{2(\lambda_{n-1} - n\iota) - 3n\iota}{b}\eta' \in (0,1).
\end{aligned}$$

Moreover, by the gap assumption $\lambda_n + 4n\iota < \lambda_{n-1}$ we have

$$\begin{aligned}
q_n - q_{-n} &\ge \eta'\left(\frac{2(\lambda_{n-1} - n\iota) - 3n\iota}{b} - \frac{2\lambda_n(1-4n\iota^2)}{b}\right) \\
&\ge \frac{2\eta'}{b}\left(\lambda_{n-1} - \lambda_n - 3n\iota\right) \\
&> 0.
\end{aligned}$$

Therefore $0 < q_{-n} < q_n < 1$. Thus we can set $\xi = \frac{4\eta'\sqrt{n}\iota}{b}$ to be small such that

$$0 < q := \frac{q_{-n}}{q_n - \xi \cdot A_0/B_0} < 1. \tag{20}$$

Moreover, since $\sqrt{n}\iota \le \text{poly}\,(\epsilon)$ and $\xi = \frac{4\eta'\sqrt{n}\iota}{b}$, we have

$$\frac{\xi}{q_n - \xi \cdot A_0/B_0} \le \frac{(1-q)\epsilon}{2}. \tag{21}$$

Now we recursively show $\frac{A_k}{B_k} \le \frac{A_0}{B_0}$. Clearly it holds for $k=0$. Suppose $\frac{A_k}{B_k} \le \frac{A_0}{B_0}$, we consider $\frac{A_{k+1}}{B_{k+1}}$ in the following

$$\begin{aligned}
\frac{A_{k+1}}{B_{k+1}} &\le \frac{q_{-n}\cdot A_k + \xi \cdot B_k}{q_n \cdot B_k - \xi \cdot A_k} \qquad \text{(by Lemma 17)} \\
&= \frac{q_{-n}\cdot \frac{A_k}{B_k} + \xi}{q_n - \xi \cdot \frac{A_k}{B_k}} \\
&\le \frac{q_{-n}\frac{A_k}{B_k} + \xi}{q_n - \xi \cdot A_0/B_0} \qquad \text{(by inductive assumption)} \\
&= \frac{q_{-n}}{q_n - \xi \cdot A_0/B_0}\frac{A_k}{B_k} + \frac{\xi}{q_n - \xi \cdot A_0/B_0} \\
&\le q \cdot \frac{A_k}{B_k} + \frac{(1-q)\epsilon}{2} \qquad \text{(by Eq. (20) and (21))} \\
&\le q \cdot \frac{A_0}{B_0} + \frac{(1-q)\epsilon}{2} \\
&\le \frac{A_0}{B_0},
\end{aligned}$$

where in the last inequality we assume $\frac{\epsilon}{2} < \frac{A_0}{B_0}$.

Moreover, from the above we have

$$\frac{A_{k+1}}{B_{k+1}} \le q \cdot \frac{A_k}{B_k} + \frac{(1-q)\epsilon}{2},$$

which implies

$$\begin{aligned}
\frac{A_{k_2}}{B_{k_2}} &\le q^{k_2} \cdot \frac{A_0}{B_0} + \frac{1}{1-q} \cdot \frac{(1-q)\epsilon}{2}, \\
&\le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

where we set $k_2 \ge \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$.

■

Next we prove the directional bias of SGD with small learning rate.

Theorem 9 (Theorem 3, formal version) *Suppose $3n\iota < \lambda_n$ and $\lambda_n + 4n\iota < \lambda_{n-1}$. Suppose v_0 is away from 0. Consider the SGD iterates given by Eq. (8) with the following small learning rate scheme*

$$\eta_k = \eta' \in \left(0,\ \frac{b}{2\lambda_1 + 2n\iota}\right), \quad k = 1, \dots, k_2.$$

Then for $0 < \epsilon < 1$ satisfying $\sqrt{n}\iota \le \mathrm{poly}(\epsilon)$, if $k_2 \ge \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$, then

$$\gamma_n \le \frac{v_{k_2}^\top H v_{k_2}}{\|P v_{k_2}\|_2^2} \le (1+\epsilon)\cdot \gamma_n.$$

Proof First by Lemma 18 we have

$$\frac{B_{k_2}^2}{A_{k_2}^2+B_{k_2}^2}=\frac{1}{\frac{A_{k_2}^2}{B_{k_2}^2}+1}\geq\frac{1}{\epsilon^2+1}\geq 1-\epsilon^2. \tag{22}$$

Next by $H=H_n+H_{-n}+H_c$ we obtain

$$\begin{aligned}
\frac{v_{k_2}^\top H v_{k_2}}{\|Pv_{k_2}\|_2^2}&=\frac{(P_n v_{k_2})^\top H_n(P_n v_{k_2})}{\|Pv_{k_2}\|_2^2}+\frac{(P_{-n} v_{k_2})^\top H_{-n}(P_{-n} v_{k_2})}{\|Pv_{k_2}\|_2^2}+\frac{(Pv_{k_2})^\top H_c(Pv_{k_2})}{\|Pv_{k_2}\|_2^2}\\
&\leq \lambda_n\cdot\frac{\|P_n v_{k_2}\|_2^2}{\|Pv_{k_2}\|_2^2}+(\lambda_1+n\iota)\cdot\frac{\|P_{-n} v_{k_2}\|_2^2}{\|Pv_{k_2}\|_2^2}+4\sqrt{n\iota} \qquad \text{by reloading Lemma 5, 6, 7}\\
&\leq \lambda_n+(\lambda_1+n\iota)\cdot\frac{A_{k_2}^2}{A_{k_2}^2+B_{k_2}^2}+4\sqrt{n\iota}\\
&\leq \gamma_n+n\iota+(\lambda_1+n\iota)\cdot\epsilon^2+4\sqrt{n\iota} \qquad \text{by reloading Lemma 4 and Eq. (22)}\\
&\leq \gamma_n+\gamma_n\cdot\epsilon. \qquad \text{since } \sqrt{n\iota}\leq \text{poly}\,(\epsilon)
\end{aligned}$$

Finally, we note $\frac{v_{k_2}^\top H v_{k_2}}{\|Pv_{k_2}\|_2^2}\geq\gamma_n$ since γ_n is the smallest eigenvalue of H restricted in the column space of P. ■

Theorem 10 (Theorem 4 third part, formal version) *Suppose $3n\iota<\lambda_n$ and $\lambda_n+4n\iota<\lambda_{n-1}$. Suppose v_0 is away from 0. Consider the SGD iterates given by Eq. (8) with the following small learning rate scheme*

$$\eta_k=\eta'\in\left(0,\ \frac{b}{2\lambda_1+2n\iota}\right),\quad k=1,\dots,k_2.$$

Then for $0<\epsilon<1$ such that $\sqrt{n\iota}\leq\text{poly}\,(\epsilon)$, if $k_2\geq\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$, then SGD outputs an M-suboptimal solution where $M=\frac{\gamma_1}{\gamma_n}(1-\epsilon)>1$ is a constant.

Proof From Eq. (8) and $\eta'<\frac{1}{2\lambda_1}$ we know that restricted in the column space of P, the eigenvalues of $\mathcal{M}_\pi$ is smaller than 1, thus v_k indeed converges to 0.

Consider an α-level set where

$$\alpha=L_{\mathcal{S}}(v_k)=\frac{1}{n}v_{k_2}^\top H v_{k_2}. \tag{23}$$

Then

$$\begin{aligned}
\frac{\Delta(u)}{\Delta_*}&=\gamma_1\frac{\|Pv_k\|_2^2}{n\alpha} \qquad \text{(by Lemma 8)}\\
&=\gamma_1\cdot\frac{\|Pv_k\|_2^2}{v_k^\top H v_k} \qquad \text{(by Eq. (23))}\\
&\geq\gamma_1\cdot\frac{1}{(1+\epsilon)\gamma_n} \qquad \text{(by Theorem 9)}\\
&\geq\frac{\gamma_1}{\gamma_n}(1-\epsilon)\\
&=:M,
\end{aligned}$$

where we have $M > 1$ by letting $\epsilon < 1 - \frac{\gamma_n}{\gamma_1}$. ■

Appendix C. Proof of Auxiliary Lemmas in Sections A and B

C.1. Proof of Lemma 1

Proof [Proof of Lemma 1] Note that $\bar{x}_i$ follows uniform distribution on the sphere $\mathcal{S}^{d-1}$. Therefore, let ξ be a random variable following distribution χ^2_d distribution and define $z_i = \xi \cdot \bar{x}_i$, we have z_i follows standard normal distribution in the d-dimensional space. Then it suffices to prove that $|\langle z_i, z_j\rangle|/(\|z_i\|_2\|z_j\|_2) \le \iota$ for all $i \neq j$.

