Provable Benefit of Mixup for Finding Optimal Decision Boundaries

Junsoo Oh [1] **Chulhee Yun** [1]

Abstract

We investigate how pair-wise data augmentation techniques like Mixup affect the sample complexity of finding optimal decision boundaries in a binary linear classification problem. For a family of data distributions with a separability constant κ, we analyze how well the optimal classifier in terms of training loss aligns with the optimal one in test accuracy (i.e., Bayes optimal classifier). For vanilla training without augmentation, we uncover an interesting phenomenon named the *curse of separability*. As we increase κ to make the data distribution more separable, the sample complexity of vanilla training *increases exponentially* in κ; perhaps surprisingly, the task of finding optimal decision boundaries becomes harder for more separable distributions. For Mixup training, we show that Mixup mitigates this problem by *significantly reducing* the sample complexity. To this end, we develop new concentration results applicable to n^2 pair-wise augmented data points constructed from n independent data, by carefully dealing with dependencies between overlapping pairs. Lastly, we study other masking-based Mixup-style techniques and show that they can *distort* the training loss and make its minimizer converge to a suboptimal classifier in terms of test accuracy.

1. Introduction

Mixup (Zhang et al., 2018) is a modern technique that augments training data with a random convex combination of a pair of training points and labels. Zhang et al. (2018) empirically show that this simple technique has various benefits, such as better generalization, robustness to label corruption and adversarial attack, and stabilization of generative adversarial network training. Inspired by the success of Mixup, several variants of Mixup have appeared in the literature; e.g., Manifold Mixup (Verma et al., 2019), Cutmix (Yun et al., 2019), Puzzle Mix (Kim et al., 2020), and Co-Mixup (Kim et al., 2021). The success of Mixup-style training schemes is not only limited to improved generalization performance in supervised learning; they are known to be helpful in other aspects including model calibration (Thulasidasan et al., 2019), semi-supervised learning (Berthelot et al., 2019; Sohn et al., 2020), contrastive learning (Kalantidis et al., 2020; Verma et al., 2021), and natural language processing (Guo et al., 2019; Sun et al., 2020).

Although Mixup and its variants demonstrate surprising empirical benefits, a concrete theoretical understanding of such benefits still remains mysterious. As a result, a recent line of research (Carratino et al., 2020; Zhang et al., 2020; 2022; Chidambaram et al., 2021; 2022; Park et al., 2022) trying to theoretically understand Mixup and its variants has appeared in the literature. Most results, including this paper, compare the training procedure with Mixup against solving the vanilla empirical risk minimization (ERM) problem without any augmentation. For the remainder of this paper, we refer to the training loss without augmentation as *ERM loss*, and the training loss with Mixup as *Mixup loss*.

The parallel works of Carratino et al. (2020) and Zhang et al. (2020) represent the Mixup loss as the ERM loss equipped with additive data-dependent regularizers which penalize the gradient and Hessian of a model with respect to data. Using these regularizers, Zhang et al. (2020) show that Mixup training can yield smaller Rademacher complexity which allows for a smaller uniform generalization bound. Also, Park et al. (2022) extend these results to several Mixup variants such as Cutmix (Yun et al., 2019). However, the benefits of Mixup-style schemes from the Rademacher complexity view shown by Zhang et al. (2020) and Park et al. (2022) stand out only if the intrinsic dimension of data is small.

Chidambaram et al. (2021) try to understand Mixup by investigating when Mixup training minimizes the ERM loss. The authors show failure cases of Mixup and also sufficient conditions for its success. Chidambaram et al. (2021) also show that ERM loss and Mixup loss can have the same optimal classifier by considering linear model and special types of data distribution. Detailed comparisons to our work are provided in Section 2.

[1]Kim Jaechul Graduate School of AI, KAIST. Correspondence to: Chulhee Yun <chulhee.yun@kaist.ac.kr>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Another recent work (Chidambaram et al., 2022) shows that Midpoint Mixup (Guo, 2021) can outperform ERM training in terms of feature learning performance in a multi-view data framework proposed by Allen-Zhu & Li (2020). However, their analysis is limited to an extreme case of Mixup using only the midpoint of two data points.

1.1. Our Contributions

In this paper, we study the problem of finding optimal decision boundaries in a binary linear classification problem with logistic loss (i.e., logistic regression). We consider a data generating distribution whose positive and negative examples come from two symmetric Gaussian distributions. The two distributions have the same covariance matrices inversely proportional to a *separability constant* κ, which controls the degree to which the two classes are separable.

Our data generating distribution has the advantage that the *Bayes optimal classifier* (i.e., the classifier with the best test accuracy) is given as a closed-form linear classifier. This motivates us to study how well the optimal classifier in terms of training loss is aligned with the Bayes optimal classifier, and how many data points are required to make sure that the two classifiers are close enough. Therefore, we focus on the sample complexity of vanilla ERM training and Mixup-style training for achieving close-to-one cosine similarity between the two classifiers.

Our results demonstrate that Mixup provably requires a much smaller number of samples to achieve the same cosine similarity compared to vanilla ERM training. Our contributions can be summarized as the following:

- In Section 3, we investigate the sample complexity of (vanilla) ERM under our data distribution. Theorem 3.1 shows that the expected value of ERM training loss has a unique optimal solution aligned with the Bayes optimal classifier. We then prove that the sample complexity for making the ERM loss optimum closely aligned with the Bayes optimal classifier grows *exponentially* with the separability constant κ; we show that the exponential growth is both sufficient (Theorem 3.4) and necessary (Theorem 3.7). Interestingly, these results demonstrate that *ERM suffers the curse of separability*, where the sample complexity increases as the data distribution becomes more separable.
- In Section 4, we study a unified class of Mixup-style training scheme that includes Mixup (Zhang et al., 2018) with various choices of hyperparameters. Theorem 4.2 shows that the expected value of Mixup loss has a unique optimal solution still aligned with the Bayes optimal classifier, which indicates that Mixup-style augmentations do not "distort" the training loss in a misleading way, at least in our setting. In Theorem 4.4, we show that the sample complexity for getting a near-Bayes-optimal Mixup loss solution grows only quadratically in κ. This result indicates that *Mixup provably mitigates the curse of separability*.
- In Section 5, we analyze how recent masking-based variants of Mixup such as CutMix (Yun et al., 2019) behave in our setting. Unfortunately, Theorem 5.3 shows that masking-based augmentation can *distort* the training loss and the expected value of masking-based Mixup loss can have optimal solutions far away from the Bayes optimal classifier. Ironically, Theorem 5.4 indicates the sample complexity for approaching such a "wrong" minimizer does not grow with separability. These results show that, at least in our setting, masking-based techniques have small sample complexity but may not converge to the Bayes optimal classifier.

2. Problem Setting and Notation

In this section, we introduce our formal problem setting and notation. We consider a binary linear classification problem with training dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ are data points and $y_i \in \{0, 1\}$ are labels. The vanilla empirical risk minimization (ERM) loss under training set S can be formulated as follows:

$$\mathcal{L}_S(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n y_i l(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1 - y_i) l(-\boldsymbol{w}^\top \boldsymbol{x}_i), \quad (1)$$

where $l(\cdot)$ is the logistic loss $l(z) := \log(1 + \exp(-z))$.

Data Generating Distribution. We mainly focus on the following family of data generating distributions, where we obtain positive and negative examples from two symmetric Gaussian distributions. More precisely, for a given constant $\kappa \in (0, \infty)$, we define a data generating distribution $\mathcal{D}_\kappa$ as the following: we say that $(\boldsymbol{x}, y) \sim \mathcal{D}_\kappa$ when y is uniformly drawn from $\{0, 1\}$ and

$$\boldsymbol{x} \mid y = 1 \sim N(\boldsymbol{\mu}, \kappa^{-1}\boldsymbol{\Sigma}),$$
$$\boldsymbol{x} \mid y = 0 \sim N(-\boldsymbol{\mu}, \kappa^{-1}\boldsymbol{\Sigma}).$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is nonzero, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is positive definite, and $\|\boldsymbol{\mu}\|^2 = \|\boldsymbol{\Sigma}\|$. Here, $\|\cdot\|$ denotes the ℓ_2 norm for a vector and the spectral norm for a matrix. We further define a data generating distribution $\mathcal{D}_\infty$ as the limiting behavior of $\mathcal{D}_\kappa$ as $\kappa \to \infty$; i.e., $(\boldsymbol{x}, y) \sim \mathcal{D}_\infty$ implies y follows uniform distribution on $\{0, 1\}$ and $\boldsymbol{x} = (2y - 1)\boldsymbol{\mu}$.

