Generalization Analysis of Linear Knowledge Distillation

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Knowledge distillation (KD), a framework in which a smaller student model is trained under the guidance of a stronger teacher, has become a popular technique for model compression. Despite its empirical success, the theoretical understanding of KD remains underexplored. In this work, we theoretically study the generalization behavior of linear knowledge distillation (LKD), a simplified setting in which the student is restricted to a linear model. We first characterize the implicit bias of gradient descent on separable training data when the student is trained with LKD. Building on the results, we derive a population zero-one risk bound for the distilled student under binary Gaussian mixture data. We quantify the provable generalization benefit of LKD distilled from various teachers compared to standard hard-label training.

1. Introduction

Knowledge distillation (KD) [13] refers to a technique in which a smaller student model is trained with the auxiliary supervision of a larger teacher model. Given its empirical success across diverse domains—from computer vision [29, 33] to large language models [1, 11]—KD has been widely adopted for improving performance and compressing models.

Accordingly, understanding the mechanisms underlying the effectiveness of KD in generalization has become an important research direction. However, most prior work falls into one of the following categories: (1) Taking a *static* viewpoint, e.g., solely analyzing the KD objective, thereby overlooking the optimization dynamics closely tied to generalization [9, 17, 19, 38, 39], (2) consider settings where the teacher and student share the same architecture (i.e., self-distillation) [10, 22, 32], or (3) analyzing a surrogate generalization measure (e.g., agreement in predictions between the teacher and student) [15, 26]. A more detailed discussion of related work is provided in Section A.

Contribution. In this work, we study the generalization properties of KD for binary classification on separable training data. For theoretical tractability and simplicity, we consider a setting in which the teacher's knowledge is distilled into a linear student model, which we refer to as *linear knowledge distillation* (LKD). We first derive the converged solution—also known as *implicit bias*—of gradient descent (GD) in LKD. Based on the result, we derive an upper bound on the population zero-one risk under binary Gaussian mixture. We further show that LKD can achieve better generalization than standard hard label training, when distilled from various form of teachers.

2. Problem setup

We consider the task of binary linear classification in the *overparametrized* regime. We want to predict the *label* $y \in \{-1, +1\}$ from the *feature* $\mathbf{x} \in \mathbb{R}^d$ for the pair $(\mathbf{x}, y)$ jointly drawn from the pop-

ulation distribution $\mathcal{D}$. The *student model* is a linear classifier $f_{\mathbf{w}}(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^\top \mathbf{x})$, parametrized by the *weight* $\mathbf{w} \in \mathbb{R}^d$. Our goal is to minimize the *zero-one risk* of the student, i.e.,

$$\mathcal{R}(\mathbf{w}) := \Pr_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq f_{\mathbf{w}}(\mathbf{x})] = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}[y \cdot \mathbf{w}^\top \mathbf{x} < 0]. \tag{1}$$

We train the student by minimizing the *empirical KD risk* over a training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $d \geq n$. For each sample $\mathbf{x}_i$, we use two forms of supervision: the *hard label* $h_i := (1 + y_i)/2$ is simply the label rescaled to have values in $\{0, 1\}$. The *soft label* $p_i := \sigma(\mathcal{T}(\mathbf{x}_i))$ is the (positive) class probability predicted by the *teacher model* $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}$, where $\sigma(x) := 1/(1 + e^{-x})$ is the sigmoid function. Given these supervisions, the empirical KD risk is defined as

$$\widehat{\mathcal{L}}(\mathbf{w}, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left[(1 - \lambda)\, \ell_{h_i}(\mathbf{w}^\top \mathbf{x}_i) + \lambda\, \ell_{p_i}(\mathbf{w}^\top \mathbf{x}_i) \right], \tag{2}$$

where $\lambda \in [0, 1]$ is the distillation weight that balances the impact of the hard and soft labels, and $\ell_q(u)$ denotes the standard cross-entropy loss $\ell_q(u) := -q \log \sigma(u) - (1 - q) \log(1 - \sigma(u))$.

The student $\mathbf{w}$ is trained via full-batch gradient descent, with learning rate $\eta > 0$:

$$\mathbf{w}_\lambda(t + 1) = \mathbf{w}_\lambda(t) - \eta \nabla_{\mathbf{w}} \widehat{\mathcal{L}}(\mathbf{w}_\lambda(t), \lambda). \tag{3}$$

For our theoretical analyses, we make the following assumptions.

Assumption 1 (Linear separability) *There exists* $\mathbf{w}_* \in \mathbb{R}^d$ *with* $y_i \mathbf{w}_*^\top \mathbf{x}_i > 0$, $\forall i \in [n]$.

Assumption 2 (Full-rank) *We have* $\mathrm{rank}(\mathbf{X}) = n$, *where* $\mathbf{X} := [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$.

Notation. For $\mathbf{z} \in \mathbb{R}^n$, $\mathrm{diag}(\mathbf{z}) \in \mathbb{R}^{n \times n}$ denotes the corresponding diagonal matrix. $\mathbf{X}^\dagger$ denotes the pseudoinverse of $\mathbf{X}$. $\Phi(\cdot)$ denotes the standard Gaussian CDF. For a symmetric, positive-definite matrix $\mathbf{A}$, $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} := \mathbf{x}^\top \mathbf{A} \mathbf{y}$ is the induced inner product, and consequently $\cos_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) := \frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}}}{\|\mathbf{x}\|_{\mathbf{A}} \|\mathbf{y}\|_{\mathbf{A}}}$. The identity matrix is denoted by $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ and the all-one vector by $\mathbf{1}_d \in \mathbb{R}^d$.

3. Main results

In this section, we provide our main theoretical results, from implicit bias of LKD to population risk bound with different teacher choices. All proofs in this section are deferred to Section C.

3.1. Implicit bias

We first characterize the exact form of the implicit bias induced by GD in LKD.

Lemma 3 *For any* $\lambda \in (0, 1]$ *and any stepsize* $0 < \eta < 8n/\|\mathbf{X}\|_2^2$, *we have*

$$\lim_{t \to \infty} \mathbf{w}_\lambda(t) = (\mathbf{I}_d - \mathbf{X}^\dagger \mathbf{X}) \mathbf{w}(0) + \mathbf{w}_\lambda^\star, \qquad \mathbf{w}_\lambda^\star := \mathbf{X}^\dagger \sigma^{-1}((1 - \lambda)\mathbf{h} + \lambda \mathbf{p}), \tag{4}$$

where $\mathbf{h} := [h_1, \cdots, h_n]^\top$ *and* $\mathbf{p} := [p_1, \cdots, p_n]^\top$.

Remark. Since $(\mathbf{I}_d - \mathbf{X}^\dagger\mathbf{X})\mathbf{w}(0)$ is unaffected by the GD iterations and λ, we will simply omit this term in the subsequent discussions, by assuming $\mathbf{w}(0) = \mathbf{0}$.

Lemma 3 extends the result of Phuong and Lampert [26]—which considered distillation-only loss and gradient flow—to the case of gradient descent with mixed supervision. By letting $\lambda = 1$, we can recover the identical solution as in [26], i.e., GD and GF converges to the same minimum.

Through the extension to GD, we can now compare the result more directly with the results of Soudry et al. [31] on the implicit bias of linear classifiers trained via GD on separable data, using only the label-training loss (i.e., $\lambda = 0$). In this case, the magnitude of the weight $\|\mathbf{w}_0(t)\|_2$ diverges and its direction $\mathbf{w}_0(t)/\|\mathbf{w}_0(t)\|_2$ converges to the hard-margin support vector machine (SVM), i.e.,

$$\lim_{t\to\infty} \mathbf{w}_0(t)/\|\mathbf{w}_0(t)\|_2 = \mathbf{w}_{\mathrm{SVM}}/\|\mathbf{w}_{\mathrm{SVM}}\|_2, \qquad \text{where} \tag{5}$$

$$\mathbf{w}_{\mathrm{SVM}} := \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\|_2 \quad \text{s.t.} \quad y_i \cdot \mathbf{w}^\top \mathbf{x}_i \geq 1, \quad \forall i \in [n]. \tag{6}$$

On the other hand, Lemma 3 reveals that for any $\lambda \in (0, 1]$, the weight converges to a *finite-norm solution* $\mathbf{w}_\lambda^\star$. Here, the solution is equivalent to the *minimum-norm interpolator* for the mixed label $z_{\lambda,i} := (1-\lambda)h_i + \lambda p_i$, i.e.,

$$\mathbf{w}_\lambda^\star = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\|_2 \quad \text{s.t.} \quad \mathbf{w}^\top \mathbf{x}_i = \sigma^{-1}(z_{\lambda,i}), \quad \forall i \in [n]. \tag{7}$$

Note that $\sigma^{-1}(z_{\lambda,i})$ diverges as $\lambda \to 0$, and thus the limit of $\mathbf{w}_\lambda^\star$ is *singular*. Nevertheless, under appropriate assumptions (see Assumption 4), the normalized direction admits a well-defined limit at $\lambda = 0$. In this case, we can write

$$\mathbf{w}_\lambda^\star = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\|_2 \quad \text{s.t.} \quad \mathbf{w}^\top \mathbf{x}_i = \begin{cases} y_i & \cdots \lambda = 0 \\ \sigma^{-1}(z_{\lambda,i}) & \cdots \lambda \in (0,1] \end{cases}, \quad \forall i \in [n]. \tag{8}$$

In this sense, Lemma 3 can be seen as providing a smooth interpolation of two classic results—on $\lambda = 0$ [31] and $\lambda = 1$ [26]—for the intermediate regime $\lambda \in (0, 1]$.