First we will bound the inner product $\langle z_i, z_j\rangle$. Note that we have each entry in z_i is 1-subgaussian, it can be direcly deduced that

$$\langle z_i, z_j\rangle = \sum_{k=1}^{d} z_i^{(k)} z_j^{(k)} = \sum_{k=1}^{d} \left(\left(\frac{z_i^{(k)} + z_j^{(k)}}{2} \right)^2 - \left(\frac{z_i^{(k)} - z_j^{(k)}}{2} \right)^2 \right)$$

is d-subexponential, where $z_i^{(k)}$ denotes the k-th of the vector z_i. Then if follows that

$$\mathbb{P}\left(|\langle z_i, z_j\rangle| \ge t \right) \le 2\exp\left(-\frac{t^2}{d} \right).$$

Next we will lower bound $\|z_i\|_2$. Note that

$$\|z_i\|_2^2 - d = \sum_{k=1}^{d} \left(\left(z_i^{(k)} \right)^2 - 1 \right).$$

Since $z_i^{(k)}$ is 1-subgaussian, we have $\|z_i\|^2 - d$ is d-subexpoential, then

$$\mathbb{P}\left(\left| \|z_i\|_2^2 - d \right| \ge t \right) \le 2\exp\left(-\frac{t^2}{d} \right).$$

Finally, applying the union bound for all possible $i, j \in [n]$, we have with probability at least $1 - \delta$, the following holds for all $i \neq j$,

$$|\langle z_i, z_j\rangle| \le \sqrt{d \log \frac{2n^2}{\delta}},$$
$$\|z_i\|_2^2 \ge d - \sqrt{d \log \frac{2n^2}{\delta}}.$$

Assume $d \ge 4\log(2n^2/\delta)$, we have $\|z_i\|^2 \ge d/2$. Then it follows that

$$|\langle \bar{x}_i, \bar{x}_j\rangle| = \frac{|\langle z_i, z_j\rangle|}{\|z_i\|_2\|z_j\|_2} < 2\sqrt{\frac{1}{d} \log \frac{2n^2}{\delta}} =: \iota.$$

This completes the proof. ■

C.2. Proof of Lemma 3

Proof [Proof of Lemma 3] Similar to the proof of Lemma 1, we consider translating $x_1, \ldots, x_n$ to $z_1, \ldots, z_n$ by introducing χ^2_d random variables. Let $Z_{-1} = (z_2, \ldots, z_n) \in \mathbb{R}^{d\times(n-1)}$, in which each entry is i.i.d. generated from Gaussian distribution $\mathcal{N}(0,1)$. Then we have

$$P_{-1}\bar{x}_1 = X_{-1}(X_{-1}^\top X_{-1})^{-1}X_{-1}^\top \bar{x}_1 = Z_{-1}(Z_{-1}^\top Z_{-1})^{-1}Z_{-1}^\top \bar{x}_1.$$

Then conditioned on $\bar{x}_1$, we have each entry in $Z_{-1}^\top \bar{x}_1$ i.i.d. follows $\mathcal{N}(0,1)$. Then it is clear that $\|Z_{-1}^\top \bar{x}_1\|_2^2$ follows from χ^2_{n-1} distribution, implying that with probability at least $1-\delta'$, we have

$$\|Z_{-1}^\top \bar{x}_1\|_2^2 \le (n-1) + \sqrt{(n-1)\log(2/\delta')}.$$

Then by Corollary 5.35 in Vershynin [33], we know that with probability at least $1-\delta'$ it holds that

$$\sqrt{d} - \sqrt{n-1} - \sqrt{2\log(2/\delta')} \le \sigma_{\min}(Z_{-1}) \le \sigma_{\max}(Z_{-1}) \le \sqrt{d} + \sqrt{n-1} + \sqrt{2\log(2/\delta')}.$$

Therefore, assume $\sqrt{(n-1)} + \sqrt{2\log(2/\delta')} \le \sqrt{d}/8$, we have with probability at least $1-\delta'$

$$\begin{aligned}\left\|Z_{-1}(Z_{-1}^\top Z_{-1})^{-1}\right\|_2 &\le \frac{\sqrt{d} + \sqrt{n-1} + \sqrt{2\log(2/\delta')}}{\left(\sqrt{d} - \sqrt{n-1} - \sqrt{2\log(2/\delta')}\right)^2} \\ &\le \frac{1}{\sqrt{d}}\left(1 + 4\left(\sqrt{\frac{n-1}{d}} + \sqrt{\frac{2\log(2/\delta')}{d}}\right)\right).\end{aligned}$$

Combining with the upper bound of $\|Z_{-1}^\top \bar{x}_1\|_2$, set $\delta' = \delta/2$, we have with probability at least $1-\delta$ that

$$\begin{aligned}\|P_{-1}\bar{x}_1\|_2 &\le \left\|Z_{-1}(Z_{-1}^\top Z_{-1})^{-1}\right\|_2 \cdot \|X_{-1}^\top \bar{x}_1\|_2 \\ &\le \left(1 + 4\left(\sqrt{\frac{n-1}{d}} + \sqrt{\frac{2\log(4/\delta)}{d}}\right)\right) \cdot \left(\sqrt{\frac{n-1}{d}} + \sqrt{\frac{\sqrt{(n-1)\log(4/\delta)}}{d}}\right) \\ &\le (1 + 4\sqrt{n\iota}) \cdot \sqrt{n\iota},\end{aligned}$$

where the last inequality follows from the definition of ι. Then assume $\sqrt{n\iota} \le 1/4$, we are able to completes the proof of the first argument. Note that $\|P_1\bar{x}_1\|_2^2 + \|P_{-1}\bar{x}_1\|_2^2 = \|\bar{x}_1\|_2^2 = 1$, we have

$$\|P_{-1}\bar{x}_1\|_2 = \sqrt{1 - \|P_1\bar{x}_1\|_2^2} \ge \sqrt{1 - 4n\iota^2} \ge 1 - 4n\iota^2.$$

This completes the proof of the second argument. The third argument holds trivially by the construction of $P_\perp$.

■

C.3. Proof of Lemma 4

Proof [Proof of Lemma 4] Clearly $XX^\top \in \mathbb{R}^{d\times d}$ is of rank n and symmetric, thus $XX^\top$ has n real, non-zero (potentially repeated) eigenvalues, denoted as $\gamma_1, \dots, \gamma_n$ in non-decreasing order. Moreover, $\gamma_1, \dots, \gamma_n$ are also eigenvalues of $X^\top X \in \mathbb{R}^{n\times n}$, thus it is sufficient to locate the eigenvalues of $X^\top X$, where $\left(X^\top X\right)_{ij} = x_i^\top x_j$.

We first calculate the diagonal entry

$$\left(X^\top X\right)_{ii} = x_i^\top x_i = \lambda_i.$$

Then we bound the off diagonal entries. For $j \neq i$,

$$\left(X^\top X\right)_{ij} = x_i^\top x_j = \sqrt{\lambda_i\lambda_j}\langle \bar{x}_i, \bar{x}_j\rangle \in (-\iota, \iota)\,,$$

where we use $0 < \lambda_1, \dots, \lambda_n \leq 1$. Thus we have

$$R_i(X^\top X) = \sum_{j\neq i}\left|\left(X^\top X\right)_{ij}\right| \leq n\iota, \quad i = 1, \dots, n,$$

Finally our conclusions hold by applying Gershgorin circle theorem. ■

C.4. Proof of Lemma 5

Proof [Proof of Lemma 5] The first conclusion is clear since by construction, we have $P_{-1}P_1 = P_{-1}P_\perp = 0$.

Note that H_{-1} is a rank $n-1$ symmetric matrix. Let $\tau_2, \dots, \tau_n$ be the $n-1$ non-zero eigenvalues of H_{-1}. Clearly, $\tau_2, \dots, \tau_n$ with $\tau_1 := 0$ give the spectrum of

$$H'_{-1} := (P_{-1}X)^\top P_{-1}X \in \mathbb{R}^{n\times n}.$$

We then bound $\tau_2, \dots, \tau_n$ by analyzing H'_{-1}.