Separability Constant κ. Our data generating distribution $\mathcal{D}_\kappa$ with larger κ is more *separable* in the sense that the two Gaussian distributions overlap less and they are more likely to generate well-separated training data. This separability constant κ is of great importance in our analysis, as we will demonstrate the curse of separability phenomenon based on the dependency of sample complexity on κ.

Bayes Optimal Classifier. The Bayes optimal classifier is a classifier achieving the lowest possible test error on the data population. In other words, any other classifier cannot outperform the Bayes optimal classifier in terms of test accuracy. For the specific form of data generating distribution $\mathcal{D}_\kappa$ that we consider in this paper, it is well-known that for any $\kappa \in (0, \infty)$, decision boundary of the Bayes optimal classifier for $\mathcal{D}_\kappa$ is a hyperplane with normal vector $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. For completeness, we provide proof for this in Appendix D.1. The fact that we have a closed-form solution of the optimal decision boundary motivates us to study finding a decision boundary close enough to that optimal one. Since we are mainly interested in linear models, for our "closeness" metric, we use cosine similarity between the two linear decision boundaries, or equivalently, the two normal vectors.

Comparison to the Setting of Chidambaram et al. (2021). Chidambaram et al. (2021) also consider the linear model and a special type of data generating distributions to compare training ERM loss vs. Mixup loss. Our setting is different from theirs in several aspects. First, there is a difference in data generating distributions; we draw positive and negative data from two symmetric Gaussian distributions while Chidambaram et al. (2021) consider the distribution that obtains data from a single spherical Gaussian distribution. Chidambaram et al. (2021) show that training ERM loss and Mixup loss can lead to the same solution by considering the highly overparametrized regime. On the other hand, we consider the underparametrized regime and show the gap between the two training methods. In addition, Chidambaram et al. (2021) only consider the optimal classifier in terms of training loss while we investigate the number of required training data to get close to the optimal solution of population loss.

Notation. We denote taking expectation on training data $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{D}_\kappa$ and any other randomness, if any, as $\mathbb{E}_\kappa$ for each $\kappa \in (0, \infty]$. We use $[k]$ for the index set $\{1, 2, \ldots, k\}$ for each $k \in \mathbb{N}$. For two nonzero vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, let us denote their cosine similarity as $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u}^\top \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}$ and the angle between them as $\angle(\boldsymbol{u}, \boldsymbol{v}) = \cos^{-1}(\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}))$. We use $\mathrm{support}(\cdot)$ to denote the support of a probability distribution or a random variable and $\mathbf{1}_{(\cdot)}$ denotes the 0-1 indicator function. We also use $\odot$ to denote element-wise multiplication between two vectors or two matrices having the same size. We use $\mathcal{O}(\cdot), \Theta(\cdot), \Omega(\cdot)$ to represent asymptotic behavior as κ grows and to hide terms related to $d, \boldsymbol{\mu}, \boldsymbol{\Sigma}$. In addition, whenever we express the κ dependency in $\mathcal{O}(\cdot), \Theta(\cdot), \Omega(\cdot)$ notation, we only write the most dominant factor; for example, we say $\kappa^m \mathrm{polylog}(\kappa) = \Theta(\kappa^m)$ and $\exp(c\kappa^m)\,\mathrm{poly}(\kappa) = \exp(\Theta(\kappa^m))$.

3. ERM Suffers the Curse of Separability

In this section, we investigate a solution of ERM loss in Equation 1. The ERM loss function $\mathcal{L}_S(\boldsymbol{w})$ is a stochastic function that depends on the random samples in the training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{D}_\kappa$. We will first characterize the minimizer of the expected value $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ of the ERM loss[1] and show that the unique optimum of the expected ERM loss has the same direction as the Bayes optimal classifier. Next, we study the sample complexity for the ERM loss optimum to align closely enough to the Bayes optimal classifier and conclude that ERM without Mixup suffers the *curse of separability*.

Our first theorem below analyzes the optimum of expected ERM loss $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$.

Theorem 3.1. *For any $\kappa \in (0, \infty)$, the expectation of ERM loss $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ has a unique minimizer $\boldsymbol{w}^*(\kappa)$. In addition, its direction is the same as the Bayes optimal solution $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.*

Proof Sketch. We will only sketch main proof ideas and full proof can be found in Appendix A.1. We can rewrite the expected ERM loss as

$$\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})] = \mathbb{E}\left[l\left(\kappa^{-1/2}\left(\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}\right)^{1/2} Z + \boldsymbol{w}^\top \boldsymbol{\mu}\right)\right],$$

where $Z \sim N(0, 1)$. The following Lemma 3.2 implies that for a fixed value of $\boldsymbol{w}^\top \boldsymbol{\mu}$, a smaller value of $\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$ induces a smaller $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$.

Lemma 3.2. *Let $X_1 \sim N(m, \sigma_1^2), X_2 \sim N(m, \sigma_2^2)$, $m \in \mathbb{R}$ and $\sigma_1 > \sigma_2 > 0$. Then, $\mathbb{E}[l(X_1)] > \mathbb{E}[l(X_2)]$.*

Lemma 3.3 below concludes that the unique minimizer should be parallel to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

Lemma 3.3. *For any constant C, a unique solution of $\min_{\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{w}^\top \boldsymbol{\mu} = C} \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$ is a rescaling of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.*

Any remaining details can be found in Appendix A.1. □

Our Theorem 3.1 implies that the minimizer $\boldsymbol{w}^*(\kappa)$ of expected ERM loss $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ induce the Bayes optimal classifier. This may look obvious to some readers because $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ is in fact equal to $\mathbb{E}_\kappa[yl(\boldsymbol{w}^\top \boldsymbol{x}) + (1 - y)l(-\boldsymbol{w}^\top \boldsymbol{x})]$, the population loss. However, notice that the optimal classifier in terms of the population loss is dependent on the loss l, and hence is not always equal to the Bayes optimal classifier. Also, we will see theorems similar to Theorem 3.1 in later sections; in particular, Theorem 5.3 reveals how masking-based augmentation can *distort* the optimal classifier of the corresponding expected training loss; thus, characterizing such optima is of importance.

[1] The expected ERM loss is equal to the population loss, which is independent of n. However, for Mixup loss, its expected value becomes *dependent* on n, as we will see later.

Notice that the minimizer characterized in Theorem 3.1 is for *expected* ERM loss $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$. Since $\mathcal{D}_\kappa$ is not known to us and we only observe training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{D}_\kappa$, we can only hope to get close to the minimizer by optimizing the training loss $\mathcal{L}_S(\boldsymbol{w})$, and sufficiently many data samples are required to obtain a "close enough" one. Thus, a natural question arises:

How many data points are required to find a near Bayes optimal classifier using the ERM loss?

We present two theorems that answer the question above. The first one (Theorem 3.4) shows that the number of samples growing exponentially with κ^2 is *sufficient*, and the next one (Theorem 3.7) proves that this exponential growth with κ is *necessary*. In other words, as the data distribution becomes more separable, the sample complexity for getting a Bayes optimal classifier grows exponentially.

Theorem 3.4. *Let $\epsilon, \delta \in (0,1)$. Suppose the training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{D}_\kappa$ with large enough $\kappa \in (0, \infty)$, and*

$$n = \frac{\exp(\Omega(\kappa^2))}{\epsilon^4}\left(1 + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right).$$

Then, with probability at least $1-\delta$, the unique minimizer $\hat{\boldsymbol{w}}_S^*$ of $\mathcal{L}_S(\boldsymbol{w})$ exists and $\operatorname{sim}(\hat{\boldsymbol{w}}_S^*, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \geq 1-\epsilon$.

Proof Sketch. A useful tool for proving Theorem 3.4 is Lemma 3.5 inspired by the proof techniques used in Dai et al. (2000) and Shapiro & Xu (2008).