3.2. Risk bound

In this section, we analyze the generalization behavior of the linear student, based on the results from Section 3.1. We first specify the data model for our analysis, i.e., a binary Gaussian mixture.

Assumption 4 *The training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are generated i.i.d. as follows:*

- *Binary label $y_i \sim \mathrm{Unif}(\{-1, +1\})$.*
- *$\mathbf{x}_i = y_i\boldsymbol{\mu} + \mathbf{z}_i$ with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ independent of y_i, where $\boldsymbol{\mu} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$.*
- *There exists an constant $C_D > 0$ such that $d \geq C_D \max\left\{n^2, n\sqrt{\log n}\|\boldsymbol{\mu}\|_2\right\}$ and $\|\boldsymbol{\mu}\|_2 \geq C_D$.*

The last condition—from Cao et al. [7]—ensures both linear separability of the training data and the equivalence between the SVM and minimum-norm interpolator in $\lambda \to 0$ limit, since all training samples become *support vectors*. Moreover, under the data model, Assumption 2 holds almost surely. With this setup, we now present the risk bound of LKD.

Theorem 5 *Let* $\widetilde{\mathbf{X}} = \mathrm{diag}(\mathbf{y})\mathbf{X}$, $\mathbf{s} = \widetilde{\mathbf{X}}\boldsymbol{\mu}$, $\mathbf{a}_\lambda = \widetilde{\mathbf{X}}\mathbf{w}_\lambda^\star$, $\mathbf{K} = (\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top)^{-1}$. *Assuming* $(\mathbf{w}_\lambda^\star)^\top\boldsymbol{\mu} \geq 0$, *for every* $\lambda \in [0,1]$, *we have*

$$\mathcal{R}(\mathbf{w}_\lambda^\star) \leq \exp\left(-C\frac{\cos^2_{\mathbf{K}}(\mathbf{a}_\lambda, \mathbf{s})}{\cos^2_{\mathbf{K}}(\mathbf{1}_n, \mathbf{s})} \cdot \frac{n\|\boldsymbol{\mu}\|_2^4}{d + n\|\boldsymbol{\mu}\|_2^2}\right). \tag{9}$$

with probability at least $1 - O(n^{-1})$ *and for some constant* $C > 0$.

By letting $\lambda = 0$ in Theorem 5, the squared ratio involving the cosine term becomes one, and the risk bound recovers the same rate as that of Cao et al. [7], i.e., $\exp(-O(n/d))$. More generally, the bound can be tighter or looser—compared to $\lambda = 0$—depending on the cosine ratio, which quantifies the *alignment between the logit vectors* under the inner product induced by $\mathbf{K}$.

3.3. LKD vs. Label Training: When Does Distillation Improve Generalization?

We now ask: *Does LKD lead to better generalization than standard label training? If so, when?* From Eq. (9), we can see that the effects of λ and teacher supervision enter only through the cosine ratio. This motivates the following definition of the *generalization improvement* factor $\mathsf{Imp}(\lambda, \mathcal{T})$.

$$\mathsf{Imp}(\lambda, \mathcal{T}) := \cos^2_{\mathbf{K}}(\mathbf{a}_\lambda, \mathbf{s}) / \cos^2_{\mathbf{K}}(\mathbf{1}_n, \mathbf{s}). \tag{10}$$

Here, the risk bound becomes tighter when $\mathsf{Imp}(\lambda, \mathcal{T}) > 1$ and looser otherwise, relative to the $\lambda = 0$ baseline. Now, we analyze the different choices of teacher, derive $\mathsf{Imp}(\lambda, \mathcal{T})$, and establish the conditions for improvement.

Example 1: Bayes teacher. Theoretical works on KD often study an idealized setting in which the teacher is given by the *Bayes posterior*, i.e., the teacher outputs the ground-truth population distribution $p = p^\star(y = +1|\mathbf{x})$ [9, 19, 23]. The result is stated in Lemma 6.

Lemma 6 *The Bayes teacher is obtained as* $\mathcal{T}(\mathbf{x}) = 2\boldsymbol{\mu}^\top\mathbf{x}$. *For any* $\lambda \in (0,1]$, *we have*

$$1 \lesssim \mathsf{Imp}(\lambda, \mathcal{T}) \lesssim 1 + 1/\|\boldsymbol{\mu}\|_2^2, \tag{11}$$

and $\mathsf{Imp}(\lambda, \mathcal{T})$ *is monotonically increasing in* $\lambda \in (0,1]$.

Example 2: Biased teacher. Next, we consider the Bayes teacher with non-zero bias, i.e., $\mathcal{T}_b(\mathbf{x}) = 2\boldsymbol{\mu}^\top\mathbf{x} + b$, where $b \in \mathbb{R} \setminus \{0\}$.

Lemma 7 *For* $\lambda \in (0,1]$, *we have* $\mathsf{Imp}(\lambda, \mathcal{T}_b) > 1$ *when*

$$|b| \lesssim \sqrt{2\|\boldsymbol{\mu}\|_2^2 + \log(1/\lambda) + 1} \tag{12}$$

Example 3: Noisy teacher. Here, we consider a *noisy* teacher, defined as a Bayes teacher whose output label is flipped independently with probability $\rho \in [0, 0.5]$. The result is stated in Lemma 8.

Lemma 8 *We have* $\mathsf{Imp}\,(1, \mathcal{T}_\rho) > 1$ *when*

$$\rho \lesssim 0.5 - \sqrt{\|\boldsymbol{\mu}\|_2^2/(4\|\boldsymbol{\mu}\|_2^2 + 4)}. \tag{13}$$

Example 4: Linear teacher. As a final example, we consider the linearly parametrized teacher by $\mathbf{v} \in \mathbb{R}^d$, written as $\mathcal{T}(\mathbf{x}) = \mathbf{v}^\top \mathbf{x}$, without assuming Bayes optimal direction.

Lemma 9 *Let* $\mathbf{v}_\perp := \mathbf{v} - \frac{\mathbf{v}^\top \boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2^2}\boldsymbol{\mu}$ *denotes the component of* $\mathbf{v}$ *orthogonal to* $\boldsymbol{\mu}$*. For* $\lambda \in (0, 1]$*, we have* $\mathsf{Imp}(\lambda, \mathcal{T}_\mathbf{v}) > 1$ *when*

$$\|\mathbf{v}_\perp\|_2^2 \lesssim \frac{(\mathbf{v}^\top \boldsymbol{\mu})^2}{\|\boldsymbol{\mu}\|_2^2}\left(1 + \frac{1}{2\|\boldsymbol{\mu}\|_2^2} + \frac{2\log(1/\lambda)}{\mathbf{v}^\top \boldsymbol{\mu}}\right). \tag{14}$$

Implication. For the *Bayes teacher*, any $\lambda \in (0, 1]$ improves the risk bound. Moreover, the improvement is maximized at $\lambda = 1$ (i.e., pure distillation): In this case, the numerator of Eq. (10) becomes one, thereby attaining the optimal improvement. For the *biased*, *noisy*, and *linear* teachers, improvement is not always guaranteed. However, in each case, we can set b, ρ, or $\mathbf{v}$ appropriately to obtain $\mathsf{Imp}\,(\lambda, \mathcal{T}) > 1$. This implies that λ and the teacher's *imperfectness* should be considered jointly for achieving better generalization in KD.

Numerical experiments. To analyze this more precisely, we conduct numerical experiments, with results and additional details provided in Section B. In Figs. 1 to 3, we plot $\mathsf{Imp}(\lambda, \mathcal{T})$ as a function of λ for the biased, noisy, and linear teachers, respectively. There, we show that the improvement factor predicted by our theory closely matches the improvement factors (1) derived from the implicit bias (Eq. (8)) and (2) obtained from standard gradient-based training.

In the case of the Bayes teacher ($b = 0$, $\rho = 0$, and $\|\mathbf{v}_\perp\|_2 = 0$), as mentioned earlier, the improvement increases monotonically as λ becomes larger.

Even when the teacher is *imperfect*, pure distillation can still be beneficial (e.g., $b = 4$ in Fig. 1). In some cases, improvement occurs only when λ falls below a certain *threshold*, as in the case of $b = 8$. Moreover, somewhat counterintuitively, even when pure distillation severely degrades generalization, a sufficiently small λ can still outperform label training.

4. Conclusion

In this work, we theoretically study the generalization behavior of LKD, a learning framework in which a linear student is trained under the supervision of a teacher. We first characterize the exact form of its asymptotic implicit bias. Building on this result, we derive a population zero-one risk bound that depends on both the distillation weight and the output of a teacher, where the tightness of the bound is determined by the alignment between the logits of the Bayes optimal classifier and those of the student. For different choices of teacher, we further show how the bound becomes tighter or looser compared with standard label training.

Limitation. Our results do not capture the effect of *temperature*, although this hyperparameter plays an important role in feature learning dynamics [2, 4, 36]. Also, our risk bound relies on the asymptotic implicit bias and thus does not capture finite-time behavior [24, 37].

References

[1] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*, 2024.

[2] Atish Agarwala, Samuel Stern Schoenholz, Jeffrey Pennington, and Yann Dauphin. Temperature check: theory and practice for training models with softmax-cross-entropy losses. *TMLR*, 2023.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *ICLR*, 2023.