From Lemma 3 we have $\|P_{-1}\bar{x}_1\|_2 \leq 2\sqrt{n}\iota$. From Lemma 2 we have

$$P_{-1}X = (P_{-1}x_1, P_{-1}x_2, \dots, P_{-1}x_n) = (P_{-1}x_1, x_2, \dots, x_n)\,.$$

Then we calculate the diagonal entries:

$$\left(H'_{-1}\right)_{ii} = \begin{cases} \|P_{-1}x_1\|_2^2 \leq \lambda_1 \cdot 4n\iota^2 \leq 4n\iota^2, & i = 1; \\ \|x_i\|_2^2 = \lambda_i, & i \neq 1. \end{cases}$$

Then we bound the off diagonal entries. Let $j \neq i$. Then at least one of them is not 1. Without loss of generality let $i \neq 1$, which yields $x_i = P_{-1}x_i$ by Lemma (2). Thus $\langle x_i, P_1x_j\rangle = \langle P_{-1}x_i, P_1x_j\rangle = 0$. Thus we have

$$\begin{aligned}\left(H'_{-1}\right)_{ij} &= (P_{-1}x_i)^\top P_{-1}x_j \\ &= x_i^\top P_{-1}x_j \\ &= x_i^\top x_j - x_i^\top P_1 x_j \\ &= x_i^\top x_j \\ &= \sqrt{\lambda_i\lambda_j}\cdot\langle \bar{x}_i, \bar{x}_j\rangle \\ &\in (-\iota, \iota)\,.\end{aligned}$$

Thus we have

$$R_i(H'_{-1}) = \sum_{j \neq i} \left| \left(H'_{-1}\right)_{ij} \right| \leq n\iota, \quad i = 1, \ldots, n.$$

Finally, we set $4n\iota^2 + 2n\iota < \lambda_n$, so that the first Geoshgorin disc does not intersect with the others, then Gershgorin circle theorem gives our second conclusion. ■

C.5. Proof of Lemma 8

Proof [Proof of Lemma 8] For the empirical loss, it is clear that

$$\begin{aligned} L_{\mathcal{S}}(v) &= \frac{1}{n} \left(w - w_*\right)^\top XX^\top \left(w - w_*\right) = \frac{1}{n} v^\top XX^\top v = \frac{1}{n} v^\top Hv \\ &= \frac{1}{n} (Pv)^\top H(Pv) \\ &= \frac{1}{n} (Pv)^\top H_1(Pv) + \frac{1}{n} (Pv)^\top H_{-1}(Pv) + \frac{1}{n} (Pv)^\top H_c(Pv) \\ &= \frac{1}{n} (P_1 v)^\top H_1(P_1 v) + \frac{1}{n} (P_{-1} v)^\top H_{-1}(P_{-1} v) + \frac{1}{n} (Pv)^\top H_c(Pv), \end{aligned}$$

where we use Lemma 4, Lemma 5, and Lemma 6. For the population loss,

$$L_{\mathcal{D}}(v) = \mu \left\| w - w_* \right\|_2^2 = \mu \left\| v \right\|_2^2.$$

For the hypothesis class $\mathcal{H}_{\mathcal{S}} = \left\{ w \in \mathbb{R}^d : P_\perp w = P_\perp w_0 \right\}$, applying $w - w_* = v$ and $w_0 - w_* = v_0$, we obtain

$$\mathcal{H}_{\mathcal{S}} = \left\{ v \in \mathbb{R}^d : P_\perp v = P_\perp v_0 \right\}.$$

For the α-level set, we note the optimal training loss is $L_{\mathcal{S}}^* = \inf_{v \in \mathcal{H}_{\mathcal{S}}} L_{\mathcal{S}}(v) = 0$.

As for the estimation error, we note that $\inf_{v \in \mathcal{H}_{\mathcal{S}}} L_{\mathcal{D}}(v) = \inf_{P_\perp v = P_\perp v_0} \mu \left\| v \right\|_2^2 = \mu \left\| P_\perp v_0 \right\|_2^2$. thus for $v \in \mathcal{H}_{\mathcal{S}}$, we have

$$\Delta(v) = L_{\mathcal{D}}(v) - \inf_{v' \in \mathcal{V}} L_{\mathcal{D}}(v') = \mu \left\| v \right\|_2^2 - \mu \left\| P_\perp v_0 \right\|_2^2 = \mu \left\| Pv \right\|_2^2.$$

Finally, consider $v \in \mathcal{V}$, i.e., $n\alpha = v^\top XX^\top v$, thus

$$\Delta_* = \inf_{v \in \mathcal{V}} \Delta(v) = \inf_{n\alpha = v^\top XX^\top v} \mu \left\| Pv \right\|_2^2 = \frac{\mu n \alpha}{\gamma_1},$$

where γ_1 is the largest eigenvalue of the matrix $XX^\top$ and the inferior is attended by setting v parallel to the first eigenvector of $XX^\top$. ■

C.6. Proof of Lemma 9

Proof [Proof of Lemma 9] From Eq. (7) we have

$$v_{k,j+1} = \left(I - \frac{2\eta}{b} H(\mathcal{B}_j) \right) v_{k,j}, \quad j = 1, \ldots, m. \tag{24}$$

Recall the following property of projection operators:

$$P_1 = P_1P_1, \quad P_{-1} = P_{-1}P_{-1}$$
$$0 = P_1P_{-1} = P_{-1}P_1.$$

Moreover since $x_i^\top P_\perp v = 0$, we have

$$H(\mathcal{B}_j)P_\perp v = \sum_{i\in\mathcal{B}_j} x_i x_i^\top P_\perp v = 0.$$

Applying P_1 to Eq. (24) we have

$$\begin{aligned}
P_1 v_{k,j+1} &= P_1\left(I - \frac{2\eta}{b}H(\mathcal{B}_j)\right)v_{k,j} \\
&= P_1\left(I - \frac{2\eta}{b}H(\mathcal{B}_j)\right)(P_1 v_{k,j} + P_{-1}v_{k,j} + P_\perp v_{k,j}) \\
&= P_1\left(I - \frac{2\eta}{b}H(\mathcal{B}_j)\right)P_1 v_{k,j} + P_1\left(I - \frac{2\eta}{b}H(\mathcal{B}_j)\right)P_{-1}v_{k,j} \\
&= \left(I - \frac{2\eta}{b}P_1H(\mathcal{B}_j)P_1\right)\cdot P_1 v_{k,j} - \left(\frac{2\eta}{b}P_1H(\mathcal{B}_j)P_{-1}\right)\cdot P_{-1}v_{k,j}.
\end{aligned}$$

Similarly applying P_{-1} to Eq. (24) we have

$$\begin{aligned}
P_{-1} v_{k,j+1} &= P_{-1}\left(I - \frac{2\eta}{b}H(\mathcal{B}_j)\right)v_{k,j} \\
&= P_{-1}\left(I - \frac{2\eta}{b}H(\mathcal{B}_j)\right)(P_1 v_{k,j} + P_{-1}v_{k,j} + P_\perp v_{k,j}) \\
&= P_{-1}\left(I - \frac{2\eta}{b}H(\mathcal{B}_j)\right)P_1 v_{k,j} + P_{-1}\left(I - \frac{2\eta}{b}H(\mathcal{B}_j)\right)P_{-1}v_{k,j} \\
&= -\left(\frac{2\eta}{b}P_{-1}H(\mathcal{B}_j)P_1\right)\cdot P_1 v_{k,j} + \left(I - \frac{2\eta}{b}P_{-1}H(\mathcal{B}_j)P_{-1}\right)\cdot P_{-1}v_{k,j}.
\end{aligned}$$

To sum up we have

$$\begin{pmatrix} P_1 v_{k,j+1} \\ P_{-1}v_{k,j+1}\end{pmatrix} = \begin{pmatrix} I - \frac{2\eta}{b}P_1H(\mathcal{B}_j)P_1 & -\frac{2\eta}{b}P_1H(\mathcal{B}_j)P_{-1} \\ -\frac{2\eta}{b}P_{-1}H(\mathcal{B}_j)P_1 & I - \frac{2\eta}{b}P_{-1}H(\mathcal{B}_j)P_{-1}\end{pmatrix}\cdot\begin{pmatrix} P_1 v_{k,j} \\ P_{-1}v_{k,j}\end{pmatrix}$$

Notice that if $1 \notin \mathcal{B}_j$, i.e., x_1 is not used in the j-th step, then we claim

$$P_1H(\mathcal{B}_j) = H(\mathcal{B}_j)P_1 = 0,$$

since $H(\mathcal{B}_j) = \sum_{i\in\mathcal{B}_j} x_i x_i^\top$ is composed by the data belonging to the column space of P_{-1}. Therefore if $1 \notin \mathcal{B}_j$ we have

$$\begin{pmatrix} P_1 v_{k,j+1} \\ P_{-1}v_{k,j+1}\end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I - \frac{2\eta}{b}P_{-1}H(\mathcal{B}_j)P_{-1}\end{pmatrix}\cdot\begin{pmatrix} P_1 v_{k,j} \\ P_{-1}v_{k,j}\end{pmatrix}$$

■

C.7. Proof of Lemma 10

Proof [Proof of Lemma 10] Clearly for each component in the production, the column space of $P_1 + P_\perp$, which is $(n-d+1)$-dimensional, belongs to its eigenspace of eigenvalue 1, which yields the first claim.