Lemma 3.5. *Let $f(\cdot,\cdot) : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}$ be a real-valued function. Define functions $F_N : \mathbb{R}^k \to \mathbb{R}$ and $\hat{F}_N : \mathbb{R}^k \to \mathbb{R}$ as*

$$F(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\boldsymbol{\theta}, \boldsymbol{\eta})],\ \hat{F}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum\nolimits_{i=1}^N f(\boldsymbol{\theta}, \boldsymbol{\eta}_i),$$

where $\mathcal{P}$ is a probability distribution on $\mathbb{R}^m$ and $\{\boldsymbol{\eta}_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{P}$. Let $\mathcal{C}$ be a nonempty compact subset of $\mathbb{R}^k$ with diameter D. Suppose the following assumptions hold:

- *The functions F and $\hat{F}_N$ have unique minimizers on $\mathcal{C}$ named $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}_N^*$, respectively.*
- *The function F is α-strongly convex on $\mathcal{C}$ ($\alpha > 0$).*
- *For any $\boldsymbol{\theta} \in \mathcal{C}$, $\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}\left[e^{|f(\boldsymbol{\theta},\boldsymbol{\eta}) - \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\boldsymbol{\theta},\boldsymbol{\eta})]|}\right] < M$.*
- *There exists a function $g(\cdot) : \mathbb{R}^k \to \mathbb{R}$ such that for any $\boldsymbol{\theta} \in \mathcal{C}$ and $\boldsymbol{\eta} \in \mathbb{R}^k$, it holds that $\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{\eta})\| \leq g(\boldsymbol{\eta})$ and $\|\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\boldsymbol{\theta}, \boldsymbol{\eta})]\| \leq \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[g(\boldsymbol{\eta})]$. In addition, $\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}\left[e^{g(\boldsymbol{\eta})}\right] < L$.*

For each $0 < \epsilon < \alpha^{-1/2}$, we have $\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}\| < \epsilon$ with probability at least $1-\delta$ if N is greater than

$$\frac{C_1 M}{\alpha^2\epsilon^4}\log\left(\frac{3}{\delta}\max\left\{1, \left(\frac{C_2 k^{1/2} D L}{\alpha\epsilon^2}\right)^k\right\}\right),$$

where $C_1, C_2 > 0$ are universal constants.

We use Lemma 3.5 by considering $\{\boldsymbol{\eta}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{P}$ as our training dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\kappa$, $\boldsymbol{\theta}$ as weight vector $\boldsymbol{w}$, and $f(\boldsymbol{\theta}, \boldsymbol{\eta}_i)$ as $(\boldsymbol{w}, (\boldsymbol{x}_i, y_i)) \mapsto y_i l(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i) l(-\boldsymbol{w}^\top \boldsymbol{x}_i)$. When we apply Lemma 3.5, several quantities are sensitive to $\|\boldsymbol{w}^*(\kappa)\|$, the ℓ_2 norm of $\boldsymbol{w}^*(\kappa)$ defined in Theorem 3.1, which we characterize in the following lemma.

Lemma 3.6. *The unique minimizer $\boldsymbol{w}^*(\kappa)$ of expected ERM loss $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ satisfies*

$$\|\boldsymbol{w}^*(\kappa)\| = \Theta(\kappa).$$

Full proof of Theorem 3.4 appears in Appendix A.4. □

Our sufficient sample complexity in Theorem 3.4 grows *exponentially* with κ. The sufficient number of data points for the minimizer $\hat{\boldsymbol{w}}_S^*$ of the ERM loss $\mathcal{L}_S(\boldsymbol{w})$ to be close to the Bayes optimal classifier becomes exponentially larger for more well-separable data distributions.

One may think that this exponential dependency may be just an artifact of our analysis and the exponential growth in κ is in fact avoidable. As an answer to this question, we introduce Theorem 3.7 indicating that the exponential growth of sample complexity is *inevitable*, dashing hopes for sub-exponential sample complexity bounds.

Theorem 3.7. *Assume $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{D}_\kappa$ with large enough $\kappa \in (0, \infty)$. If $n = \exp(\mathcal{O}(\kappa))$, then $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is linearly separable and $\operatorname{sim}(\bar{\boldsymbol{w}}_S, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) < \frac{1+\operatorname{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2}$ with probability at least 0.99, where $\bar{\boldsymbol{w}}_S$ is the ℓ_2 max margin solution:*

$$\begin{aligned} \bar{\boldsymbol{w}}_S = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad & \|\boldsymbol{w}\|^2 \\ \text{subject to} \quad & (2y_i - 1)\boldsymbol{w}^\top \boldsymbol{x}_i \geq 1. \end{aligned} \tag{2}$$

Proof Sketch. If κ is large, then $(2y-1)\boldsymbol{x}$ will be concentrated near $\boldsymbol{\mu}$ where $(\boldsymbol{x}, y) \sim \mathcal{D}_\kappa$. Hence, if n is not large enough, $(2y_i - 1)\boldsymbol{x}_i$'s will be located inside a small ball centered at $\boldsymbol{\mu}$, with high probability. Therefore, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is likely to be linearly separable. From the KKT condition of the ℓ_2 max margin problem (Equation 2), $\bar{\boldsymbol{w}}_S$ is a homogeneous combination of $(2y_i - 1)\boldsymbol{x}_i$'s. It implies $\bar{\boldsymbol{w}}_S$ is directionally close to $\boldsymbol{\mu}$, not $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. The detailed proof is in Appendix A.7. We note that the numbers $\frac{1+\operatorname{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2}$ and 0.99 in the statement are not strictly necessary; they can be replaced by any other feasible constants. □

Theorem 3.7 states that if we do not have sufficiently many data points, then the training dataset becomes linearly separable: there exists a direction $\boldsymbol{w}$ such that $(2y_i - 1)\boldsymbol{w}^\top \boldsymbol{x}_i > 0$ for all $i \in [n]$. However, this in fact means that there exist infinitely many directions $\boldsymbol{w}$ that classify the data perfectly. Then why do we care about the specific ℓ_2 max margin classifier in Theorem 3.7? Soudry et al. (2018); Ji & Telgarsky

(2019) study the implicit bias of gradient descent on a linear model with logistic loss and show that this algorithm converges in direction to the ℓ_2 max margin classifier when training data is linearly separable. In other words, if we run gradient descent on $\mathcal{L}_S(\boldsymbol{w})$, then the algorithm will return a linear classifier defined by the direction of $\bar{\boldsymbol{w}}_S$. This is why we analyze the ℓ_2 max margin solution.

The Curse of Separability. Theorem 3.7 implies that without the number of samples exponentially growing with κ, the solution found by gradient descent can be far from the Bayes optimal classifier. Combining Theorems 3.4 and 3.7, we have an interesting conclusion that even though it is easier to correctly classify training dataset when κ is larger, finding the theoretically optimal model becomes much harder. We refer to this interesting phenomenon as the *curse of separability*; without Mixup, ERM training suffers the curse of separability due to its sample complexity growing exponentially in κ.

Intuitive Explanations. What causes this phenomenon? When the data distribution is well-separable, limited training data can result in many decision boundaries having high training accuracy. However, among these, there is only one optimal decision boundary in terms of test accuracy, which is difficult to locate due to the scarcity of data points near it; this causes the curse of separability. We believe that this intuition extends beyond our simple setup; see Section 6.

4. Mixup Provably Mitigates the Curse of Separability in ERM

In this section, we study a unifying framework of Mixup-style data augmentation techniques and show that Mixup significantly alleviates the curse of separability. We will first define the unifying framework along with Mixup loss, study the location of the minimizer of expected Mixup loss, and then study the sample complexity for Mixup training to achieve near Bayes optimal classifier.

We start by defining the Mixup loss in the following framework. From this point on, we assume $n \geq 2$.

Definition 4.1 (Mixup Loss)**.** Mixup loss with training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is defined by

$$\mathcal{L}_S^{\text{mix}}(\boldsymbol{w}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{y}_{i,j} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}) + (1-\tilde{y}_{i,j}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}),$$

where

$$\begin{aligned} \lambda_{i,j} &\overset{\text{i.i.d}}{\sim} \Lambda, \\ \tilde{\boldsymbol{x}}_{i,j} &= g(\lambda_{i,j})\boldsymbol{x}_i + (1 - g(\lambda_{i,j}))\boldsymbol{x}_j, \\ \tilde{y}_{i,j} &= \lambda_{i,j} y_i + (1 - \lambda_{i,j}) y_j. \end{aligned}$$

The probability distribution Λ satisfies $\text{support}(\Lambda) \subset [0,1]$ and $\mathbb{P}_{\lambda \sim \Lambda}\left[\lambda \notin \{0,1\} \wedge g(\lambda) \neq \frac{1}{2}\right] > 0$. Also, the function $g : [0,1] \to [0,1]$ satisfies $g(z) > \frac{1}{2}$ if and only if $z > \frac{1}{2}$.

Our definition is a broad framework that covers the original Mixup by choosing Λ as the Beta distribution and $g(\cdot)$ as the identity function.