[4] Alexander Atanasov, Alexandru Meterez, James B Simon, and Cengiz Pehlevan. The optimization landscape of SGD across the feature learning strength. In *ICLR*, 2025.

[5] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *PNAS*, 2020.

[6] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

[7] Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In *NeurIPS*, 2021.

[8] Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *JMLR*, 2021.

[9] Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. In *ICLR*, 2021.

[10] Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. In *ICML*, 2023.

[11] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *ICLR*, 2024.

[12] Hrayr Harutyunyan, Ankit Singh Rawat, Aditya Krishna Menon, Seungyeon Kim, and Sanjiv Kumar. Supervision complexity and its role in knowledge distillation. In *ICLR*, 2023.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

[15] Guangda Ji and Zhanxing Zhu. Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. In *NeurIPS*, 2020.

[16] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, 2000.

[17] Xin-Chun Li, Wen-Shu Fan, Shaoming Song, Yinchuan Li, Shao Yunfeng, De-Chuan Zhan, et al. Asymmetric temperature scaling makes larger networks teach well again. In *NeurIPS*, 2022.

[18] Saptarshi Mandal, Xiaojun Lin, and Rayadurgam Srikant. A theoretical analysis of soft-label vs hard-label training in neural networks. In *L4DC*, 2025.

[19] Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *ICML*, 2021.

[20] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[21] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *ICML*, 2020.

[22] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. In *NeurIPS*, 2020.

[23] Itai Morad, Nir Shlezinger, and Yonina C. Eldar. SGD-based knowledge distillation with bayesian teachers: Theory and guidelines. In *ICLR*, 2026.

[24] Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. In *NeurIPS*, 2020.

[25] Vaishnavh Nagarajan, Aditya K Menon, Srinadh Bhojanapalli, Hossein Mobahi, and Sanjiv Kumar. On student-teacher deviations in distillation: does it pay to disobey? In *NeurIPS*, 2023.

[26] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *ICML*, 2019.

[27] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *ICML*, 2021.

[28] Luca Saglietti and Lenka Zdeborová. Solvable model for inheriting the regularization through knowledge distillation. In *MSML*, 2022.

[29] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

[30] Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *COLT*, 2022.

[31] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *JMLR*, 2018.

[32] Kaito Takanami, Takashi Takahashi, and Ayaka Sakata. The effect of optimal self-distillation in noisy gaussian mixture model. In *NeurIPS*, 2025.

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[34] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

[35] Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIMODS*, 2022.

[36] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *COLT*, 2020.

[37] Taesun Yeom, Taehyeok Ha, and Jaeho Lee. Over-alignment vs over-fitting: The role of feature learning strength in generalization. *arXiv preprint arXiv:2602.00827*, 2026.

[38] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, 2020.

[39] Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. In *ICLR*, 2024.

Contents

Appendix A. Related Work

Theoretical analysis on knowledge distillation. Since knowledge distillation (KD) [13] has demonstrated empirical success across various domains, understanding its internal mechanisms has become an important and active research direction in deep learning [3, 12, 18, 25].

A seminal work by Phuong and Lampert [26] studies the implicit bias of KD with only soft labels (in our case, $\lambda = 1$), when both the teacher and student are linear models. They also analyze the *transfer risk* of the distilled student, defined as the difference in prediction between the student and the teacher. Ji and Zhu [15] extend this framework to the setting where the student is in the neural tangent kernel regime [14]. However, transfer risk is not aligned with standard generalization metrics, such as test loss or accuracy, because it merely measures agreement in predictions or parameters between teacher and student. In contrast, our generalization bound is directly tied to test accuracy. Moreover, our work provides provable quantitative performance gains compared to standard label training.

Another research direction focus on characterizing the *quality of teacher* theoretically. For example, Menon et al. [19] analyze KD from a statistical perspective, characterizing the optimal teacher as one that outputs the *Bayes class probabilities* and explaining its benefits through a generalization bound. Their results suggest that the minimizing the calibration error (i.e., ℓ_2 distance between Bayes class probability and the teacher prediction) improves generalization. Dao et al. [9] take a similar viewpoint while relaxing assumptions on the loss functions. However, these *static* viewpoint does not consider the gradient-based optimization dynamic of the student, which is closely tied to generalization behavior. Compared to these works, we derive generalization bound based on the converged solution (i.e., implicit bias) depending on both λ and teacher predictions.

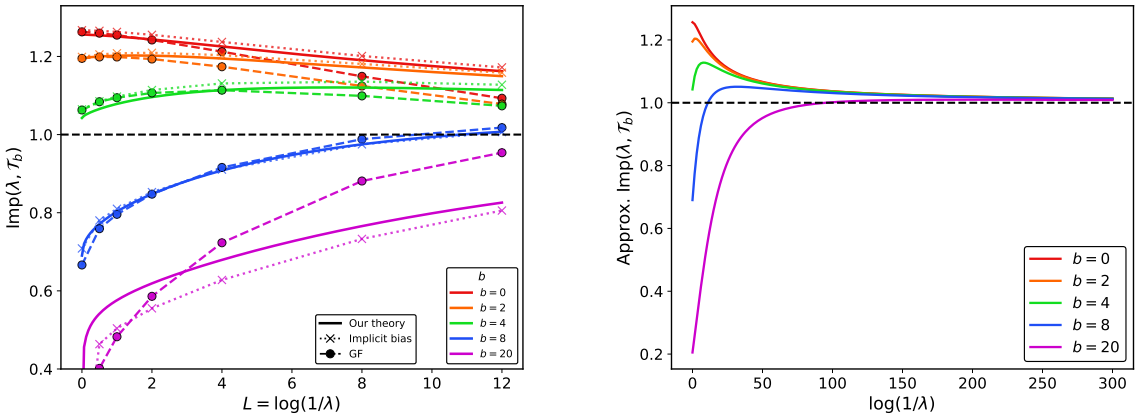
Classification on Gaussian mixture data. In theoretical machine learning, (sub-)Gaussian mixture is among the most tractable and widely used data model for studying various phenomena that we observed in practical scenarios.

For example, Chatterji and Long [8] and Wang and Thrampoulidis [35] derive the finite-sample risk bound on Gaussian mixture in linear classification, and discovered the condition when the interpolating classifier can be overfitting *benignly* (also known as "benign overfitting" [5]). Cao et al. [7] relax the data assumption and provide tighter risk bound. These results focus solely on the maximum ℓ_2-margin classifier, which corresponds to the implicit bias when $\lambda = 0$ in our paper. In contrast, our work extends this setup to distillation scenarios.

In the domain of KD, several studies have leveraged the *replica method* [20] from statistical physics to derive closed-form expressions for the generalization error in the proportional limit regime, where $d/n \to \alpha \in (0, \infty)$ as $d, n \to \infty$ [21, 27]. Within this framework, Saglietti and Zdeborová [28] provide an analysis of KD, demonstrating that the distillation process inherits the implicit regularization of the teacher model. Recently, Takanami et al. [32] study self-distillation on a noisy Gaussian mixture and claim that the effects of *dark knowledge* are negligible; Rather, the distillation process *denoises* the noisy labels, thereby improving generalization. Compared to these works, we do not assume proportional limit: To our knowledge, we provide the first theoretical analysis of KD under a Gaussian mixture data with finite number of training samples.

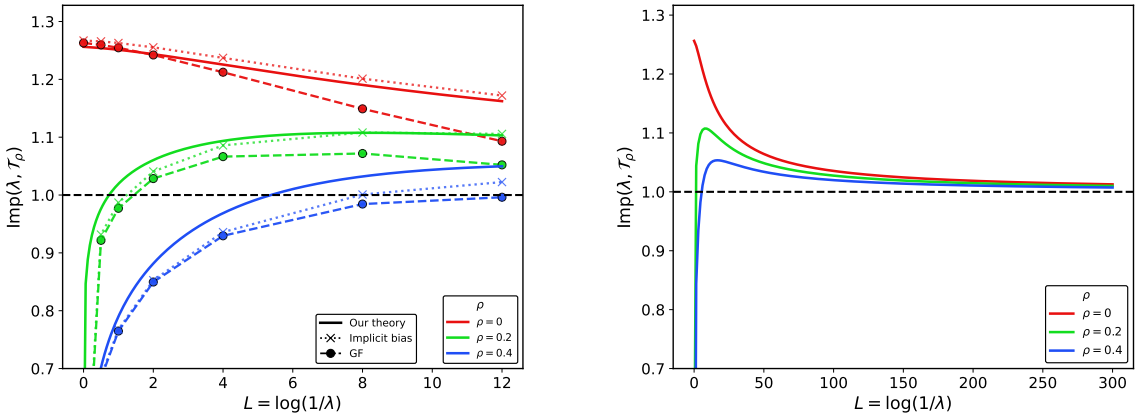
Appendix B. Numerical Experiments and Analysis

In this section, we present numerical experiments investigating how different factors—the distillation weight λ, the pair $(\lambda, \mathcal{T})$, the bias b, the flip probability ρ, and the ℓ_2-norm of the orthogonal component $\|\mathbf{v}_\perp\|_2$—affect the improvement factor $\mathsf{Imp}(\lambda, \mathcal{T})$.