In the following, we restrict ourselves in the column space of P_{-1}. Let us expand $\mathcal{M}_{-1}$:

$$\begin{aligned}
\mathcal{M}_{-1} &= \prod_{j=1}^{m} \big(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1}\big) \\
&= \big(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_m) P_{-1}\big) \cdots \big(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_1) P_{-1}\big) \\
&= I - \underbrace{\frac{2\eta}{b} \sum_{j=1}^{m} P_{-1} H(\mathcal{B}_j) P_{-1}}_{H_{-1}} \\
&\qquad + \underbrace{\left(\frac{2\eta}{b}\right)^2 \sum_{1 \le i < j \le n} P_{-1} H(\mathcal{B}_j) P_{-1} H(\mathcal{B}_i) P_{-1} + \ldots.}_{C}
\end{aligned}$$

We first analyze matrix H_{-1}. Since $H(\mathcal{B}_j) = \sum_{i \in \mathcal{B}_j} x_i x_i^\top$ and $\pi = \{\mathcal{B}_1, \ldots, \mathcal{B}_m\}$ is a partition for index set $[n]$, we have

$$\begin{aligned}
H_{-1} &= \sum_{j=1}^{m} P_{-1} H(\mathcal{B}_j) P_{-1} \\
&= \sum_{j=1}^{m} P_{-1} \sum_{i \in \mathcal{B}_j} x_i x_i^\top P_{-1} \\
&= P_{-1} \sum_{i=1}^{n} x_i x_i^\top P_{-1} \\
&= P_{-1} X X^\top P_{-1},
\end{aligned}$$

which is exactly the matrix we studied in Lemma 5, and from where we have H_{-1} has eigenvalue zero (with multiplicity being $n-d+1$) in the column space of $P_1 + P_\perp$, and restricted in the column space of P_{-1}, the eigenvalues of H_{-1} belong to $(\lambda_n - n\iota, \lambda_2 + n\iota)$.

Then we analyze matrix C.

$$\begin{aligned}
P_{-1} H(\mathcal{B}_j) P_{-1} H(\mathcal{B}_i) P)_{-1} &= \left(P_{-1} \sum_{i' \in \mathcal{B}_i} x_{i'} x_{i'}^\top P_{-1}\right) \left(P_{-1} \sum_{j' \in \mathcal{B}_j} x_{j'} x_{j'}^\top P_{-1}\right) \\
&= \sum_{i' \in \mathcal{B}_i} \sum_{j' \in \mathcal{B}_j} (P_{-1} x_{i'}) \langle P_{-1} x_{i'}, P_{-1} x_{j'} \rangle (P_{-1} x_{j'})^\top. \qquad (25)
\end{aligned}$$

Remember that $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ for $i \neq j$, thus $x_{i'} \neq x_{j'}$ for $i' \in \mathcal{B}_i$ and $j' \in \mathcal{B}_j$. Then from Lemma 1 we have,

$$\big|\langle P_{-1} x_{i'}, P_{-1} x_{j'} \rangle\big| \le \big|\langle x_{i'}, x_{j'} \rangle\big| \le \sqrt{\lambda_{i'} \lambda_{j'}} \cdot \iota \le \iota.$$

Inserting this into Eq. (25) we obtain

$$\|P_{-1}H(\mathcal{B}_j)P_{-1}H(\mathcal{B}_i)P)_{-1}\|_F \le b^2 \cdot \max\left|\langle P_{-1}x_{i'}, P_{-1}x_{j'}\rangle\right|^2 \le b^2\iota^2.$$

We can bound the Frobenius norm of the higher degree terms in matrix C in a similar manner; in sum for the Frobenius norm of C, we have

$$\begin{aligned}
\|C\|_F &\le \sum_{s=2}^{m}\left(\frac{2\eta}{b}\right)^s \cdot b^s \cdot \iota^s \cdot \binom{m}{s} \\
&= \sum_{s=2}^{m}(2\eta\iota)^s \cdot \binom{n}{s} \\
&= \sum_{s=0}^{m}(2\eta\iota)^s \cdot \binom{n}{s} - 1 - 2m\eta\iota \\
&= (1+2\eta\iota)^m - 1 - 2m\eta\iota \\
&\le 1 + m\cdot 2\eta\iota + \frac{m^2\sqrt{e}}{2}\cdot(2\eta\iota)^2 - 1 - 2m\eta\iota \qquad (\text{for } 2\eta\iota < \frac{1}{2m}) \\
&\le 4m^2\eta^2\iota^2,
\end{aligned}$$

where for the second to the last inequality we notice that for $f(t) = (1+t)^m$ and $t \in [0, \frac{1}{2n}]$, we have $f''(t) = m(m-1)(1+t)^{m-2} \le m(m-1)(1+\frac{1}{2m})^{m-2} \le m(m-1)\cdot\sqrt{e}$, which implies $f(t)$ is $(m^2\sqrt{e})$-smooth for $t \in [0, \frac{1}{2m}]$; moreover, by the assumption that $3n\iota < \lambda_n$ and $\eta < \frac{b}{\lambda_n + 3n\iota}$, we can indeed verify that

$$2\eta\iota < \frac{2b\iota}{\lambda_n + 3n\iota} \le \frac{2b\iota}{6n\iota} \le \frac{1}{2m}. \tag{26}$$

Now we rephrase $\mathcal{M}_{-1}^\top\mathcal{M}_{-1}$ as

$$\begin{aligned}
\mathcal{M}_{-1}^\top\mathcal{M}_{-1} &= \left(I - \frac{2\eta}{b}H_{-1} + C^\top\right)\cdot\left(I - \frac{2\eta}{b}H_{-1} + C\right) \\
&= \left(I - \frac{2\eta}{b}H_{-1}\right)^2 + \underbrace{C^\top\left(I - \frac{2\eta}{b}H_{-1}\right) + \left(I - \frac{2\eta}{b}H_{-1}\right)C + C^\top C}_{D}.
\end{aligned} \tag{27}$$

Restricting ourselves in the column space of P_{-1}, the eigenvalues of H_{-1} belong to $(\lambda_n - n\iota, \lambda_2 + n\iota)$, thus the eigenvalues of $\left(I - \frac{2\eta}{b}H_{-1}\right)^2$ are upper bounded by

$$\max\left\{\left(1 - \frac{2\eta}{b}(\lambda_2 + n\iota)\right)^2,\ \left(1 - \frac{2\eta}{b}(\lambda_n - n\iota)\right)^2\right\} < 1, \tag{28}$$

where the last inequality is guaranteed by our assumptions on η and ι. For simplicity we defer the verification to the end of the proof.

Consider the following eigen decomposition $I - 2\eta H_{-1} = U\,\mathrm{diag}\,(\mu_1, \ldots, \mu_{n-1}, 1, \ldots, 1)\,U^\top$, where $\mu_1, \ldots, \mu_{n-1} \in (-1, 1)$ by Eq. (28). Then we have

$$\begin{aligned}
\|(I - \eta H_{-1})C\|_F &= \left\|\mathrm{diag}\,(\mu_1, \ldots, \mu_{n-1}, 1, \ldots, 1)\,U^\top CU\right\|_F \\
&\le \left\|U^\top CU\right\|_F = \|C\|_F.
\end{aligned}$$

Therefore we can bound the Frobenius norm of D by

$$\begin{aligned}\|D\|_F &\le 2\|(I-2\eta H_{-1})C\|_F + \|C\|_F^2 \\ &\le 2\|C\|_F + \|C\|_F^2 \\ &\le 8m^2\eta^2\iota^2 + 16m^4\eta^4\iota^4 \\ &\le 9m^2\eta^2\iota^2, \end{aligned} \tag{29}$$

where the last inequality follows from $2\eta\iota \le 1/(2m)$ proved in Eq. (26).