Similar to ERM loss $\mathcal{L}_S(\boldsymbol{w})$, Mixup loss $\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})$ is also a stochastic function depending on the training set S. Unlike ERM loss, the expectation of Mixup loss *depends* on n, as can be checked in Appendix B.1. As we did in Section 3, we first characterize the minimizer of expected Mixup loss $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$.

Theorem 4.2. *For each $\kappa \in (0, \infty)$ and $n \in \mathbb{N}$, the expectation of Mixup loss $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ has a unique minimizer $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$. In addition, its direction is the same as the Bayes optimal solution $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.*

Proof Sketch. We can rewrite $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ as the form

$$\mathbb{E}\left[\sum_{i=1}^{k} a_i l\left(b_i \kappa^{-1/2} (\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w})^{1/2} Z + c_i \boldsymbol{w}^\top \boldsymbol{\mu}\right)\right],$$

where $Z \sim N(0,1)$ and a_i, b_i, c_i's are real-valued random variables depending on Λ; in particular, a_i, b_i's are positive. Then, the same proof idea of Theorem 3.1 works. □

Mixup Does Not Distort Training Loss. Theorem 4.2 shows that the expected Mixup loss also has its unique minimizer pointing to the Bayes optimal direction. In other words, this theorem implies that the pair-wise mixing done in Mixup does not introduce any bias or distortion in the training loss, at least in our setting. This is one benefit that Mixup has compared to other masking-based augmentations, as we will see in Section 5.

In order to investigate the sample complexity for achieving a near Bayes optimal classifier when we train with Mixup loss, one could speculate that the same approach using Lemma 3.5 should work. However, this is not the case; analysis of the Mixup loss has to overcome a significant barrier because the mixed data points $\{(\tilde{\boldsymbol{x}}_{i,j}, \tilde{y}_{i,j})\}_{i,j=1}^n$ are no longer independent of one another. To overcome this difficulty, we prove the following lemma in Appendix B.4, which could be of independent interest:

Lemma 4.3. *Let $f(\cdot,\cdot) : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}$ be a real-valued function. Define functions $F_N : \mathbb{R}^k \to \mathbb{R}$ and $\hat{F}_N : \mathbb{R}^k \to \mathbb{R}$ as*

$$F_N(\boldsymbol{\theta}) = \frac{1}{N^2} \sum_{i,j \in [N]} f(\boldsymbol{\theta}, \boldsymbol{\eta}_{i,j}), \; F_N(\boldsymbol{\theta}) = \mathbb{E}\left[\hat{F}_N(\boldsymbol{\theta})\right]$$

where $\mathcal{P}_{i,j}$ are probability distributions on $\mathbb{R}^m$, $\boldsymbol{\eta}_{i,j} \sim \mathcal{P}_{i,j}$, and expectation is taken over all randomness. Let $\mathcal{C}$ be a nonempty compact subset of $\mathbb{R}^k$ with diameter D. Suppose the following assumptions hold:

- *For $i_1, i_2, j_1, j_2 \in [N]$, if $\{i_1\} \cup \{j_1\}$ and $\{i_2\} \cup \{j_2\}$ are disjoint, then $\boldsymbol{\eta}_{i_1,j_1}$ and $\boldsymbol{\eta}_{i_2,j_2}$ are independent.*

- *The functions F_N and $\hat{F}_N$ have unique minimizers on $\mathcal{C}$ named $\boldsymbol{\theta}_N^*$ and $\hat{\boldsymbol{\theta}}_N^*$, respectively.*
- *The function F_N is α-strongly convex on $\mathcal{C}$ ($\alpha > 0$).*
- *For any $\boldsymbol{\theta} \in \mathcal{C}$ and $i, j \in [N]$,*

$$\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}\left[e^{|f(\boldsymbol{\theta},\boldsymbol{\eta})-\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}[f(\boldsymbol{\theta},\boldsymbol{\eta})]|}\right] < M.$$

- *There exists a function $g(\cdot) : \mathbb{R}^k \to \mathbb{R}$ such that for any $\boldsymbol{\theta} \in \mathcal{C}$ and $\boldsymbol{\eta} \in \mathbb{R}^k$, it holds that $\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{\eta})\| \le g(\boldsymbol{\eta})$. In addition, $\|\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}[f(\boldsymbol{\theta}, \boldsymbol{\eta})]\| \le \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}[g(\boldsymbol{\eta})]$ and $\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}\left[e^{g(\boldsymbol{\eta})}\right] < L$ for each $\boldsymbol{\theta} \in \mathcal{C}$ and $i, j \in [N]$.*

For each $0 < \epsilon < \alpha^{-1/2}$, we have $\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}\| < \epsilon$ with probability at least $1 - \delta$ if N is greater than

$$\frac{C_1' M}{\alpha^2 \epsilon^4} \log\left(\frac{3}{\delta} \max\left\{1, \left(\frac{C_2' k^{1/2} D L}{\alpha \epsilon^2}\right)^k\right\}\right),$$

where $C_1', C_2' > 0$ are universal constants.

Proof Sketch. When we follow the proof of Lemma 3.5, the challenging part is that we cannot use the fact that the expected value of a product of independent random variables is equal to a product of expectations of individual random variables. We overcome this by partitioning the N^2 random variables into batches such that random variables belonging to the same batch are independent (Lemma D.9) and then applying a generalized Cauchy-Schwartz inequality (Lemma D.10) to bound an expectation of a product of dependent random variables (each corresponding to a batch) with a product of expectations of the random variables. A formal proof can be found in Appendix B.4. □

Similar to the proof of Theorem 3.4, considering $\{\boldsymbol{\eta}_{i,j}\}_{i,j=1}^n$ as the "mixed" dataset $\{(\tilde{\boldsymbol{x}}_{i,j}.\tilde{y}_{i,j})\}_{i.j=1}^n$, $\boldsymbol{\theta}$ as weight vector $\boldsymbol{w}$, and $f(\boldsymbol{\theta}, \boldsymbol{\eta}_{i,j})$ as $(\boldsymbol{w}, (\tilde{\boldsymbol{x}}_{i,j}, \tilde{y}_{i,j})) \mapsto \tilde{y}_i l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j} + (1 - \tilde{y}_i) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j})$ induces the following theorem.

Theorem 4.4. *Let $\epsilon, \delta \in (0, 1)$. Suppose the training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{D}_\kappa$ with large enough $\kappa \in (0, \infty)$ and*

$$n = \frac{\Omega(\kappa^2)}{\epsilon^4}\left(1 + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right).$$

Then, with probability at least $1 - \delta$, the unique minimizer $\hat{\boldsymbol{w}}_{\text{mix},S}^*$ of $\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})$ exists and $\text{sim}(\hat{\boldsymbol{w}}_{\text{mix},S}^*, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \ge 1 - \epsilon$.

Theorem 4.4 indicates that the sample complexity for finding a near Bayes optimal classifier with Mixup training grows only quadratically in κ. Compared to the necessity of exponential growth demonstrated in Theorem 3.7, Theorem 4.4 shows that there is a provable exponential gap between ERM training and Mixup training.

Intuition on the Smaller Sample Complexity of Mixup. We would like to provide some intuition on our result before we introduce technical aspects. Unlike ERM training, Mixup training uses mixed training points and these can be located near the optimal decision boundary when we mix two data points having distinct labels. This closeness of mixed points to the Bayes optimal decision boundary makes it easier to correctly locate the boundary.

Different Scaling of $\boldsymbol{w}^*(\kappa)$ and $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$. On the technical front, the difference in sample complexity from Theorem 3.4 stems from the difference between the norm of expected loss minimizers $\boldsymbol{w}^*(\kappa)$ and $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$, which determine several meaningful terms when we apply Lemma 4.3. The following lemma characterizes the ℓ_2 norm of $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$, defined in Theorem 4.2:

Lemma 4.5. *The unique minimizer $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$ of expected Mixup loss $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ satisfies*

$$\|\boldsymbol{w}_{\text{mix}}^*(n, \kappa)\| = \Theta(1).^2$$

Comparison with Lemma 3.6 reveals that the minimizer $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$ of expected Mixup loss is much closer to zero compared to that of the expected ERM loss. For ERM loss, large scaling of weight leads to smaller loss for correctly classified data points. Also, for larger κ, larger portion of population data will be correctly classified by $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Hence, $\|\boldsymbol{w}^*(\kappa)\|$ increases as κ increases. However, in case of Mixup, mixed labels prevent $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$ from growing with κ. To illustrate why, consider the case $\tilde{y}_{i,j} = 0.5$, which leads to a Mixup loss

$$0.5 l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}) + 0.5 l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}).$$

Notice here that the loss becomes large whenever $\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}$ is large in magnitude, no matter the sign is. For this reason, $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$ should not increase with κ and this leads to the smaller sample complexity of Mixup.