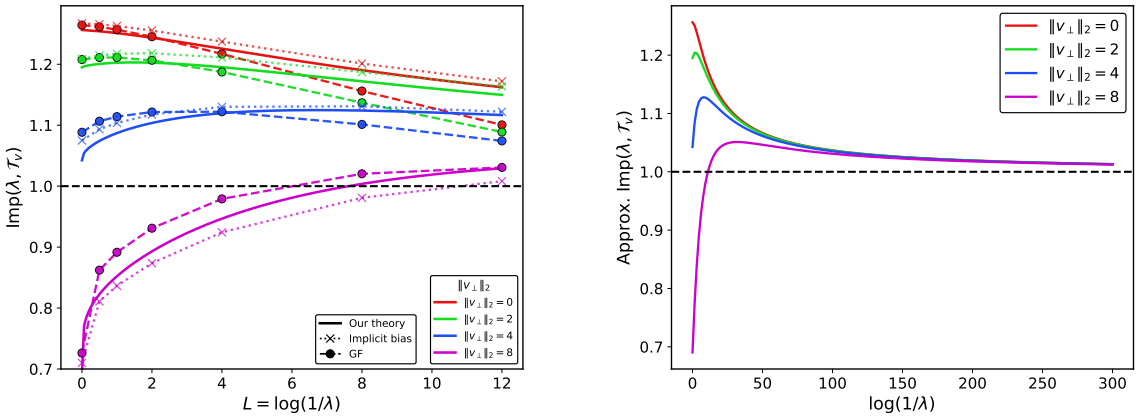
(a) Zoomed-in view. (a) Full view.

Figure 1: **Experiments on biased teacher:** λ vs. $\mathsf{Imp}\,(\lambda, \mathcal{T})$ for varying b. **Our theory** denotes the theoretical prediction derived in our analysis, **Implicit bias** denotes the solution from Eq. (8), and **GF** denotes the result obtained by training the student via gradient flow.



(a) Zoomed-in view. (b) Full view.

Figure 2: **Experiments on noisy teacher:** λ vs. $\mathsf{Imp}\,(\lambda, \mathcal{T})$ for varying ρ. **Our theory** denotes the theoretical prediction derived in our analysis, **Implicit bias** denotes the solution from Eq. (8), and **GF** denotes the result obtained by training the student via gradient flow.



(a) Zoomed-in view.

(b) Full view.

Figure 3: **Experiments on linear teacher:** λ vs. $\mathsf{Imp}\,(\lambda, \mathcal{T})$ for varying $\|\mathbf{v}_\perp\|_2$. **Our theory** denotes the theoretical prediction derived in our analysis, **Implicit bias** denotes the solution from Eq. (8), and **GF** denotes the result obtained by training the student via gradient flow.

In Figs. 1 to 3, we plot $\log(1/\lambda)$ versus approximated $\mathsf{Imp}(\lambda, \mathcal{T})$ (i.e., without $o(1)$ error term) plot, for the "Example 2", "Example 3", and "Example 4" respectively. Here, we can interpret the results as follows:

- **Generalization improvement with Bayes teacher.** When $b = 0$ in Fig. 1, $\rho = 0$ in Fig. 2, and $\|\mathbf{v}_\perp\|_2 = 0$ in Fig. 3, the teacher becomes Bayes teacher. Here, the results suggest that using the Bayes teacher *consistently outperforms* label training, as the improvement factor never falls below 1. Moreover, $\mathsf{Imp}(\lambda, \mathcal{T})$ increases as λ increases, as we proved in Lemma 6.

- **Emergence of an optimal λ.** For moderate values of b and ρ, the improvement is maximized at an optimal $\lambda^\star$, corresponding to the peak observed in the full-view figure. For example, in Fig. 1, as b increases—meaning that the deviation from the Bayes teacher becomes larger—$\lambda^\star$ decreases. This aligns with the intuition that as the teacher becomes more uninformative, we should assign a smaller weight to the soft labels.

- **(Moderately) Bad teacher can make good student.** Even for small levels of teacher's *imperfectness* (e.g., $b \in \{2, 4\}$), LKD with any λ outperforms the hard-label baseline. Interestingly, when $b = 8$, there exists a threshold λ_{thres}: For examp,e, when $\lambda < \lambda_{\text{thres}}$, LKD outperforms hard-label training; Otherwise, hard-label training achieves a better generalization bound.

- **(Very) Bad teacher harms generalization.** When $b = 20$, LKD cannot achieve better generalization than the hard-label baseline, in an approximate sense.

These analyses indicate that there is a *permissible level of teacher imperfection* for which LKD achieves better generalization than the hard-label baseline.

Our findings align with phenomena observed in practical KD scenarios. For example, (1) even when the teacher is imperfect, KD with a moderate λ can improve generalization compared to label training; and (2) a weaker teacher can produce a stronger student, a phenomenon also known as *weak-to-strong generalization* [6].

Experimental details. For all numerical experiments, we set $n = 50$, $d = 10^4$, and $\|\boldsymbol{\mu}\|_2 = 4$. To approximate gradient descent, we numerically integrate the gradient flow ODE using a Runge-Kutta method: Since the implicit bias characterizes the asymptotic limit of training, finite-time GD iterates can deviate noticeably from this limit (especially for small λ, see [24, 30] for related discussions). Solving the gradient flow (GF) ODE allows us to efficiently reach this asymptotic regime (in an approximate sense) that would otherwise require prohibitively many GD steps, since the implicit bias of GD and GF are identical in our case.

Appendix C. Deferred Proofs

C.1. Proof of Lemma 3

Before we begin, let us define the *mixed label* as

$$z_{\lambda,i} := (1-\lambda)h_i + \lambda p_i. \tag{15}$$

Then, due to the linearity of $\ell_q(\cdot)$ with respect to the q, we know that the following identity holds:

$$(1-\lambda)\ell_{h_i}(u) + \lambda \ell_{p_i}(u) = \ell_{z_{\lambda,i}}(u). \tag{16}$$

Thus, the problem boils down to the analysis of gradient descent for another soft label $z_{\lambda,i}$.

Risk convergence. Let $\mathbf{z}_\lambda = [z_{\lambda,1}, \ldots, z_{\lambda,n}]^\top \in \mathbb{R}^n$ denote the vector of mixture labels on the training dataset. The empirical KD risk and its gradient can be expressed as

$$\widehat{L}(\mathbf{w}, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \ell_{z_{\lambda,i}}(\mathbf{x}_i^\top \mathbf{w}), \tag{17}$$

$$\nabla_{\mathbf{w}} \widehat{L}(\mathbf{w}, \lambda) = \frac{1}{n}\mathbf{X}^\top \left(\sigma(\mathbf{X}\mathbf{w}) - \mathbf{z}_\lambda\right), \tag{18}$$

where $\sigma(\cdot)$ is applied entrywise. Moreover, since

$$\ell_q'(u) = \sigma(u) - q \tag{19}$$

$$\ell_q''(u) = \sigma(u)(1-\sigma(u)) \le 1/4 \tag{20}$$

holds, the Hessian of the empirical KD risk can be written as

$$\nabla_{\mathbf{w}}^2 \widehat{L}(\mathbf{w}, \lambda) = \frac{1}{n}\mathbf{X}^\top \mathbf{D}(\mathbf{w})\mathbf{X}, \tag{21}$$

where

$$\mathbf{D}(\mathbf{w}) := \mathrm{diag}\Big(\left[\sigma(\mathbf{x}_1^\top \mathbf{w})(1-\sigma(\mathbf{x}_1^\top \mathbf{w})), \ldots, \sigma(\mathbf{x}_n^\top \mathbf{w})(1-\sigma(\mathbf{x}_n^\top \mathbf{w}))\right]\Big). \tag{22}$$

As the sigmoid function satisfies $\sigma(x)(1-\sigma(x)) \le 1/4$, we have

$$\|\nabla_{\mathbf{w}}^2 \widehat{L}(\mathbf{w}, \lambda)\|_2 \le \frac{1}{4n}\|\mathbf{X}\|_2^2. \tag{23}$$

In other words, $\widehat{L}(\cdot, \lambda)$ is $\|\mathbf{X}\|_2^2/4n$-smooth. Moreover, we know that $\hat{L}(\cdot, \lambda)$ is also convex, due to the convexity of $\ell_q(u)$. Thus, we have the risk convergence:

$$\lim_{t\to\infty} \widehat{L}(\mathbf{w}_\lambda(t), \lambda) = \inf_{\mathbf{w}\in\mathbb{R}^d} \widehat{L}(\mathbf{w}, \lambda), \qquad \forall \lambda \in (0,1],\ \eta \in (0, 8n/\|\mathbf{X}\|_2^2). \tag{24}$$

Implicit bias. It remains to characterize the minimizer. Since $\ell_q(u)$ is strictly convex in u for every $q \in (0,1)$, the first-order optimality condition implies that the logit vector $\mathbf{u} = \mathbf{X}\mathbf{w}$ has a unique minimizer

$$\mathbf{u}_\lambda^\star = \sigma^{-1}(\mathbf{z}_\lambda). \tag{25}$$

When $d = n$, we can simply invert the matrix to get the unique optimal weight $\mathbf{w}^\star = \mathbf{X}^{-1}\sigma^{-1}(\mathbf{z}_\lambda)$, to which the gradient descent trajectory converges to. For $d > n$, consider the decomposition $\mathbb{R}^d = \mathrm{range}(\mathbf{X}^\top) \oplus \mathrm{null}(\mathbf{X})$. Then, the set of risk minimizers can be written as

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \widehat{L}(\mathbf{w}, \lambda) = \left\{ \mathbf{X}^\dagger \sigma^{-1}(\mathbf{z}_\lambda) + \mathbf{v} \mid \mathbf{v} \in \mathrm{null}(\mathbf{X}) \right\}. \tag{26}$$