Finally, applying Hoffman-Wielandt theorem with Eq. (27), (28) and (29), we conclude that, restricted in the column space of P_{-1}, the eigenvalues of $\mathcal{M}_{-1}^\top \mathcal{M}_{-1}$ are upper bounded by

$$\begin{aligned}&\max\left\{\left(1-\frac{2\eta}{b}(\lambda_2+n\iota)\right)^2 + 9m^2\eta^2\iota^2,\ \left(1-\frac{2\eta}{b}(\lambda_n-n\iota)\right)^2 + 9m^2\eta^2\iota^2\right\} \\ &\le \max\left\{\left(\left|1-\frac{2\eta}{b}(\lambda_2+n\iota)\right| + 3m\eta\iota\right)^2,\ \left(\left|1-\frac{2\eta}{b}(\lambda_n-n\iota)\right| + 3m\eta\iota\right)^2\right\} \\ &:= \left(q_{-1}(\eta)\right)^2. \end{aligned} \tag{30}$$

At this point we left to verify Eq. (28) and

$$q_{-1}(\eta) := \max\left\{\left|1-\frac{2\eta}{b}(\lambda_2+n\iota)\right| + \frac{3n\eta\iota}{b},\ \left|1-\frac{2\eta}{b}(\lambda_n-n\iota)\right| + \frac{3n\eta\iota}{b}\right\} < 1. \tag{31}$$

Clearly it suffices to verify Eq. (31).

$$\begin{aligned}&\left|1-\frac{2\eta}{b}(\lambda_2+n\iota)\right| + \frac{3n\eta\iota}{b} < 1 \\ \Leftrightarrow\quad &\frac{3n\iota}{b}\eta - 1 < 1 - \frac{2(\lambda_2+n\iota)}{b}\eta < 1 - \frac{3n\iota}{b}\eta \\ \Leftrightarrow\quad &\begin{cases}\frac{2\lambda_2-n\iota}{b}\eta > 0 \\ \frac{2\lambda_2+5n\iota}{b}\eta < 2\end{cases} \\ \Leftarrow\quad &\begin{cases}\eta > 0 \\ 2\lambda_2 - n\iota > 0 \\ \eta < \frac{b}{\lambda_2+2.5n\iota}\end{cases} \\ \Leftarrow\quad &\begin{cases}3n\iota < \lambda_n \qquad (\text{since } \lambda_2 \ge \lambda_n) \\ 0 < \eta < \frac{b}{\lambda_2+3n\iota}\end{cases}\end{aligned}$$

Similarly, we verify that

$$
\begin{aligned}
& \left|1-\frac{2\eta}{b}(\lambda_n - n\iota)\right| + \frac{3n\eta\iota}{b} < 1 \\
\Leftrightarrow \quad & \frac{3n\iota}{b}\eta - 1 < 1 - \frac{2(\lambda_n - n\iota)}{b}\eta < 1 - \frac{3n\iota}{b}\eta \\
\Leftrightarrow \quad & \begin{cases} \frac{2\lambda_n - 5n\iota}{b}\eta > 0 \\ \frac{2\lambda_n + n\iota}{b}\eta < 2 \end{cases} \\
\Leftarrow \quad & \begin{cases} \eta > 0 \\ 2\lambda_n - 5n\iota > 0 \\ \eta < \frac{b}{\lambda_n + 0.5n\iota} \end{cases} \\
\Leftarrow \quad & \begin{cases} 3n\iota < \lambda_n \\ 0 < \eta < \frac{b}{\lambda_2 + 3n\iota} \end{cases} \qquad \text{(since } \lambda_2 \geq \lambda_n\text{)}
\end{aligned}
$$

These complete our proof. ■

C.8. Proof of Lemma 11

Proof [Proof of Lemma 11]

Note that during one epoch of SGD updates, x_1 is used for only once. Without loss of generality, assume SGD uses x_1 at the l-th step, i.e., $1 \in \mathcal{B}_l$ and $1 \notin B_j$ for $j \neq l$. Recursively applying Lemma 9, we have

$$
\begin{aligned}
\begin{pmatrix} P_1 v_{k,m+1} \\ P_{-1} v_{k,m+1} \end{pmatrix} = & \begin{pmatrix} I & 0 \\ 0 & \prod_{j=l+1}^{m} \left(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1}\right) \end{pmatrix} \times \\
& \begin{pmatrix} I - \frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_1 & -\frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_{-1} \\ -\frac{2\eta}{b} P_{-1} H(\mathcal{B}_l) P_1 & I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1} \end{pmatrix} \times \\
& \begin{pmatrix} I & 0 \\ 0 & \prod_{j=1}^{l-1} \left(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1}\right) \end{pmatrix} \times \begin{pmatrix} P_1 v_{k,1} \\ P_{-1} v_{k,1} \end{pmatrix}
\end{aligned}
$$

Let $v_{k+1} = v_{k,m+1},\ v_k = v_{k,1}$ and

$$
\begin{aligned}
\mathcal{M}_l &:= I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_l) P_{-1} \\
\mathcal{M}_{>l} &:= \prod_{j=l+1}^{m} \left(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1}\right) \\
\mathcal{M}_{<l} &:= \prod_{j=1}^{l-1} \left(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1}\right) \\
\mathcal{M}_{-1} &:= \mathcal{M}_{>l} \cdot \mathcal{M}_l \cdot \mathcal{M}_{<l} = \prod_{j=1}^{m} \left(I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1}\right)
\end{aligned}
$$

then we have

$$\begin{aligned}\begin{pmatrix} P_1 v_{k+1} \\ P_{-1} v_{k+1} \end{pmatrix} &= \begin{pmatrix} I & 0 \\ 0 & \mathcal{M}_{>l} \end{pmatrix} \begin{pmatrix} I - \frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_1 & -\frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_{-1} \\ -\frac{2\eta}{b} P_{-1} H(\mathcal{B}_l) P_1 & \mathcal{M}_l \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & \mathcal{M}_{<l} \end{pmatrix} \begin{pmatrix} P_1 v_k \\ P_{-1} v_k \end{pmatrix} \\ &= \begin{pmatrix} I - \frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_1 & -\left(\frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_{-1}\right) \mathcal{M}_{<l} \\ -\mathcal{M}_{>l} \left(\frac{2\eta}{b} P_{-1} H(\mathcal{B}_l) P_1\right) & \mathcal{M}_{-1} \end{pmatrix} \begin{pmatrix} P_1 v_k \\ P_{-1} v_k \end{pmatrix} \end{aligned} \tag{32}$$

In the following we bound the norm of each entries in the above coefficient matrix.

According to Lemma 5, we have the eigenvalues of $P_{-1} H(\mathcal{B}_j) P_{-1}$ are upper bounded by $\lambda_2 + n\iota$. Thus the assumption $\eta < \frac{b}{\lambda_2 + 2n\iota}$ yields

$$\left\| I - \frac{2\eta}{b} P_{-1} H(\mathcal{B}_j) P_{-1} \right\|_2 \le 1,$$

which further yields

$$\|\mathcal{M}_{>l}\|_2 \le 1, \quad \|\mathcal{M}_{<l}\|_2 \le 1. \tag{33}$$

On the other hand notice that $P_1 x_i = 0$ for $i \neq 1$, thus

$$\begin{aligned} P_1 H(\mathcal{B}_l) P_{-1} &= P_1 \sum_{i \in \mathcal{B}_l} x_i x_i^\top P_{-1} = P_1 x_1 x_1^\top P_{-1}, \\ P_{-1} H(\mathcal{B}_l) P_1 &= P_{-1} \sum_{i \in \mathcal{B}_l} x_i x_i^\top P_1 = P_{-1} x_1 x_1^\top P_1, \end{aligned}$$

which yield

$$\max\left\{ \|P_1 H(\mathcal{B}_l) P_{-1}\|_2, \|P_{-1} H(\mathcal{B}_l) P_1\|_2 \right\} \le \|P_1 x_1\|_2 \cdot \|P_{-1} x_1\|_2 \le 2\sqrt{n}\iota, \tag{34}$$

where the last inequality is from Lemma 3 and $\lambda_1 = \|x_1\|^2 \le 1$. Eq. (33) and (34) imply

$$\max\left\{ \left\| \left(\frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_{-1}\right) \mathcal{M}_{<l} \right\|_2, \left\| \mathcal{M}_{>l} \left(\frac{2\eta}{b} P_{-1} H(\mathcal{B}_l) P_1\right) \right\|_2 \right\} \le \frac{4\eta\sqrt{n}\iota}{b} =: \xi(\eta) \tag{35}$$