In this section, we showed that Mixup training does not distort the training loss (Theorem 4.2) and also that Mixup provides a great remedy to the curse of separability phenomenon (Theorem 4.4), because the sample complexity only grows in κ^2 while ERM suffers at least exponential growth in κ. Thus, Mixup provably mitigates the curse of separability and helps us find a model with the best generalization performance.

5. Masking-based Mixup Can Distort Training

Recent Mixup variants for image data (Yun et al., 2019; Kim et al., 2020; 2021; Liu et al., 2022) use masking on input data. In this section, we investigate how masking-based augmentation techniques work in our data distribution setting. We consider the class of masking-based Mixup variants formulated as follows.

[2] We stress that the upper/lower bounds are independent of n.

Definition 5.1. (Masking-based Mixup Loss) Masking-based Mixup loss with training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is defined as

$$\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}) := \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{y}_{i,j}^{\text{mask}} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) + (1 - \tilde{y}_{i,j}^{\text{mask}}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}),$$

where

$$(\boldsymbol{M}_{i,j}, \lambda_{i,j}) \overset{\text{i.i.d}}{\sim} \mathcal{M},$$
$$\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}} = \boldsymbol{M}_{i,j} \odot \boldsymbol{x}_i + (\boldsymbol{1} - \boldsymbol{M}_{i,j}) \odot \boldsymbol{x}_j,$$
$$\tilde{y}_{i,j}^{\text{mask}} = \lambda_{i,j} y_i + (1 - \lambda_{i,j}) y_j.$$

Here, $\text{support}(\mathcal{M}) \subset \{0,1\}^d \times [0,1]$.
In our definition of masking-based mixup loss, we formulate the masking operation on data points by element-wise multiplication with vectors having entries from only 0 and 1. This formulation includes CutMix (Yun et al., 2019) which is simplest type of masking-based Mixup. State-of-the-art Mixup variants having more complex masking strategies (Kim et al., 2020; 2021; Liu et al., 2022) are out of the scope of this paper. We also introduce the following assumption on masking.

Assumption 5.2. The set $\text{support}(\boldsymbol{\mu} \odot (2\boldsymbol{M} - \boldsymbol{1}))$ spans $\mathbb{R}^d$ and $\mathbb{P}[\lambda \boldsymbol{1}_{\boldsymbol{M}=\boldsymbol{M}_0} \notin \{0,1\}] > 0$ for each $\boldsymbol{M}_0 \in \text{support}(\boldsymbol{M})$ where $(\boldsymbol{M}, \lambda) \sim \mathcal{M}$.

Before we move on to our main results, we demonstrate why our formulation and assumption hold for CutMix. CutMix samples mixing ratio $\lambda_{i,j}$ from beta distribution and masking vector $\boldsymbol{M}_{i,j}$ is uniformly sampled from vectors in which the number of 1's is proportional to $\lambda_{i,j}$. Since the support of beta distribution is $[0,1]$, support of $\boldsymbol{M}_{i,j}$ contains the standard basis of $\mathbb{R}^d$. Hence if all the components of $\boldsymbol{\mu}$ are nonzero, Assumption 5.2 holds.

Recall that we defined $\mathcal{D}_\infty$ as a limit behavior of $\mathcal{D}_\kappa$ as $\kappa \to \infty$ and it is independent of $\boldsymbol{\Sigma}$. Hence, $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ is independent of $\boldsymbol{\Sigma}$. The following theorem investigates the minimizer of the expected masking-based Mixup loss $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$, focusing on large enough κ.

Theorem 5.3. *Suppose Assumption 5.2 holds and the training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{D}_\kappa$ with $\kappa \in (0, \infty]$. Then, the expected loss $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ has a unique minimizer $\boldsymbol{w}_{\text{mask}}^*(n, \kappa)$. In addition, $\lim_{\kappa \to \infty} \boldsymbol{w}_{\text{mask}}^*(n, \kappa) = \boldsymbol{w}_{\text{mask}}^*(n, \infty)$.*

Proof Sketch. For the uniqueness of the minimizer, we use almost the same strategy as the uniqueness parts of Theorem 3.1 and Theorem 4.2. The only part that requires a different strategy is the uniqueness for $\kappa = \infty$; in this case, we exploit Assumption 5.2. Also, from Assumption 5.2, we can show that $\mathbb{E}_\infty\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\right]$ is α-strongly convex with some $\alpha > 0$. Using strong convexity constant α, we establish upper bound on $\|\boldsymbol{w}_{\text{mask}}^*(n, \kappa) - \boldsymbol{w}_{\text{mask}}^*(n, \infty)\|$ represented by $\mathbb{E}_\kappa\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\right] - \mathbb{E}_\infty\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\right]$ for several values of $\boldsymbol{w}$ contained in a bounded set. We finish up by showing $\mathbb{E}_\kappa\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\right] - \mathbb{E}_\infty\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\right] \to 0$ uniformly on the bounded set as $\kappa \to \infty$. □

Masking-based Mixup Can Distort Training Loss. Unlike ERM loss and Mixup loss, characterizing the exact direction of $\boldsymbol{w}_{\text{mask}}^*(n, \kappa)$ is challenging since $\mathbb{E}_\kappa\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\right]$ has a more complicated form because of masking. However, our Theorem 5.3 implies that $\boldsymbol{w}_{\text{mask}}^*(n, \kappa)$ leads to a solution only depending on $\boldsymbol{\mu}$ and deviates from the Bayes optimal solution for sufficiently large κ. Even though Theorem 5.3 guarantees deviation of $\boldsymbol{w}_{\text{mask}}^*(n, \kappa)$ from the Bayes optimal direction only for large κ, our experimental results in Section 6.1 suggest that our result holds even for moderately sized κ.

One might be wondering whether the same thing can be said for the minimizer of the expected Mixup loss $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$; we illustrate why the proof idea of Theorem 5.3 does not work for Mixup. While $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ is strongly convex and has a unique minimizer, minimizers of $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ are not unique since if $\boldsymbol{u} \in \mathbb{R}^d$ is a minimizer of $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$, $\boldsymbol{u} + \boldsymbol{v}$ is also a minimizer of $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ for any $\boldsymbol{v} \in \mathbb{R}^d$ orthogonal to $\boldsymbol{\mu}$. Therefore, $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$ maintains its direction even though $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ is independent of $\boldsymbol{\Sigma}$; since there are many minimizers in $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$, $\boldsymbol{w}_{\text{mix}}^*(n, \kappa)$ converges to the Bayes optimal minimizer (among the many) as $\kappa \to \infty$.

While masking-based Mixup training does not necessarily lead to Bayes optimal classifiers, the sample complexity proof still works. In fact, we can find the solution near the minimizer $\boldsymbol{w}_{\text{mask}}^*(n, \kappa)$ of the expected loss $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ with fewer samples.

Theorem 5.4. *Let $\epsilon, \delta \in (0,1)$. Suppose Assumption 5.2 holds and the training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{D}_\kappa$ with large enough $\kappa \in (0, \infty)$. If*

$$n = \frac{\Omega(1)}{\epsilon^4} \left(1 + \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right),$$

then with probability at least $1 - \delta$, the unique minimizer $\hat{\boldsymbol{w}}_{\text{mask},S}^*$ of $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})$ exists and $\text{sim}(\hat{\boldsymbol{w}}_{\text{mask},S}^*, \boldsymbol{w}_{\text{mask}}^*(n, \kappa)) \geq 1 - \epsilon$.

Theorem 5.4 indicates masking-based Mixup also mitigates the curse of separability (even better than Mixup).