Since

$$\nabla_{\mathbf{w}} \widehat{L}(\mathbf{w}, \lambda) = \frac{1}{n}\mathbf{X}^\top(\sigma(\mathbf{X}\mathbf{w}) - \mathbf{z}_\lambda) \in \mathrm{range}(\mathbf{X}^\top), \tag{27}$$

i.e., every GD update lies in $\mathrm{range}(\mathbf{X}^\top)$. In other words, the component of the iterate in $\mathrm{null}(\mathbf{X})$ is invariant under GD. Since $\mathbf{I}_d - \mathbf{X}^\dagger\mathbf{X}$ is the orthogonal projection onto $\mathrm{null}(\mathbf{X})$, we have, for all GD iteration $t \geq 0$,

$$(\mathbf{I}_d - \mathbf{X}^\dagger\mathbf{X})\mathbf{w}(t) \ = (\mathbf{I}_d - \mathbf{X}^\dagger\mathbf{X})\mathbf{w}(0). \tag{28}$$

The remaining component lies in $\mathrm{range}(\mathbf{X}^\top)$ and converges to the unique minimum-norm interpolating solution, i.e.,

$$\lim_{t\to\infty} \mathbf{X}^\dagger\mathbf{X}\mathbf{w}_\lambda(t) = \mathbf{X}^\dagger\sigma^{-1}(\mathbf{z}_\lambda) \tag{29}$$

Combining the results, we obtain

$$\lim_{t\to\infty} \mathbf{w}_\lambda(t) = (\mathbf{I}_d - \mathbf{X}^\dagger\mathbf{X})\mathbf{w}(0) + \mathbf{X}^\dagger\sigma^{-1}(\mathbf{z}_\lambda), \tag{30}$$

and this completes the proof.

C.2. Proof of Theorem 5

We prove the theorem in the following steps.

- First, we derive the upper bound of the population zero-one risk, including the (1) cosine ratio and the (2) ratio of the inner product term w.r.t. $\mathbf{K}$.
- Next, we define the matrix $\widehat{\mathbf{K}}$, which is a surrogate for $\mathbf{K}$, and lower bound the inner product term w.r.t. $\widehat{\mathbf{K}}$.
- Finally, we transfer this lower bound to $\mathbf{K}$.

Throughout the proof, we write all positive absolute constant as C_k for brevity, where $k \in \mathbb{N}$.

Risk bound in terms of the cosine ratio. Under Assumption 4, the logit of the student model with parameters $\mathbf{w} \in \mathbb{R}^d$ can be written as

$$y \cdot \mathbf{w}^\top \mathbf{x} = \mathbf{w}^\top \boldsymbol{\mu} + y \cdot \mathbf{w}^\top \mathbf{z} \tag{31}$$

Since $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d)$ and y is independent of $\mathbf{z}$, we have $y \cdot \mathbf{w}^\top \mathbf{z} \sim \mathcal{N}\left(0, \|\mathbf{w}\|_2^2\right)$. Therefore,

$$\begin{aligned}
\mathcal{R}(\mathbf{w}) = \Pr[y \cdot \mathbf{w}^\top \mathbf{x} \leq 0] &= \Pr[y \cdot \mathbf{w}^\top \mathbf{z} \leq -\mathbf{w}^\top \boldsymbol{\mu}] && (32)\\
&= \Phi\left(-\frac{\mathbf{w}^\top \boldsymbol{\mu}}{\|\mathbf{w}\|_2}\right) && (33)\\
&\leq \exp\left(-\frac{\left(\mathbf{w}^\top \boldsymbol{\mu}\right)^2}{2\|\mathbf{w}\|_2^2}\right) && (34)
\end{aligned}$$

Here, the last inequality is due the standard Gaussian tail bound, from the assumption of $\mathbf{w}^\top \boldsymbol{\mu} \geq 0$.[1] Plugging in $\mathbf{w}_\lambda^\star$, we have

$$\begin{aligned}
\mathcal{R}\left(\mathbf{w}\right) &\leq \exp\left(-\frac{\left(\mathbf{w}_\lambda^{\star\top} \boldsymbol{\mu}\right)^2}{2\|\mathbf{w}_\lambda^\star\|_2^2}\right) && (35)\\
&= \exp\left(-\frac{\left(\mathbf{a}_\lambda^\top \mathbf{K}\mathbf{s}\right)^2}{2\mathbf{a}_\lambda^\top \mathbf{K}\mathbf{a}_\lambda}\right) && (36)\\
&= \exp\left(-\frac{1}{2}\frac{\cos_{\mathbf{K}}^2\left(\mathbf{a}_\lambda, \mathbf{s}\right)}{\cos_{\mathbf{K}}^2\left(\mathbf{1}_n, \mathbf{s}\right)} \cdot \frac{(\mathbf{1}_n^\top \mathbf{K}\mathbf{s})^2}{\mathbf{1}_n^\top \mathbf{K}\mathbf{1}_n}\right) && (37)\\
&=: \exp\left(-\frac{1}{2}\frac{\cos_{\mathbf{K}}^2\left(\mathbf{a}_\lambda, \mathbf{s}\right)}{\cos_{\mathbf{K}}^2\left(\mathbf{1}_n, \mathbf{s}\right)} \cdot \frac{A^2}{B}\right), && (38)
\end{aligned}$$

where

$$A := \mathbf{1}_n^\top \mathbf{K}\mathbf{s}, \qquad B := \mathbf{1}_n^\top \mathbf{K}\mathbf{1}_n. \tag{39}$$

1. This assumption is satisfied when the distilled student performs no worse than random guessing, and is therefore a mild one.

Lower bound on the surrogate ratio. It remains to derive a lower bound on A^2/B. By rotational invariance of the Gaussian noise, we can assume without loss of generality that

$$\boldsymbol{\mu} = \sqrt{m}\mathbf{e}_1 \quad \text{where} \quad \mathbf{e}_1 := [1, 0, \cdots, 0]^\top, \;\; m := \|\boldsymbol{\mu}\|_2^2. \tag{40}$$

Moreover, we can write as

$$\mathbf{s}_i = \widetilde{\mathbf{x}}_i^\top \boldsymbol{\mu} = m + \sqrt{m}\zeta_i, \qquad \zeta_i \sim \mathcal{N}(0, 1). \tag{41}$$

Then, we may decompose the matrix $\widetilde{\mathbf{X}}$ as

$$\widetilde{\mathbf{X}} = \left[\frac{\mathbf{s}}{\sqrt{m}}, \mathbf{H}\right], \tag{42}$$

where $\mathbf{H} \in \mathbb{R}^{n\times(d-1)}$ has i.i.d. $\mathcal{N}(0, 1)$ entries and is independent of $\mathbf{s}$. Therefore,

$$\mathbf{K} = (\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top)^{-1} \tag{43}$$

$$= (\mathbf{H}\mathbf{H}^\top + \mathbf{s}\mathbf{s}^\top/m)^{-1}. \tag{44}$$

Next, let us define the surrogate matrix and scalars

$$\widehat{\mathbf{K}} := \Big((d-1)\mathbf{I}_n + \mathbf{s}\mathbf{s}^\top/m\Big)^{-1}, \tag{45}$$

$$= \mathbf{I}_n/(d-1) - \mathbf{s}\mathbf{s}^\top/((d-1)D). \tag{46}$$

$$\widehat{A} := \mathbf{1}_n^\top \widehat{\mathbf{K}}\mathbf{s}, \tag{47}$$

$$\widehat{B} := \mathbf{1}_n^\top \widehat{\mathbf{K}}\mathbf{1}_n. \tag{48}$$

Here, Eq. (46) can be obtained from using the Sherman–Morrison formula. For notational brevity, we additionally define the following scalars:

$$T := \mathbf{1}_n^\top \mathbf{s}, \qquad S := \|\mathbf{s}\|_2^2, \qquad D := (d-1)m + S. \tag{49}$$

Using the notations, we can write as

$$\widehat{A} = \frac{mT}{D}, \quad \widehat{B} = \frac{nD - T^2}{(d-1)D} \quad \text{hence} \quad \frac{\widehat{A}^2}{\widehat{B}} = \frac{(d-1)T^2m^2}{D\,(nD - T^2)}. \tag{50}$$

By the Gaussian tail bound and the Lemma 1 from Laurent and Massart [16], with probability at least $1 - O(n^{-1})$ (see Section C.2.1 for the detailed derivation procedure),

$$\left|\sum_{i=1}^{n} \zeta_i\right| \le C_1\sqrt{n\log n}, \qquad \frac{n}{2} \le n\mathrm{Var}_n(\zeta_i) \le 2n, \tag{51}$$

where $\mathrm{Var}_n(\cdot)$ denotes the empirical variance. Since $m \ge C$ and C is sufficiently large, this event implies

$$0.5nm \le T \le 1.5nm \tag{52}$$

$$S = \frac{T^2}{n} + mn\mathrm{Var}_n(\zeta_i) \le \frac{9}{4}nm^2 + 2nm \le C_2 nm\,(m+1) \tag{53}$$