Next, by $P_1 x_i = 0$ for $i \neq 1$ we have

$$P_1 H(\mathcal{B}_l) P_1 = P_1 \sum_{i \in \mathcal{B}_l} x_i x_i^\top P_1 = P_1 x_1 x_1^\top P_1 = (P_1 x_1)(P_1 x_1)^\top,$$

from where we know $\|P_1 x_1\|_2^2$ is the only non-zero eigenvalue of the rank-1 matrix $P_1 H(\mathcal{B}_l) P_1$, and the corresponding eigenspace is the column space of P_1. Therefore $1 - \frac{2\eta}{b}\|P_1 x_1\|_2^2$ is an eigenvalue of the matrix $I - \frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_1$, and the corresponding eigenspace is the column space of P_1, which implies

$$\begin{aligned} \left\| \left(I - \frac{2\eta}{b} P_1 H(\mathcal{B}_l) P_1\right) P_1 v_k \right\|_2 &= \left\| \left(1 - \frac{2\eta}{b}\|P_1 x_1\|_2^2\right) P_1 v_k \right\| \\ &= \left| 1 - \frac{2\eta}{b}\|P_1 x_1\|_2^2 \right| \cdot \|P_1 v_k\|_2 \\ &=: q_1(\eta) \cdot \|P_1 v_k\|_2 . \end{aligned} \tag{36}$$

Finally, according to Lemma 10, we have, restricted in the column space of P_{-1}, the right eigenvalues of $\mathcal{M}_{-1}$ is upper bounded by $(q_{-1}(\eta))^2$, which implies

$$\|\mathcal{M}_{-1}P_{-1}v_k\|_2 \leq q_{-1}(\eta)\cdot\|P_{-1}v_k\|_2\,. \tag{37}$$

Note we have $q_{-1}(\eta)<1$ by Lemma 10.

Combining Eq. (32) with Eq. (35), (36), (37), and letting $B_k := \|P_1 v_k\|_2\,,\ A_k := \|P_{-1}v_k\|_2$, we obtain

$$\begin{aligned}
B_{k+1} &\leq q_1(\eta)\cdot B_k + \xi(\eta)\cdot A_k\\
B_{k+1} &\geq q_1(\eta)\cdot B_k - \xi(\eta)\cdot A_k\\
A_{k+1} &\leq q_{-1}\cdot(\eta)A_k + \xi(\eta)\cdot B_k.
\end{aligned}$$

■

C.9. Proof of Lemma 12

Proof [Proof of Lemma 12] Let

$$\begin{aligned}
\xi := \xi(\eta) &= \frac{4\eta\sqrt{n}\iota}{b},\\
q_1 := q_1(\eta) &= \left|1-\frac{2\eta\lambda_1}{b}\left\|P_1\bar{x}_1\right\|_2^2\right|,\\
q_{-1} := q_{-1}(\eta) &= \max\left\{\left|1-\frac{2\eta}{b}(\lambda_2+n\iota)\right|+\frac{3n\eta\iota}{b},\ \left|1-\frac{2\eta}{b}(\lambda_n-n\iota)\right|+\frac{3n\eta\iota}{b}\right\}.
\end{aligned} \tag{38}$$

Then for $0<k\leq k_1$, Lemma 11 gives us

$$B_k \geq q_1 B_{k-1} - \xi A_{k-1}, \tag{39}$$

$$\begin{pmatrix}A_k\\B_k\end{pmatrix} \leq \begin{pmatrix}q_{-1} & \xi\\ \xi & q_1\end{pmatrix}\cdot\begin{pmatrix}A_{k-1}\\B_{k-1}\end{pmatrix}, \tag{40}$$

where "$\leq$" means "entry-wisely smaller than".

Let $\theta, \rho_{-1}, \rho_1$ determine the eigen decomposition of the coefficient matrix, i.e.,

$$\begin{pmatrix}q_{-1} & \xi\\ \xi & q_1\end{pmatrix} = \begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\rho_{-1} & 0\\ 0 & \rho_1\end{pmatrix}\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}. \tag{41}$$

Then Eq. (40) and Eq. (41) yield

$$\begin{aligned}
\begin{pmatrix} A_k \\ B_k \end{pmatrix} &\le \begin{pmatrix} q_{-1} & \xi \\ \xi & q_1 \end{pmatrix}^k \cdot \begin{pmatrix} A_0 \\ B_0 \end{pmatrix} \\
&= \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \rho_{-1}^k & 0 \\ 0 & \rho_1^k \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} A_0 \\ B_0 \end{pmatrix} \\
&= \begin{pmatrix} \rho_{-1}^k + (\rho_1^k - \rho_{-1}^k)\sin^2\theta & (\rho_1^k - \rho_{-1}^k)\cos\theta\sin\theta \\ (\rho_1^k - \rho_{-1}^k)\cos\theta\sin\theta & \rho_1^k - (\rho_1^k - \rho_{-1}^k)\sin^2\theta \end{pmatrix} \begin{pmatrix} A_0 \\ B_0 \end{pmatrix} \\
&= \begin{pmatrix} A_0 \cdot \rho_{-1}^k + \left(\rho_1^k - \rho_{-1}^k\right)(A_0\sin\theta + B_0\cos\theta)\sin\theta \\ B_0 \cdot \rho_1^k + \left(\rho_1^k - \rho_{-1}^k\right)(A_0\cos\theta - B_0\sin\theta)\sin\theta \end{pmatrix} \\
&\le \begin{pmatrix} A_0 \cdot \rho_{-1}^k + \left|\rho_1^k - \rho_{-1}^k\right| \sqrt{A_0^2 + B_0^2}\sin\theta \\ B_0 \cdot \rho_1^k + \left|\rho_1^k - \rho_{-1}^k\right| \sqrt{A_0^2 + B_0^2}\sin\theta \end{pmatrix} \\
&= \begin{pmatrix} A_0 \cdot \rho_{-1}^k + \left|\rho_1^k - \rho_{-1}^k\right| \cdot \|Pv_0\|_2 \cdot \sin\theta \\ B_0 \cdot \rho_1^k + \left|\rho_1^k - \rho_{-1}^k\right| \cdot \|Pv_0\|_2 \cdot \sin\theta \end{pmatrix}. \qquad (42)
\end{aligned}$$

We claim the following inequalities hold by our assumptions:

$$0 < \rho_{-1} < 1 < \rho_1 \le q_1 + \xi \qquad (43a)$$

$$\rho_{-1}^{k_1} \|Pv_0\|_2 \le \frac{\epsilon}{2} \cdot \beta \qquad (43b)$$

$$\rho_1^{k_1} \|Pv_0\|_2 \sin\theta \le \frac{\epsilon}{2} \cdot \beta, \qquad (43c)$$

$$\xi \cdot \left(A_0 + \frac{\epsilon\beta_0}{2}\right) < (q_1 - 1)\beta_0. \qquad (43d)$$

The verification of Eq. (43) is left later. In the following we prove the conclusions using Eq. (43).

We first bound A_{k_1} using Eq. (42) and Eq. (43):

$$\begin{aligned}
A_{k_1} &\le A_0 \cdot \rho_{-1}^{k_1} + \left|\rho_1^{k_1} - \rho_{-1}^{k_1}\right| \cdot \|Pv_0\|_2 \cdot \sin\theta \\
&\le \|Pv_0\|_2 \cdot \rho_{-1}^{k_1} + \rho_1^{k_1} \cdot \|Pv_0\|_2 \cdot \sin\theta \\
&\le \frac{\epsilon}{2} \cdot \beta + \frac{\epsilon}{2} \cdot \beta \\
&= \epsilon \cdot \beta,
\end{aligned}$$

which justifies the first conclusion. In addition we can obtain an uniform upper bound for A_k for $k = 0, 1, \ldots, k_1$:

$$\begin{aligned}
A_k &\le A_0 \cdot \rho_{-1}^k + \left|\rho_1^k - \rho_{-1}^k\right| \cdot \|Pv_0\|_2 \cdot \sin\theta \\
&\le A_0 + \rho_1^k \cdot \|Pv_0\|_2 \cdot \sin\theta \\
&\le A_0 + \frac{\epsilon}{2} \cdot \beta. \qquad (44)
\end{aligned}$$

Next we bound B_{k_1} using Eq. (42) and Eq. (43):

$$\begin{aligned}
B_{k_1} &\le B_0 \cdot \rho_1^{k_1} + \left|\rho_1^{k_1} - \rho_{-1}^{k_1}\right| \cdot \|Pv_0\|_2 \cdot \sin\theta \\
&\le \|Pv_0\|_2 \cdot \rho_1^{k_1} + \rho_1^{k_1} \cdot \|Pv_0\|_2 \cdot \sin\theta \\
&\le \|Pv_0\|_2 \cdot \rho_1^{k_1} + \frac{\epsilon}{2} \cdot \beta,
\end{aligned}$$

which justifies the second conclusion.