Why Even Smaller Sample Complexity? We consider a simple case $\boldsymbol{\mu} = [1, 1]^\top$, which is sufficient to convey our intuition. When κ is large, most of the data points are likely to be concentrated around $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$. Since Mixup uses linearly interpolated data points, all the mixed training data will be close to a line that passes through the origin and has direction $\boldsymbol{\mu}$. It means that all raw data points or mixed points

are almost orthogonal to $[1, -1]^\top$. Consequently, both ERM and Mixup training loss will be less sensitive to perturbations orthogonal to $\boldsymbol{\mu}$: i.e., $\mathcal{L}_S(\boldsymbol{w}) \approx \mathcal{L}_S(\boldsymbol{w} + t[1, -1]^\top)$ and $\mathcal{L}_S^{\text{mix}}(\boldsymbol{w}) \approx \mathcal{L}_S^{\text{mix}}(\boldsymbol{w} + t[1, -1]^\top)$ for any small t, which makes it difficult to locate the exact minimizer of the objective loss. In contrast, masking-based Mixup uses cut-and-pasted data points such as $\boldsymbol{z}_1 \approx [1, -1]^\top$, constructed by pasting the first coordinate of positive data and the second coordinate of negative data. Masking-based Mixup also uses $\boldsymbol{z}_2 \approx [1, 1]^\top$, obtained from mixing two positive data points. We observe that $\boldsymbol{z}_1, \boldsymbol{z}_2$ span the whole space $\mathbb{R}^2$. Therefore, for any perturbation $\boldsymbol{v} \in \mathbb{R}^2$, $\boldsymbol{v}^\top \boldsymbol{z}_1 \not\approx 0$ and/or $\boldsymbol{v}^\top \boldsymbol{z}_2 \not\approx 0$ should hold, which implies that $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}) \not\approx \mathcal{L}_S^{\text{mask}}(\boldsymbol{w}+\boldsymbol{v})$ for any perturbation $\boldsymbol{v}$. In other words, the masking-based Mixup loss is sensitive to perturbations in any direction. This makes it easier to locate the exact minimizer of the objective loss, even when κ is large.

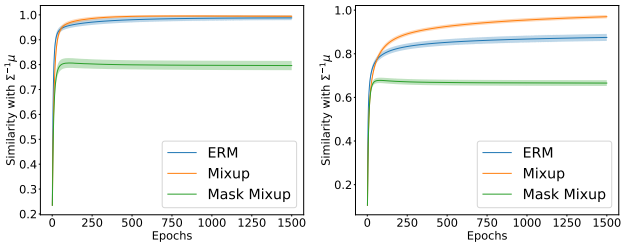
In this section, we showed that masking-based Mixup mitigates the curse of separability even better than Mixup, but unfortunately, making-based Mixup can find a classifier that is far from being Bayes optimal due to Theorem 5.3. One may think that these results are contradictory to the empirical success of masking-based Mixup such as CutMix (Yun et al., 2019) on image data. However, the regularization effect of Mixup variants is highly dependent on the data, as also noted by Park et al. (2022). Therefore, our conclusion in Section 5 does not necessarily contradict the success of masking-based Mixup on practical image data. We speculate that the distortion effect of masking-based Mixup on complex image data may be small or even beneficial (e.g., by increasing the chance of co-occurrence of some useful features). In this case, the small sample complexity of masking-based Mixup would be helpful. However, a rigorous theoretical analysis is beyond the scope of our current work and is an essential direction for future research.

6. Experiments

In this section, we present several experimental results to support our findings.

6.1. Experiments on Our Setting $\mathcal{D}_\kappa$

First, we provide empirical results on our setting. We properly choose $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with $\|\boldsymbol{\mu}\|^2 = \|\boldsymbol{\Sigma}\|$ so that $\boldsymbol{\mu}$ and the Bayes optimal direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ have different directions, i.e., $\boldsymbol{\mu}$ is not an eigenevector of $\boldsymbol{\Sigma}$. We provide exact values of our choice of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in Appendix E.1. We compare three training methods ERM, Mixup, and Mask Mixup. Mask Mixup we considered is a kind of masking-based Mixup with $\mathcal{M}$ defined by the following: we say $(\boldsymbol{M}, \lambda) \sim \mathcal{M}$ when λ is drawn from beta distribution $\text{Beta}(\alpha, \alpha)$ and each component of $\boldsymbol{M}$ follows $\text{Bernoulli}(\lambda)$.



(a) $\kappa = 0.5$ (b) $\kappa = 2.0$

Figure 1. Cosine similarity between learned weight and the Bayes optimal direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. ERM successfully finds the Bayes optimal classifier when κ is small ($\kappa = 0.5$) and fails when κ is larger ($\kappa = 2.0$) because of the curse of separability. Meanwhile, Mixup succeeds in both cases since Mixup mitigates the curse of separability. Mask Mixup fails in both cases due to distortion.

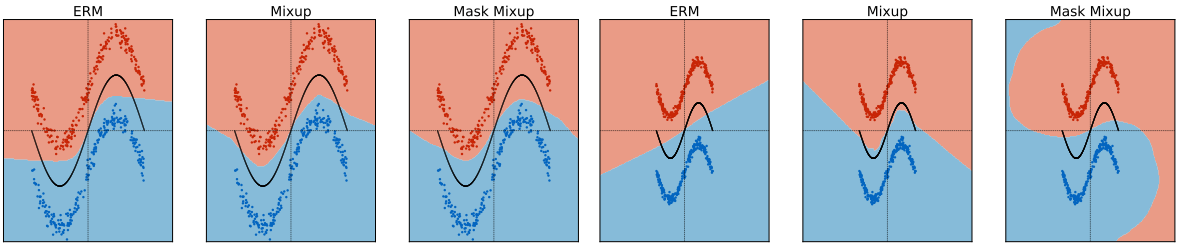
We compare two different values of κ; $\kappa = 0.5$ and $\kappa = 2.0$. We train for 1500 epochs using randomly sampled 500 training samples from each $\mathcal{D}_\kappa$ and full gradient descent with learning rate 1 and we choose $\alpha = 1$ for the hyperparameter of Mixup and Mask Mixup. We run 500 times with fixed initial weight but different samples of training sets and we plot cosine similarity between the trained weight and the Bayes optimal direction during training in Figure 1.

For the case $\kappa = 0.5$, ERM and Mixup lead to the Bayes optimal classifier. However, for the case $\kappa = 2.0$, ERM finds a solution deviating from the Bayes optimal solution, while Mixup still finds almost accurate solutions. This result is predicted by our theoretical findings; ERM suffers the curse of separability and Mixup mitigates it. Also, we can check the minimizers of the Mask Mixup loss deviate significantly from the Bayes optimal direction for both cases $\kappa = 0.5$ and $\kappa = 2.0$, even though our theoretical result (Theorem 5.3) focus on large κ. We provide additional results on more various values of κ, the number of samples n, and the dimension of data d in Appendix E.2.

6.2. 2D Classification on Synthetic Data

We also provide empirical results supporting that the intuitions gained from our analysis extend beyond our settings. We consider training a two-layer ReLU network with 500 hidden nodes on 2D synthetic data with binary labels having sine-shaped noise[3] from its mean for each class. As a result of the noise, the optimal decision boundary is also sine-shaped. We consider two settings with different magnitudes of noise while keeping the means the same. Using three methods ERM, Mixup, and Mask Mixup (which we introduced in the previous subsection), we train for 1500 epochs using 500 samples of data points and Adam (Kingma & Ba, 2014) with full batch, learning rate 0.001 and using default hyperparameters of $\beta_1 = 0.9, \beta_2 = 0.999$. We also use $\alpha = 1$ for the hyperparameter of Mixup and Mask Mixup.

[3]The noise consists of uniformly sampled x-coordinate and y-coordinate having sine value with additional Gaussian noise.



(a) Large noise, less separable setting (b) Small noise, well separable setting

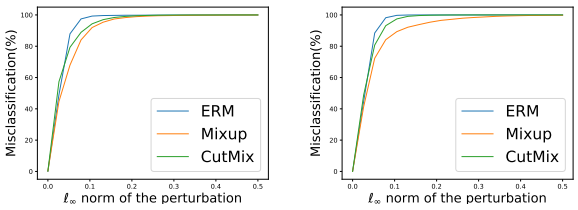
Figure 2. Boundary decision of trained models with ERM loss, Mixup loss and Mask Mixup loss

Figure 2(a) plots the decision boundaries (red vs. blue) of trained models in the setting with larger noise, which corresponds to a less separable setting. We also draw the Bayes optimal boundaries with black solid lines. All ERM, Mixup, and Mask Mixup find decision boundaries that reflect the sine-shaped optimal decision boundary. Figure 2(b) shows the results with smaller noise, i.e., a more separable setting. The decision boundary of ERM degenerates to a linear boundary, ignoring the sine-shaped noise. However, even though Mixup slightly distorts training,[4] Mixup finds a nonlinear boundary that captures sine shape even when data is highly separated. This result is consistent with our findings that *Mixup mitigates the curse of separability*, even outside our simple settings. Also, the decision boundary of models trained by using Mask Mixup is nonlinear, which may come from smaller sample complexity, but it seems to suffer more distortion compared to Mixup.