Hence

$$D = (d-1)m + S \tag{54}$$
$$\leq (d-1)m + C_2 nm\,(m+1) \tag{55}$$
$$\leq C_3 m\,(d+nm)\,, \tag{56}$$

where we used $d = \Omega(n^2)$. Therefore,

$$\frac{\widehat{A}^2}{\widehat{B}} = \frac{(d-1)T^2m^2}{D\,(nD - T^2)} \tag{57}$$
$$\geq \frac{(d-1)\,(nm/2)^2\,m^2}{C_3 m\,(d+nm)\cdot nm\,((d-1)+2n)} \tag{58}$$
$$\geq C_4 \frac{d-1}{d-1+n}\cdot\frac{nm^2}{d+nm} \tag{59}$$
$$\geq C_5 \frac{nm^2}{d+nm}. \tag{60}$$

Transferring the lower bound. Define

$$\mathbf{E} := \mathbf{K}^{-1} - \widehat{\mathbf{K}}^{-1} \tag{61}$$
$$= \mathbf{H}\mathbf{H}^\top - (d-1)\mathbf{I}_n. \tag{62}$$

Using the singular value bound for the Gaussian random matrix [34, Theorem 4.6.1], with probability at least $1 - O(n^{-1})$,

$$\|\mathbf{E}\|_2 \leq 2\sqrt{d-1}(\sqrt{n} + \sqrt{2\log n}) + (\sqrt{n} + \sqrt{2\log n})^2 \tag{63}$$
$$\leq C_6(\sqrt{nd} + n + \sqrt{d\log n} + \log n) \tag{64}$$
$$\leq C_6(\sqrt{nd} + n) + C_7(\sqrt{nd} + n) \tag{65}$$
$$\leq C_8(\sqrt{nd} + n). \tag{66}$$

Moreover, since $\widehat{\mathbf{K}}^{-1} \succeq (d-1)\mathbf{I}_n$, we have $\|\widehat{\mathbf{K}}\|_2 \leq 1/(d-1)$. Let $\mathbf{F} := \widehat{\mathbf{K}}^{1/2}\mathbf{E}\widehat{\mathbf{K}}^{1/2}$. From the submultiplicativity of spectral norm, we obtain

$$\|\mathbf{F}\|_2 \leq \|\widehat{\mathbf{K}}\|_2 \cdot \|\mathbf{E}\|_2 \tag{67}$$
$$\leq C_8(\sqrt{nd} + n)\cdot 1/(d-1) \tag{68}$$
$$\leq C_9\left(\sqrt{n/d} + n/d\right) \tag{69}$$
$$\leq 1/2. \tag{70}$$

Here, since $d \geq Cn^2$, taking C sufficiently large, we get the last inequality. Moreover,

$$\mathbf{K} = \left(\widehat{\mathbf{K}}^{-1} + \mathbf{E}\right)^{-1} = \widehat{\mathbf{K}}^{1/2}\left(\mathbf{I}_n + \mathbf{F}\right)^{-1}\widehat{\mathbf{K}}^{1/2}, \tag{71}$$

and this implies

$$\widehat{\mathbf{K}}/2 \preceq \mathbf{K} \preceq 2\widehat{\mathbf{K}}. \tag{72}$$

From the inequality, we obtain

$$B = \mathbf{u}^\top \mathbf{K}\mathbf{u} \leq 2\mathbf{u}^\top \widehat{\mathbf{K}}\mathbf{u} = 2\widehat{B}. \tag{73}$$

We next compare A and $\widehat{A}$.

$$\begin{aligned}
\left|A - \widehat{A}\right| &= \left|\mathbf{u}^\top \left(\mathbf{K} - \widehat{\mathbf{K}}\right)\mathbf{s}\right| && (74)\\
&= \left|\left\langle \widehat{\mathbf{K}}^{1/2}\mathbf{u}, \left((\mathbf{I}_n + \mathbf{F})^{-1} - \mathbf{I}_n\right)\widehat{\mathbf{K}}^{1/2}\mathbf{s}\right\rangle\right| && (75)\\
&\leq \frac{\|\mathbf{F}\|_2}{1 - \|\mathbf{F}\|_2}\|\mathbf{u}\|_{\widehat{\mathbf{K}}}\|\mathbf{s}\|_{\widehat{\mathbf{K}}} && (76)\\
&\leq 2\|\mathbf{F}\|_2\|\mathbf{u}\|_{\widehat{\mathbf{K}}}\|\mathbf{s}\|_{\widehat{\mathbf{K}}} && (77)\\
&\leq C_{10}\|\mathbf{F}\|_2\widehat{A} && (78)\\
&\leq \widehat{A}/2. && (79)
\end{aligned}$$

where the detailed derivation of the inequalities can be found in Section C.2.1. Combining this with $B \leq 2\widehat{B}$ gives

$$\begin{aligned}
A^2/B &\geq (\widehat{A}/2)^2/2\widehat{B} && (80)\\
&= \widehat{A}^2/8\widehat{B} && (81)\\
&\geq C_{11}\frac{nm^2}{d + nm}, && (82)
\end{aligned}$$

Plugging Eq. (82) into Eq. (37) and taking a union bound over the all events gives probability at least $1 - O(n^{-1})$, and this concludes the proof.

C.2.1. DETAILED DERIVATION

Derivation of Eq. (51) We have $\sum_{i=1}^{n} \zeta_i \sim \mathcal{N}(0, n)$. Hence, by the standard Gaussian tail bound, for any $t > 0$,

$$\Pr\left(\left|\sum_{i=1}^{n} \zeta_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2n}\right). \tag{83}$$

Taking $t = C_1\sqrt{n\log n}$ gives

$$\Pr\left(\left|\sum_{i=1}^{n} g_i\right| \geq C_1\sqrt{n\log n}\right) \leq 2\exp\left(-\frac{C_1^2\log n}{2}\right) \tag{84}$$

$$= 2n^{-C_1^2/2}. \tag{85}$$

Therefore, by choosing $C_1 > 2$, we obtain

$$\left|\sum_{i=1}^{n} g_i\right| \leq C_1\sqrt{n\log n} \tag{86}$$

with probability at least $1 - O(n^{-1})$. Next, let

$$\bar{\zeta} := \frac{1}{n}\sum_{i=1}^{n} \zeta_i, \qquad \mathrm{Var}_n(\zeta_i) := \frac{1}{n}\sum_{i=1}^{n} \left(\zeta_i - \bar{\zeta}\right)^2 \sim \chi^2_{n-1}, \tag{87}$$

since ζ_i are i.i.d. standard Gaussian random variables. Here, χ^2_{n-1} denotes the Chi-square random variable with $(n-1)$ degree of freedom. From Laurent and Massart [16], for $Q \sim \chi^2_k$ and any $x > 0$, we have

$$\Pr\left(Q \geq k + 2\sqrt{kx} + 2x\right) \leq e^{-x}, \tag{88}$$

$$\Pr\left(Q \leq k - 2\sqrt{kx}\right) \leq e^{-x}. \tag{89}$$

Applying this with $Q = n\mathrm{Var}_n(\zeta_i)$, $k = n-1$, and $x = c\log n$, we get

$$n - 1 - 2\sqrt{(n-1)c\log n}. \leq n\mathrm{Var}_n(\zeta_i) \leq n - 1 + 2\sqrt{(n-1)c\log n} + 2c\log n, \tag{90}$$

Thus, for sufficiently large n,

$$\frac{n}{2} \leq n\mathrm{Var}_n(\zeta_i) \leq 2n \tag{91}$$

with probability at least $1 - O(n^{-1})$.

Combining the two events by a union bound, we get what we want.

Derivation of Eq. (79) From $\mathbf{s}^\top \widehat{\mathbf{K}}\mathbf{s} = mS/D$, we have

$$\cos^2_{\widehat{\mathbf{K}}}(\mathbf{u}, \mathbf{s}) = \frac{\widehat{A}^2}{\widehat{B}\left(\mathbf{s}^\top \widehat{\mathbf{K}}\mathbf{s}\right)} \tag{92}$$

$$= \frac{(d-1)T^2 m}{S\left(nD - T^2\right)} \tag{93}$$

$$\geq C_{12} \frac{d-1}{d-1+n} \cdot \frac{m}{m+1} \tag{94}$$

$$\geq C_{13}, \tag{95}$$

for an absolute constant $c_{\text{angle}} > 0$. This implies $\|\mathbf{u}\|_{\widehat{\mathbf{K}}} \|\mathbf{s}\|_{\widehat{\mathbf{K}}} \leq C_{14}\widehat{A}$, and we get Eq. (79). By increasing C if necessary, we may assume $2C_{14}\|\mathbf{F}\|_2 \leq 1/2$, hence $A \geq \widehat{A}/2$.

C.3. Proof of Lemma 6

We prove the lemma in the following steps. Similar to Section C.2, we write an positive absolute constant as C_k.