We proceed to derive the uniform lower bound for B_k for $k = 0, 1, \ldots, k_1$. We do it by induction. For $k = 0$, by assumption we have $B_0 \geq \beta_0$. Suppose $B_{k-1} \geq \beta_0$, then by Eq. (39), (43) and (44) we have

$$\begin{aligned} B_k &\geq q_1 \cdot B_{k-1} - \xi \cdot A_{k-1} \\ &\geq q_1 \cdot \beta_0 - \xi \cdot \left(A_0 + \frac{\epsilon}{2}\beta\right) \\ &\geq q_1 \cdot \beta_0 - \xi \cdot \left(A_0 + \frac{\epsilon}{2}\beta_0\right) \\ &\geq \beta_0, \end{aligned}$$

which justifies the third conclusion.

Verification of Eq. (43)

From Eq. (41) and Gershgorin circle theorem we have

$$\begin{aligned} q_1 - \xi \leq \rho_1 \leq q_1 + \xi, \\ q_{-1} - \xi \leq \rho_{-1} \leq q_{-1} + \xi. \end{aligned} \tag{45}$$

Moreover, reformatting Eq. (41) as

$$\begin{aligned} \begin{pmatrix} q_{-1} & \xi \\ \xi & q_1 \end{pmatrix} &= \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \rho_{-1} & 0 \\ 0 & \rho_1 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \\ &= \begin{pmatrix} \rho_{-1}\cos^2\theta + \rho_1 \sin^2\theta & (\rho_1 - \rho_{-1})\cos\theta\sin\theta \\ (\rho_1 - \rho_{-1})\cos\theta\sin\theta & \rho_{-1}\sin^2\theta + \rho_1\cos^2\theta \end{pmatrix} \\ &= \begin{pmatrix} \rho_{-1} + (\rho_1 - \rho_{-1})\sin^2\theta & (\rho_1 - \rho_{-1})\cos\theta\sin\theta \\ (\rho_1 - \rho_{-1})\cos\theta\sin\theta & \rho_1 - (\rho_1 - \rho_{-1})\sin^2\theta \end{pmatrix}, \end{aligned}$$

we then have

$$\frac{\xi}{q_1 - q_{-1}} = \frac{(\rho_1 - \rho_{-1})\cos\theta\sin\theta}{(\rho_1 - \rho_{-1})(1 - 2\sin^2\theta)} = \frac{1}{2}\tan 2\theta. \tag{46}$$

For Eq. (43a), using Eq. (45) it suffices to show

$$0 < q_1 - \xi, \tag{47a}$$

$$q_{-1} + \xi < 1, \tag{47b}$$

$$1 < q_1 - \xi. \tag{47c}$$

Notice the definitions of q_1, q_{-1} and ξ are given in Eq. (38). Firstly, Eq. (47c) holds trivially when $\sqrt{n} > 4/3$. Secondly, for Eq. (47b), noticing that $\xi = \frac{4\eta\sqrt{n}\iota}{b} \leq \frac{\eta n \iota}{b}$ when $n \geq 16$, it suffices to

show

$$
\begin{aligned}
& \max\left\{\left|1-\frac{2\eta}{b}(\lambda_2+n\iota)\right|+\frac{4n\eta\iota}{b}, \left|1-\frac{2\eta}{b}(\lambda_n-n\iota)\right|+\frac{4n\eta\iota}{b}\right\}<1 \\
\Leftrightarrow \quad & \begin{cases} \frac{4n\iota}{b}\eta-1<1-\frac{2(\lambda_2+n\iota)}{b}\eta<1-\frac{4n\iota}{b}\eta \\ \frac{4n\iota}{b}\eta-1<1-\frac{2(\lambda_n-n\iota)}{b}\eta<1-\frac{4n\iota}{b}\eta \end{cases} \\
\Leftrightarrow \quad & \begin{cases} \frac{2\lambda_2-2n\iota}{b}\eta>0 \\ \frac{2\lambda_2+6n\iota}{b}\eta<2 \\ \frac{2\lambda_n-6n\iota}{b}\eta>0 \\ \frac{2\lambda_n+2n\iota}{b}\eta<2 \end{cases} \\
\Leftarrow \quad & \begin{cases} \eta>0 \\ \lambda_2-n\iota>0 \\ \lambda_n-3n\iota>0 \\ \eta<\frac{b}{\lambda_2+3n\iota} \\ \eta<\frac{b}{\lambda_n+n\iota} \end{cases} \\
\Leftarrow \quad & \begin{cases} 3n\iota<\lambda_n \\ 0<\eta<\frac{b}{\lambda_2+3n\iota} \end{cases}
\end{aligned}
$$

which are given in assumptions. Thirdly, for Eq. (47c) it suffices to show

$$
\begin{aligned}
& \frac{2\eta\lambda_1}{b}\left\|P_1\bar{x}_1\right\|_2^2-1-\frac{4\eta\sqrt{n\iota}}{b}>1 \\
\Leftarrow \quad & \frac{2\lambda_1(1-4n\iota^2)}{b}\eta-\frac{4\sqrt{n\iota}}{b}\eta>2 \qquad \text{(by Lemma 3)} \\
\Leftarrow \quad & \eta>\frac{b}{\lambda_1(1-4n\iota^2)-2\sqrt{n\iota}} \\
\Leftarrow \quad & \eta>\frac{b}{\lambda_1-3\sqrt{n\iota}}, \qquad \text{(since } n\iota<1\text{)}
\end{aligned}
$$

which are given in assumptions.

For Eq. (43b), it suffices to show set

$$
k_1=1+\frac{\log\frac{0.5\epsilon\beta}{\|Pv_0\|_2}}{\log\rho_{-1}}=\mathcal{O}\left(\log\frac{1}{\epsilon\beta}\right),
$$

as given in assumptions.

For Eq. (43c), using Eq. (43b) it suffices to show

$$\begin{aligned}
& \sin\theta \le \left(\frac{\rho_{-1}}{\rho_1}\right)^{k_1} = \frac{\rho_{-1}}{\rho_1} \cdot \left(\frac{0.5\epsilon\beta}{\|Pv_0\|_2}\right)^{1-\frac{\log\rho_1}{\log\rho_{-1}}} \\
\Leftarrow \quad & \sin\theta \le \frac{q_{-1}-\xi}{q_1+\xi} \cdot \left(\frac{0.5\epsilon\beta}{\|Pv_0\|_2}\right)^{1-\frac{\log(q_1+\xi)}{\log(q_{-1}-\xi)}} \\
\Leftarrow \quad & \xi \le 0.9\,(q_1-q_{-1}) \cdot \frac{q_{-1}-\xi}{q_1+\xi} \cdot \left(\frac{0.5\epsilon\beta}{\|Pv_0\|_2}\right)^{1-\frac{\log(q_1+\xi)}{\log(q_{-1}-\xi)}} \qquad \text{(by Eq. (46))} \\
\Leftarrow \quad & \sqrt{n}\iota \le \mathrm{poly}\,(\epsilon\beta)\,. \qquad \text{(by Eq. (38))}
\end{aligned}$$

For Eq. (43c), it suffices to show

$$\begin{aligned}
& \xi \le \frac{(q_1-1)\beta_0}{A_0+0.5\epsilon\beta_0} \\
\Leftarrow \quad & \sqrt{n}\iota \le \mathcal{O}\,(1)\,. \qquad \text{(by Eq. (38))}
\end{aligned}$$