6.3. Classification on CIFAR-10

We also conduct experiments on the real-world data CIFAR-10 (Krizhevsky et al., 2009). To compare three methods ERM, Mixup, and CutMix, we train VGG19 (Simonyan & Zisserman, 2014) and ResNet18 (He et al., 2016) for 300 epochs on the training set with batch size 256 using SGD with weigh decay 10^{-4} and we choose $\alpha = 1$ for the hyperparameter of Mixup and CutMix. Also, we use a learning rate 0.1 at the beginning and divide it by 10 after 100 and 150 epochs. Unlike linear models and 2D classification tasks, the decision boundaries of deep neural networks trained with complex data are intractable. Hence, following the method considered in Nar et al. (2019); Pezeshki et al. (2021), we use the norm of input perturbation[5] required to cross the decision boundary to investigate the complex decision boundary.

Figure 3 indicates that Mixup tends to find decision boundaries farther from overall data points than decision boundaries obtained by ERM. This is consistent with our intuition on the curse of separability and how Mixup mitigates it. In addition, the plots of CutMix are placed between the plots of ERM and Mixup. As also observed in Figure 2(b), we believe that the combination of distortion and smaller sample complexity results in such a trend.



(a) VGG19 (b) ResNet18

Figure 3. Estimation of decision boundaries of (a) VGG19 and (b) ResNet18 trained with CIFAR-10

7. Conclusion

We analyzed how Mixup-style training influences the sample complexity for getting optimal decision boundaries in logistic regression on data drawn from two symmetric Gaussian distributions with separability constant κ. Interestingly, we proved that vanilla training suffers the curse of separability. More precisely, ERM requires an exponentially increasing number of data with κ for finding a near Bayes optimal classifier. We proved that Mixup mitigates the curse of separability and the sample complexity for finding optimal classifier grows only quadratically with κ. We further investigated masking-based Mixup methods and showed that they can cause training loss distortion and find a suboptimal decision boundary while having a small sample complexity. One interesting future direction is analyzing when and how state-of-the-art masking-based Mixup works, by considering more sophisticated data distributions capturing image data and complicated models such as neural networks.

Acknowledgements

This paper was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)) funded by the Korea government (MSIT), two National Research Foundation of Korea (NRF) grants (No. NRF-2019R1A5A1028324, RS-2023-00211352) funded by the Korea government (MSIT), and a grant funded by Samsung Electronics Co., Ltd.

[4] One may check that the optimal decision boundary becomes similar to a sine-shaped curve with a slightly smaller amplitude.

[5] We apply the projected gradient descent attack implemented by Rauber et al. (2017) to compute the perturbation on input.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Carratino, L., Cissé, M., Jenatton, R., and Vert, J.-P. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

Chidambaram, M., Wang, X., Hu, Y., Wu, C., and Ge, R. Towards understanding the data dependency of mixup-style training. *arXiv preprint arXiv:2110.07647*, 2021.

Chidambaram, M., Wang, X., Wu, C., and Ge, R. Provably learning diverse features in multi-view data with midpoint mixup. *arXiv preprint arXiv:2210.13512*, 2022.

Dai, L., Chen, C.-H., and Birge, J. R. Convergence properties of two-stage stochastic programming. *Journal of Optimization Theory and Applications*, 106(3):489–509, 2000.

Guo, H. Midpoint regularization: from high uncertainty training to conservative classification. *arXiv preprint arXiv:2106.13913*, 2021.

Guo, H., Mao, Y., and Zhang, R. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*, 2019.

Hanson, D. L. and Wright, F. T. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ji, Z. and Telgarsky, M. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798. PMLR, 2019.

Kalantidis, Y., Sariyildiz, M. B., Pion, N., Weinzaepfel, P., and Larlus, D. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

Kim, J.-H., Choo, W., and Song, H. O. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pp. 5275–5285. PMLR, 2020.

Kim, J.-H., Choo, W., Jeong, H., and Song, H. O. Co-mixup: Saliency guided joint mixup with supermodular diversity. *arXiv preprint arXiv:2102.03065*, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Liu, Z., Li, S., Wu, D., Liu, Z., Chen, Z., Wu, L., and Li, S. Z. Automix: Unveiling the power of mixup for stronger classifiers. In *European Conference on Computer Vision*, pp. 441–458. Springer, 2022.

Lugosi, G. and Mendelson, S. Sub-gaussian estimators of the mean of a random vector. *The annals of statistics*, 47 (2):783–794, 2019.

Nar, K., Ocal, O., Sastry, S. S., and Ramchandran, K. Cross-entropy loss and low-rank features have responsibility for adversarial examples. *arXiv preprint arXiv:1901.08360*, 2019.

Park, C., Yun, S., and Chun, S. A unified analysis of mixed sample data augmentation: A loss function perspective. In *Advances in Neural Information Processing Systems*, 2022.

Pezeshki, M., Kaba, O., Bengio, Y., Courville, A. C., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Rauber, J., Brendel, W., and Bethge, M. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

Shapiro, A. and Xu, H. Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. *Optimization*, 57(3):395–418, 2008.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Sun, L., Xia, C., Yin, W., Liang, T., Yu, P. S., and He, L. Mixup-transformer: dynamic data augmentation for nlp tasks. *arXiv preprint arXiv:2010.02394*, 2020.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. PMLR, 2019.

Verma, V., Luong, T., Kawaguchi, K., Pham, H., and Le, Q. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, pp. 10530–10541. PMLR, 2021.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020.

Zhang, L., Deng, Z., Kawaguchi, K., and Zou, J. When and how mixup improves calibration. In *International Conference on Machine Learning*, pp. 26135–26160. PMLR, 2022.

A. Proofs for Section 3

A.1. Proof of Theorem 3.1

We prove the existence and uniqueness of a minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ first. Next, we characterize a direction of a unique minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$.

Step 1: Existence and Uniqueness of a Minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$

From convexity of $l(\cdot)$, for any $z \in \mathbb{R}$, $l(z) \geq -\frac{1}{2}z + \log 2 \geq -\frac{1}{2}z$ since $z \mapsto -\frac{1}{2}z + \log 2$ is tangent line of the graph of $l(z)$ at $z = 0$. Thus, we have

$$\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})] = \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[l(\boldsymbol{w}^\top\mathbf{x})] \geq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[l(\boldsymbol{w}^\top\mathbf{x})\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}] \geq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[-\frac{1}{2}\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}\right].$$

Claim A.1. *For any $\kappa \in (0,\infty)$, a mapping $\boldsymbol{w} \mapsto \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}]$ is continuous.*

Proof of Claim A.1. For any $\epsilon > 0$, let $\delta_\epsilon := \left(\kappa^{-1}\|\boldsymbol{\Sigma}\| + \|\boldsymbol{\mu}\|^2\right)^{-1/2}\cdot\epsilon/2 > 0$. Then, for any $\boldsymbol{\Delta} \in \mathbb{R}^d$ with $\|\boldsymbol{\Delta}\| \leq \delta_\epsilon$ and $\boldsymbol{w} \in \mathbb{R}^d$, we are going to show that

$$\left|\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}\cdot\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0}] - \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}]\right| \leq \epsilon,$$

to conclude that the mapping $\boldsymbol{w} \mapsto \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}]$ is continuous.

To this end, we start by

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}\cdot\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0}] - \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}]\right| \\
&\leq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}\cdot\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0} - \boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}\right|\right] \\
&= \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|\boldsymbol{\Delta}^\top\mathbf{x}\cdot\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0} + \boldsymbol{w}^\top\mathbf{x}\cdot(\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0} - \mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0})\right|\right] \\
&\leq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|\boldsymbol{\Delta}^\top\mathbf{x}\cdot\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0}\right|\right] + \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|\boldsymbol{w}^\top\mathbf{x}\cdot(\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0} - \mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0})\right|\right].
\end{aligned}$$

It is clear that $\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|\boldsymbol{\Delta}^\top\mathbf{x}\cdot\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0}\right|\right] \leq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|\boldsymbol{\Delta}^\top\mathbf{x}\right|\right]$. Also, for each $\boldsymbol{x} \in \mathbb{R}^d$,

$$\boldsymbol{w}^\top\boldsymbol{x}\cdot(\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\boldsymbol{x}<0} - \mathbf{1}_{\boldsymbol{w}^\top\boldsymbol{x}<0}) = \begin{cases} \boldsymbol{w}^\top\boldsymbol{x} \leq -\boldsymbol{\Delta}^\top\boldsymbol{x} = |\boldsymbol{\Delta}^\top\boldsymbol{x}| & \text{if}(\boldsymbol{w}+\boldsymbol{\Delta})^\top\boldsymbol{x} < 0, \boldsymbol{w}^\top\boldsymbol{x} \geq 0, \\ -\boldsymbol{w}^\top\boldsymbol{x} \leq \boldsymbol{\Delta}^\top\boldsymbol{x} = |\boldsymbol{\Delta}^\top\boldsymbol{x}| & \text{if}(\boldsymbol{w}+\boldsymbol{\Delta})^\top\boldsymbol{x} \geq 0, \boldsymbol{w}^\top\boldsymbol{x} < 0, \\ 0 \leq |\boldsymbol{\Delta}^\top\boldsymbol{x}| & \text{otherwise.} \end{cases}$$