Characterizing the Bayes teacher. Since the data model is symmetric Gaussian mixture, output of the Bayes teacher can be characterized as

$$\begin{aligned} p_i &= \Pr(y=+1|\mathbf{x}) && (96)\\ &= \frac{\Pr(\mathbf{x}|y=+1)}{\Pr(\mathbf{x}|y=+1)+\Pr(\mathbf{x}|y=-1)} && (97)\\ &= \frac{\Pr(\mathbf{x}|y=+1)/\Pr(\mathbf{x}|y=-1)}{1+\Pr(\mathbf{x}|y=+1)/\Pr(\mathbf{x}|y=-1)} && (98)\\ &= \frac{1}{1+\exp(-2\boldsymbol{\mu}\mathbf{x})} && (99)\\ &= \sigma(2\boldsymbol{\mu}^\top\mathbf{x}). && (100)\end{aligned}$$

In other words, the Bayes teacher is parametrized as $\mathcal{T}(\mathbf{x}) = 2\boldsymbol{\mu}^\top\mathbf{x}$. In this case, by letting $s_i = y_i\boldsymbol{\mu}^\top\mathbf{x}_i$, we get

$$\begin{aligned} g_\lambda(s_i) &:= y_i\sigma^{-1}((1-\lambda)h_i+\lambda p_i) && (101)\\ &= \log\left(\frac{1-\lambda(1+\exp(2s_i))^{-1}}{\lambda(1+\exp(2s_i))^{-1}}\right) && (102)\\ &= \log\left(\frac{1+\exp(2s_i)-\lambda}{\lambda}\right). && (103)\end{aligned}$$

We now write $\mathbf{a}_\lambda := [g_\lambda(s_1),\cdots,g_\lambda(s_n)]^\top$. Now, we follow the similar approaches as in Section C.2, we use the surrogate matrix $\widehat{\mathbf{K}}$. Using the definition in Section C.2, we can write as

$$\begin{aligned} \mathbf{a}_\lambda^\top\widehat{\mathbf{K}}\mathbf{s} &= m\mathbf{a}_\lambda^\top\mathbf{s}/D, && (104)\\ \mathbf{s}^\top\widehat{\mathbf{K}}\mathbf{s} &= mS/D && (105)\\ \mathbf{a}_\lambda^\top\widehat{\mathbf{K}}\mathbf{a}_\lambda &= (D\|\mathbf{a}_\lambda\|_2^2-(\mathbf{a}_\lambda^\top\mathbf{s})^2)/((d-1)D). && (106)\end{aligned}$$

Let us define the surrogate improvement factor

$$\begin{aligned} \widehat{\mathsf{Imp}}(\lambda,\mathcal{T}) &:= \cos^2_{\widehat{\mathbf{K}}}(\mathbf{a}_\lambda,\mathbf{s})/\cos^2_{\widehat{\mathbf{K}}}(\mathbf{1}_n,\mathbf{s}) && (107)\\ &= \frac{(\mathbf{a}^\top\mathbf{s})^2(nD-(\mathbf{u}^\top\mathbf{s})^2)}{(\mathbf{u}^\top\mathbf{s})^2(D\|\mathbf{a}_\lambda\|_2^2-(\mathbf{a}^\top\mathbf{s})^2)} && (108)\\ &= \frac{\bar{A}_\lambda^2\left((d-1)m+n(\bar{S}-\bar{T}^2)\right)}{\bar{T}^2\left((d-1)m\bar{V}_\lambda+n(\bar{S}\bar{V}_\lambda-\bar{A}_\lambda^2)\right)}. && (109)\end{aligned}$$

where $\bar{\cdot}$ denotes the empirical average over n training samples.

Let $Z \sim \mathcal{N}(m, m)$ denote a random variable with the same distribution as each s_i. By Gaussian and sub-exponential concentration inequalities, for any fixed $\lambda \in (0, 1]$, we have

$$\bar{T} = m\left(1 \pm C_1\sqrt{\log n/n}\right) \tag{110}$$

$$\bar{S} = (m^2 + m)\left(1 \pm C_1\sqrt{\log n/n}\right) \tag{111}$$

$$\bar{A}_\lambda = \mathbb{E}[Zg_\lambda(Z)]\left(1 \pm C_1\sqrt{\log n/n}\right) \tag{112}$$

$$\bar{V}_\lambda = \mathbb{E}[g_\lambda(Z)^2]\left(1 \pm C_1\sqrt{\log n/n}\right). \tag{113}$$

Substituting these estimates into Eq. (109), we obtain

$$\widehat{\mathsf{Imp}}(\lambda, \mathcal{T}) = \frac{(d-1+n)J_\lambda}{d-1+n(m+1-mJ_\lambda)}\left(1 \pm C_2\sqrt{\frac{\log n}{n}}\right), \qquad J_\lambda := \frac{\mathbb{E}[Zg_\lambda(Z)]^2}{m^2\mathbb{E}[g_\lambda(Z)^2]}. \tag{114}$$

Transferring the estimate. We now transfer the surrogate $\widehat{\mathsf{Imp}}(\lambda, \mathcal{T})$ into $\mathsf{Imp}(\lambda, \mathcal{T})$. Using the same inverse-perturbation argument as in the Section C.2, where $\mathbf{K} = (\widehat{\mathbf{K}}^{-1}+\mathbf{E})^{-1}$ and $\|\widehat{\mathbf{K}}^{1/2}\mathbf{E}\widehat{\mathbf{K}}^{1/2}\|_2 = O(\sqrt{n/d})$, we obtain

$$\mathsf{Imp}(\lambda, \mathcal{T}) = \frac{(d-1+n)J_\lambda}{d-1+n(m+1-mJ_\lambda)}\left(1 \pm C_3\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{n}{d}}\right)\right). \tag{115}$$

To derive the bound for Imp, we derive the upper bound of J_λ:

$$J_\lambda = \frac{\mathbb{E}[Zg_\lambda(Z)]^2}{m^2\mathbb{E}[g_\lambda(Z)^2]} \tag{116}$$

$$\leq \frac{\mathbb{E}[Z^2]\mathbb{E}[g_\lambda(Z)^2]}{m^2\mathbb{E}[g_\lambda(Z)^2]} \tag{117}$$

$$= 1 + 1/m. \tag{118}$$

Next, we establish the lower bound for J_λ:

$$J_\lambda = \frac{\mathbb{E}[Zg_\lambda(Z)]^2}{m^2\mathbb{E}[g_\lambda(Z)^2]} \tag{119}$$

$$= \frac{(\mathbb{E}[g_\lambda(Z)] + \mathbb{E}[g'_\lambda(Z)])^2}{\mathbb{E}[g_\lambda(Z)]^2 + \mathrm{Var}(g_\lambda(Z))} \tag{120}$$

$$\geq \frac{(\mathbb{E}[g_\lambda(Z)] + \mathbb{E}[g'_\lambda(Z)])^2}{\mathbb{E}[g_\lambda(Z)]^2 + \mathbb{E}[g_\lambda(Z)]\mathbb{E}[g'_\lambda(Z)]} \tag{121}$$

$$= 1 + \frac{\mathbb{E}[g'_\lambda(Z)]}{\mathbb{E}[g_\lambda(Z)]} > 1 \tag{122}$$

Here, we use Stein's lemma in Eq. (120): $\mathbb{E}[Zg_\lambda(Z)] = m\mathbb{E}[g_\lambda(Z)] + m\mathbb{E}[g'_\lambda(Z)]$. Eq. (121) follows from the Gaussian Poincaré inequality and the facts that $0 < g'_\lambda(t) \leq 2$ and $g_\lambda(t) \geq 2t$, which imply $\mathrm{Var}(g_\lambda(Z)) \leq m\mathbb{E}[g'_\lambda(Z)^2] \leq 2m\mathbb{E}[g'_\lambda(Z)] \leq \mathbb{E}[g_\lambda(Z)]\mathbb{E}[g'_\lambda(Z)]$. Combining the lower and the upper bound, we obtain

$$1 < J_\lambda \leq 1 + 1/m. \tag{123}$$

From Eq. (115), let

$$f(J_\lambda, n, d) := \frac{(d-1+n)J_\lambda}{d-1+n(m+1-mJ_\lambda)}. \tag{124}$$

Since this term is increasing w.r.t J_λ since $f'(J_\lambda, n, d) > 0$, we have

$$f(1, n, d) < f(J_\lambda, n, d) < f(1+1/m, n, d), \tag{125}$$

where

$$f(1, n, d) = 1, \qquad f(1+1/m, n, d) = \frac{d-1+n}{d-1}(1+1/m) \tag{126}$$

Plugging this into Eq. (115), for some $C > 0$, we get

$$1 - C\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{n}{d}}\right) < \mathsf{Imp}(\lambda, \mathcal{T}) < \frac{d-1+n}{d-1}\left(1+\frac{1}{m}\right)\left(1 + C\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{n}{d}}\right)\right), \tag{127}$$

with probability at least $1 - O(n^{-1})$. Since $d = \Omega(n^2)$ and $\log(n)/n = o(1)$, in a approximate sense, we obtain

$$1 \lesssim \mathsf{Imp}(\lambda, \mathcal{T}) \lesssim 1 + \frac{1}{\|\boldsymbol{\mu}\|_2^2}, \tag{128}$$

and this concludes the proof.

C.4. Proof of Lemma 7

We using the same notation as in Sections C.2 and C.3. Since the teacher is $\mathcal{T}(\mathbf{x}) = 2\boldsymbol{\mu}^\top \mathbf{x} + b$, we have

$$y_i \mathcal{T}(\mathbf{x}_i) = 2s_i + by_i. \tag{129}$$

In this case, J_λ in Eq. (114) becomes $J_{\lambda,b}$, which is

$$J_{\lambda,b} := \frac{\mathbb{E}[Z g_\lambda(2Z + bY)]^2}{m^2 \mathbb{E}[g_\lambda(2Z + bY)^2]} \quad \text{where} \quad g_\lambda(t) = \log\left(\frac{\exp(t) + 1 - \lambda}{\lambda}\right), \tag{130}$$

and $Y \sim \mathrm{Unif}(\{-1, +1\})$.