■

C.10. Proof of Lemma 13

Proof [Proof of Lemma 13] Let

$$\begin{aligned}
\xi' &:= \xi(\eta') = \frac{4\eta'\sqrt{n}\iota}{b}, \\
q_1' &:= q_1(\eta') = \left|1 - \frac{2\eta'\lambda_1}{b}\,\|P_1\bar{x}_1\|_2^2\right|, \\
q_{-1}' &:= q_{-1}(\eta') = \max\left\{\left|1-\frac{2\eta'}{b}(\lambda_2+n\iota)\right| + \frac{3n\eta'\iota}{b},\ \left|1-\frac{2\eta'}{b}(\lambda_n-n\iota)\right| + \frac{3n\eta'\iota}{b}\right\}.
\end{aligned} \tag{48}$$

Then for $k_1 < k \le k_2$, Lemma 11 gives us

$$\begin{pmatrix} A_k \\ B_k \end{pmatrix} \le \begin{pmatrix} q_{-1}' & \xi' \\ \xi' & q_1' \end{pmatrix} \cdot \begin{pmatrix} A_{k-1} \\ B_{k-1} \end{pmatrix}, \tag{49}$$

where "$\le$" means "entry-wisely smaller than". Denote

$$B := \|Pv_0\|_2 \cdot \rho_1^{k_1} + \frac{\epsilon}{2}\cdot\beta = \mathrm{poly}\left(\frac{1}{\epsilon\beta}\right). \tag{50}$$

We claim the following inequalities hold by our assumptions:

$$0 < q_1' < q_{-1}' < 1, \tag{51a}$$

$$\xi' \cdot \epsilon \le q_{-1}' - q_1', \tag{51b}$$

$$\xi' \cdot B \le (1-q_{-1}')\cdot\epsilon\cdot\beta. \tag{51c}$$

The verification of Eq. (51) is left later. In the following we prove the main conclusions in the lemma using Eq. (51). We proceed by induction. Clearly the conclusions are true for $k = k_1$. Suppose for $k_1, \dots, k-1$, the conclusions are also true. Then the induction assumptions give us

$$A_{k-1} \le \epsilon \cdot \beta, \tag{52}$$
$$B_{k-1} \le B_{k_1} \le B, \tag{53}$$

where the last inequality is due to $B \ge B_{k_1} \ge \beta_0 > \beta$. Then by Eq. (49) we have

$$\begin{aligned} B_k &\le q_1' \cdot B_{k-1} + \xi' \cdot A_{k-1} \\ &\le q_1' \cdot B_{k-1} + \xi' \cdot \epsilon \cdot \beta \qquad \text{(by Eq. (52))} \\ &\le q_1' \cdot B_{k-1} + (q_{-1}' - q_1') \cdot \beta \qquad \text{(by Eq. (51b))} \\ &\le \begin{cases} q_{-1}' \cdot B_{k-1}, & B_{k-1} > \beta, \\ \beta, & B_{k-1} \le \beta. \end{cases} \qquad \text{(by Eq. (51a))} \end{aligned}$$

Also by Eq. (49) we have

$$\begin{aligned} A_k &\le q_{-1}' \cdot A_{k-1} + \xi' \cdot B_k \\ &\le q_{-1}' \cdot \epsilon \cdot \beta + \xi' \cdot B \qquad \text{(by Eq. (52) and (53))} \\ &\le q_{-1}' \cdot \epsilon \cdot \beta + (1 - q_{-1}') \cdot \epsilon \cdot \beta \qquad \text{(by Eq. (51c))} \\ &= \epsilon \cdot \beta. \end{aligned}$$

Verification of Eq. (51) Notice the definitions of q_1', q_{-1}' and ξ' are given in Eq. (38). Recall $q_{-1}' < 1$ is already justified by the choice of learning rate $\eta' < \frac{1}{\lambda_2 + 3n\iota}$ (e.g., see Lemma 10), thus for Eq. (51a), it suffices to show

$$\begin{aligned} & 0 < 1 - \frac{2\eta'\lambda_1}{b} \left\| P_1 \bar{x}_1 \right\|_2^2 < 1 - \frac{2\eta'}{b}(\lambda_n - n\iota) + \frac{3n\eta'\iota}{b} \\ \Leftarrow \quad & \begin{cases} 1 - \frac{2\eta'\lambda_1}{b} > 0 \\ 2\lambda_1(1 - 4n\iota^2) > 2(\lambda_n - n\iota) + 3n\iota \end{cases} \qquad \text{(by Lemma 3)} \\ \Leftarrow \quad & \begin{cases} \eta' < \frac{b}{2\lambda_1} \\ \lambda_1 > \lambda_n + 2n\iota \end{cases} \end{aligned}$$

which are given in assumptions.

For Eq. (51b), it suffices to show

$$\begin{aligned} & \xi' \le \frac{1}{\epsilon} \cdot \left(q_{-1}' - q_1' \right) \\ \Leftarrow \quad & \sqrt{n\iota} \le \mathcal{O}\left(\frac{1}{\epsilon} \right), \end{aligned}$$

which is implied by $n\iota \le \text{poly}\,(\epsilon\beta)$.

For Eq. (51c), it suffices to show

$$\begin{aligned} & \xi' \le \frac{1}{B} \cdot \epsilon \cdot \left(1 - q_{-1}' \right) \beta \\ \Leftarrow \quad & \sqrt{n\iota} \le \text{poly}\,(\epsilon\beta)\,. \qquad \text{(by Eq. (50))} \end{aligned}$$

We complete our proof. ■

C.11. Proof of Lemma 14

Proof [Proof of Lemma 14] For the empirical loss,

$$L_{\mathcal{S}}(u) = \frac{1}{n}\left(w - w_*\right)^\top XX^\top \left(w - w_*\right) = \frac{1}{n}u^\top G^\top XX^\top Gu = \frac{1}{n}u^\top \Gamma u = \frac{1}{n}\sum_{i=1}^{n}\gamma_i\left(u^{(i)}\right)^2.$$

For the population loss,

$$L_{\mathcal{D}}(u) = \mu\left\|w - w_*\right\|_2^2 = \mu\left\|Gu\right\|_2^2 = \mu\left\|u\right\|_2^2.$$

For the hypothesis class $\mathcal{H}_{\mathcal{S}} = \left\{w \in \mathbb{R}^d : P_\perp w = P_\perp w_0\right\}$, Note $P_\perp G = \operatorname{diag}\left(0, \ldots, 0, 1, \ldots, 1\right)$. Apply $w - w_* = Gu$ and notice $w_0 - w_* = Gu_0$, then we obtain

$$\begin{aligned}\mathcal{H}_{\mathcal{S}} &= \left\{u \in \mathbb{R}^d : P_\perp Gu = P_\perp Gu_0\right\} \\ &= \left\{u \in \mathbb{R}^d : u^{(i)} = u_0^{(i)},\ \text{for } i = n+1, \ldots, d\right\}.\end{aligned}$$

For the level set, we only need to note that $L_{\mathcal{S}}^* = \inf_{u\in\mathcal{H}_{\mathcal{S}}} L_{\mathcal{S}}(u) = 0$.

As for the estimation error, we note that

$$\inf_{u\in\mathcal{U}} L_{\mathcal{D}}(u) = \mu\sum_{i=n+1}^{d}\left(u_0^{(i)}\right)^2,$$

thus for $u \in \mathcal{U}$, we have

$$\begin{aligned}\Delta(u) = L(u) - \inf_{u'\in\mathcal{U}} L(u') &= \mu\left\|u\right\|_2^2 - \mu\sum_{i=n+1}^{d}\left(u_0^{(i)}\right)^2 \\ &= \mu\sum_{i=1}^{n}\left(u^{(i)}\right)^2 + \mu\sum_{i=n+1}^{d}\left(u^{(i)}\right)^2 - \mu\sum_{i=n+1}^{d}\left(u_0^{(i)}\right)^2 \\ &= \mu\sum_{i=1}^{n}\left(u^{(i)}\right)^2.\end{aligned}$$

Now consider $u \in \mathcal{U}$, i.e., $\frac{1}{n}\sum_{i=1}^{n}\gamma_i\left(u^{(i)}\right)^2 = \alpha$, then

$$\Delta_* = \inf_{u\in\mathcal{U}}\Delta(u) = \inf_{n\alpha=\sum_{i=1}^{n}\gamma_i\left(u^{(i)}\right)^2}\mu\sum_{i=1}^{n}\left(u^{(i)}\right)^2 = \frac{\mu n\alpha}{\gamma_1},$$

where the inferior is attended when, e.g., $u^{(1)} = \pm\sqrt{\frac{n\alpha}{\gamma_1}}$ and $u^{(2)} = \cdots = u^{(n)} = 0$. ■