For the denominator of $J_{\lambda,b}$, we have

$$\mathbb{E}[Z g_\lambda(2Z + bY)] = m\mathbb{E}[g_\lambda(2Z + bY)] + m\mathbb{E}\left[\frac{\partial}{\partial Z} g_\lambda(2Z + bY)\right] \tag{131}$$

$$= m(\mathbb{E}[g_\lambda(2Z + bY)] + 2\mathbb{E}[g'_\lambda(2Z + bY)]). \tag{132}$$

From now on, we write $Q_b := g_\lambda(2Z + bY)$ and $Q'_b := g'_\lambda(2Z + bY)$ for simplicity. Then, we have

$$J_{\lambda,b} = \frac{m(\mathbb{E}[Q_b] + 2\mathbb{E}[Q'_b])^2}{m^2 \mathbb{E}[Q_b^2]} = \frac{(\mathbb{E}[Q_b] + 2\mathbb{E}[Q'_b])^2}{\mathbb{E}[Q_b]^2 + \mathrm{Var}[Q_b]}. \tag{133}$$

For $\mathrm{Var}[Q_b]$, we have

$$\mathrm{Var}[Q_b] = \mathbb{E}_Y[\mathrm{Var}[Q_b|Y]] + \mathrm{Var}_Y[\mathbb{E}[Q_b|Y]] \tag{134}$$

$$\leq 4m\mathbb{E}[Q'_b] + \mathrm{Var}_Y[\mathbb{E}[Q_b|Y]] \tag{135}$$

$$\leq 4m\mathbb{E}[Q'_b] + b^2. \tag{136}$$

Here, we use Gaussian Poincaré inequality in Eq. (135), and 1-Lipschitzness of $g_\lambda(x)$ in Eq. (136).

Next, we compute $\mathbb{E}[Q_b]$:

$$\mathbb{E}[Q_b] = \mathbb{E}[g_\lambda(2Z + bY)] \geq \mathbb{E}[2Z + bY + \log(1/\lambda)] = 2m + \log(1/\lambda). \tag{137}$$

Since $2Z + bY$ is symmetric random variable, with Lemma 10, we get $\mathbb{E}[Q'_b] \geq 0.5$. Thus, when plugging the derived terms, we have $J_{\lambda,b} > 1$ when

$$|b| < \sqrt{2\|\boldsymbol{\mu}\|_2^2 + 2\log(1/\lambda) + 1}, \tag{138}$$

and we get what we want. Here, note that $\mathsf{lmp} > 1$ when $J_{\lambda,b} > 1$, up to $o(1)$ error term.

Remark. For the proofs of Lemma 8 and Lemma 9, we note that we follow the same derivation as in Section C.4. That is, we use the same notation J_λ, and obtain $\mathsf{Imp}(\lambda, \mathcal{T}) > 1$ when $J_\lambda > 1$.

C.5. Proof of Lemma 8

First, let us define the random variable that indicates the label-flip probability ρ as F, i.e., $\Pr(F = -1) = \rho$. Recall that $Z \sim \mathcal{N}(m, m)$. Then, for the label-flipped teacher, the logit is

$$y_i \cdot 2F_i \boldsymbol{\mu}^\top \mathbf{x}_i = 2F_i \mathbf{s}_i. \tag{139}$$

To derive Imp, we follow the same approaches as in Section C.4: Recall that $g_\lambda(t) = \log\left(\frac{\exp(t)+1-\lambda}{\lambda}\right)$. Then we have

$$J_{\lambda,\rho} = \frac{\mathbb{E}[Z g_\lambda(2FZ)]^2}{m^2 \mathbb{E}[g_\lambda(2FZ)^2]} \tag{140}$$

$$= \frac{((1-\rho)\mathbb{E}[Z g_\lambda(2Z)] + \rho \mathbb{E}[Z g_\lambda(-2Z)])^2}{m^2((1-\rho)\mathbb{E}[g_\lambda(2Z)^2] + \rho \mathbb{E}[g_\lambda(-2Z)^2])}. \tag{141}$$

However, directly analyzing Eq. (141) is hard, thus we consider the case of pure distillation (i.e., $\lambda = 1$). When we set $\lambda = 1$, we have $g_1(t) = t$. Then

$$J_{1,\rho} = \frac{\mathbb{E}[Z \cdot 2FZ]^2}{m^2 \mathbb{E}[(2FZ)^2]} \tag{142}$$

$$= \frac{4\mathbb{E}[F]^2 \mathbb{E}[Z^2]^2}{4m^2 \mathbb{E}[Z^2]} \tag{143}$$

$$= (1-2\rho)^2(1+1/m). \tag{144}$$

Thus, up to $o(1)$ error term, we have $\mathsf{Imp}(1, \mathcal{T}_\rho) > 1$ whenever

$$\rho < 0.5 - \sqrt{\frac{\|\boldsymbol{\mu}\|_2^2}{4\|\boldsymbol{\mu}\|_2^2 + 4}}, \tag{145}$$

and this concludes the proof.

C.6. Proof of Lemma 9

Before begin, let us decompose the teacher's weight vector $\mathbf{v} \in \mathbb{R}^d$ into the component aligned with $\boldsymbol{\mu}$ and the orthogonal component $\mathbf{v}_\perp$, i.e.,

$$\mathbf{v} = \frac{\mathbf{v}^\top \boldsymbol{\mu}}{m} \boldsymbol{\mu} + \mathbf{v}_\perp. \tag{146}$$

From $\mathcal{T}_\mathbf{v}(\mathbf{x}) = \mathbf{v}^\top \mathbf{x}$, for simplicity, we write

$$r_i = y_i \mathcal{T}_\mathbf{v}(\mathbf{x}_i) = y_i \mathbf{v}^\top \mathbf{x}_i. \tag{147}$$

Recall the definition of Z (i.e., $Z \sim \mathcal{N}(m, m)$) and define $\nu \sim \mathcal{N}(0, \|\mathbf{v}_\perp\|_2^2)$. Then, r_i has the same distribution as

$$R := \frac{\mathbf{v}^\top \boldsymbol{\mu}}{m} Z + \nu. \tag{148}$$

Similar to Section C.4, let us define $Q_\mathbf{v} := g_\lambda(R)$ and $Q'_\mathbf{v} := g'_\lambda(R)$. Next, we define $J_{\lambda,\mathbf{v}}$ and proceed as

$$\begin{aligned} J_{\lambda,\mathbf{v}} &:= \frac{\mathbb{E}[ZQ_\mathbf{v}]}{m^2 \mathbb{E}[Q_\mathbf{v}^2]} && (149)\\ &= \frac{m\mathbb{E}[Q_\mathbf{v}] + (\mathbf{v}^\top \boldsymbol{\mu})\mathbb{E}[Q'_\mathbf{v}]}{m^2 \mathbb{E}[Q_\mathbf{v}^2]} && (150)\\ &= \frac{m\mathbb{E}[Q_\mathbf{v}] + (\mathbf{v}^\top \boldsymbol{\mu})\mathbb{E}[Q'_\mathbf{v}]}{\mathbb{E}[Q_\mathbf{v}]^2 + \mathrm{Var}[Q_\mathbf{v}]}. && (151) \end{aligned}$$

By Gaussian Poincaré inequality, we have

$$\mathrm{Var}[Q_\mathbf{v}] \le \left(\frac{(\mathbf{v}^\top \boldsymbol{\mu})^2}{m} + \|\mathbf{v}_\perp\|_2^2 \right)^2 \mathbb{E}[Q'_\mathbf{v}]. \tag{152}$$

Thus, taking a similar approach to Section C.4, we have $J_{\lambda,\mathbf{v}} > 1$ when

$$\begin{aligned} \|\mathbf{v}_\perp\|_2^2 &< \frac{2(\mathbf{v}^\top \boldsymbol{\mu})\mathbb{E}[Q_\mathbf{v}]}{m} + \frac{(\mathbf{v}^\top \boldsymbol{\mu})^2 \mathbb{E}[Q'_\mathbf{v}]}{m^2} - \frac{(\mathbf{v}^\top \boldsymbol{\mu})^2}{m} && (153)\\ &\le \frac{2(\mathbf{v}^\top \boldsymbol{\mu})(\mathbf{v}^\top \boldsymbol{\mu} + \log(1/\lambda))}{m} + \frac{(\mathbf{v}^\top \boldsymbol{\mu})^2}{2m^2} - \frac{(\mathbf{v}^\top \boldsymbol{\mu})^2}{m}, && (154) \end{aligned}$$

up to $o(1)$ error term. Rearranging the term, we get what we want.

C.7. Technical Lemma

Lemma 10 *Let X be a symmetric random variable and $g_\lambda(x) = \log\left((\exp(x) + 1 - \lambda)/x\right)$. Then, for any $y \geq 0$, we have*

$$\mathbb{E}[g'_\lambda(X + y)] \geq 0.5. \tag{155}$$

Proof Since g'_λ is an increasing function, we have $\mathbb{E}[g'_\lambda(X + y)] \geq \mathbb{E}[g'_\lambda(X)]$. Also, since X is symmetric, we have $\mathbb{E}[g'_\lambda(X)] = 0.5\mathbb{E}[g'_\lambda(X) + g'_\lambda(-X)]$. Since

$$g'_\lambda(x) + g'_\lambda(-x) = \frac{\exp(x)}{\exp(x) + 1 - \lambda} + \frac{1}{(1 - \lambda)\exp(x) + 1} \geq 1, \tag{156}$$

we finally have

$$\mathbb{E}[g'_\lambda(X + y)] \geq \mathbb{E}[g'_\lambda(X)] \geq 0.5 \tag{157}$$

This concludes the proof. ■