Therefore, $\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|\boldsymbol{w}^\top\mathbf{x}\cdot(\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0} - \mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0})\right|\right] \leq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|\boldsymbol{\Delta}^\top\mathbf{x}\right|\right]$. Also, by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\left|\boldsymbol{\Delta}^\top\mathbf{x}\right|\right] &\leq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[(\boldsymbol{\Delta}^\top\mathbf{x})^2\right]^{1/2} \\
&= \left(\kappa^{-1}\boldsymbol{\Delta}^\top\boldsymbol{\Sigma}\boldsymbol{\Delta} + (\boldsymbol{\Delta}^\top\boldsymbol{\mu})^2\right)^{1/2} \\
&\leq \left(\kappa^{-1}\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Delta}\|^2 + \|\boldsymbol{\mu}\|^2\|\boldsymbol{\Delta}\|^2\right)^{1/2} \\
&\leq \left(\kappa^{-1}\|\boldsymbol{\Sigma}\| + \|\boldsymbol{\mu}\|^2\right)^{1/2}\cdot\delta_\epsilon \\
&= \epsilon/2.
\end{aligned}$$

Hence, we have $\left|\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}\cdot\mathbf{1}_{(\boldsymbol{w}+\boldsymbol{\Delta})^\top\mathbf{x}<0}] - \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}]\right| \leq \epsilon$, as desired. □

From Claim A.1 and compactness of the unit sphere $\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| = 1\} \subset \mathbb{R}^d$, it follows that for any given $\kappa \in (0,\infty)$, a mapping $\boldsymbol{w} \mapsto \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}]$ has the maximum value (over the unit sphere) $-m_\kappa$ with $m_\kappa > 0$. For any $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}\| > 2m_\kappa^{-1}$, we have

$$\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})] \geq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[-\frac{1}{2}\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}\right]$$

$$
\begin{aligned}
&= \|\boldsymbol{w}\| \cdot \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[-\frac{1}{2}\cdot\frac{\boldsymbol{w}^\top\mathbf{x}}{\|\boldsymbol{w}\|}\cdot\mathbf{1}_{\frac{\boldsymbol{w}^\top\mathbf{x}}{\|\boldsymbol{w}\|}<0}\right] \\
&\geq \frac{m_\kappa\|\boldsymbol{w}\|}{2} > 1 > \log 2 \\
&= \mathbb{E}_\kappa[\mathcal{L}_S(\mathbf{0})].
\end{aligned} \tag{3}
$$

Therefore, a minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ has to be necessarily contained in a compact set $\{\boldsymbol{w}\in\mathbb{R}^d : \|\boldsymbol{w}\|\leq 2m_\kappa^{-1}\}$. Since $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ is a continuous function of $\boldsymbol{w}$, there must exist a minimizer. The existence part is hence proved.

To show uniqueness, we will prove strict convexity of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$. From strict convexity of $l(\cdot)$, for any $t\in[0,1]$, $\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{R}^d$ with $\boldsymbol{w}_1\neq\boldsymbol{w}_2$ and $y\in\{0,1\}$, we have

$$
\begin{aligned}
&t[yl(\boldsymbol{w}_1^\top\boldsymbol{x})+(1-y)l(-\boldsymbol{w}_1^\top\boldsymbol{x})]+(1-t)[yl(\boldsymbol{w}_2^\top\boldsymbol{x})+(1-y)l(-\boldsymbol{w}_2^\top\boldsymbol{x})] \\
&> yl((t\boldsymbol{w}_1+(1-t)\boldsymbol{w}_2)^\top\boldsymbol{x})+(1-y)l(-(t\boldsymbol{w}_1+(1-t)\boldsymbol{w}_2)^\top\boldsymbol{x}),
\end{aligned}
$$

and any $\boldsymbol{x}\in\mathbb{R}^d$ except for a Lebesgue measure zero set (i.e., the set of points $\boldsymbol{x}\in\mathbb{R}^d$ satisfying $\boldsymbol{w}_1^\top\boldsymbol{x}=\boldsymbol{w}_2^\top\boldsymbol{x}$).

By taking expectation, we have for any $t\in[0,1]$, and $\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{R}^d$,

$$
t\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w}_1)]+(1-t)\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w}_2)] > \mathbb{E}_\kappa[\mathcal{L}_S(t\boldsymbol{w}_1+(1-t)\boldsymbol{w}_2)].
$$

Therefore, $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ is strictly convex. Since a strictly convex function has at most one minimizer, we conclude that $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ has a unique minimizer $\boldsymbol{w}^*(\kappa)$ for any given $\kappa\in(0,\infty)$. □

Step 2: Direction of a Unique Minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$

We rewrite $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ as

$$
\begin{aligned}
\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})] &= \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[l\left(\boldsymbol{w}^\top\mathbf{x}\right)\right] \\
&= \mathbb{E}_{X\sim N(\boldsymbol{w}^\top\boldsymbol{\mu},\kappa^{-1}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w})}\left[l(X)\right] \\
&= \mathbb{E}_{Z\sim N(0,1)}\left[l\left(\kappa^{-1/2}\left(\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right)^{1/2}Z+\boldsymbol{w}^\top\boldsymbol{\mu}\right)\right].
\end{aligned} \tag{4}
$$

We recall two lemmas we described in our proof sketch.

Lemma 3.2. *Let $X_1\sim N(m,\sigma_1^2), X_2\sim N(m,\sigma_2^2)$, $m\in\mathbb{R}$ and $\sigma_1>\sigma_2>0$. Then, $\mathbb{E}[l(X_1)]>\mathbb{E}[l(X_2)]$.*

Lemma 3.3. *For any constant C, a unique solution of $\min_{\boldsymbol{w}\in\mathbb{R}^d,\boldsymbol{w}^\top\boldsymbol{\mu}=C}\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}$ is a rescaling of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.*

Let $C_\kappa := \boldsymbol{w}^*(\kappa)^\top\boldsymbol{\mu}$. By Equation 4 and Lemma 3.2, $\boldsymbol{w}^*(\kappa)$ is a solution for the problem $\min_{\boldsymbol{w}^\top\boldsymbol{\mu}=C_\kappa}\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}$. Hence, Lemma 3.3 implies that there exists $c_\kappa\in\mathbb{R}$ such that $\boldsymbol{w}^*(\kappa):=c_\kappa\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. The only remaining part is showing $c_\kappa>0$. If $c_\kappa<0$, we have

$$
\begin{aligned}
\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w}^*(\kappa))] &= \mathbb{E}_{Z\sim N(0,1)}\left[l\left(\kappa^{-1/2}(\boldsymbol{w}^*(\kappa)^\top\boldsymbol{\Sigma}\boldsymbol{w}^*(\kappa))^{1/2}Z+c_\kappa\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right] \\
&> \mathbb{E}_{Z\sim N(0,1)}\left[l\left(\kappa^{-1/2}(-\boldsymbol{w}^*(\kappa))^\top\boldsymbol{\Sigma}(-\boldsymbol{w}^*(\kappa)))^{1/2}Z-c_\kappa\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right)\right] \\
&= \mathbb{E}\left[\mathcal{L}_S\left(-\boldsymbol{w}^*(\kappa)\right)\right],
\end{aligned}
$$

where the inequality holds because $l(\cdot)$ is strictly decreasing and $c_\kappa\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}<0$. It is contradictory to $\boldsymbol{w}^*(\kappa)$ being a unique minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$, so we conclude $c_\kappa\geq 0$. Showing c_κ is strictly positive will be handled in the proof of Lemma 3.6 which can be found in Appendix A.6. □

A.2. Proof of Lemma 3.2

Define a function $f:(0,\infty)\to\mathbb{R}$ as $f(\sigma):=\mathbb{E}_{Z\sim N(0,1)}[l(m+\sigma Z)]$. It suffices to show that f is strictly increasing. For each $\sigma\in(0,\infty)$ and $z\in\mathbb{R}$, $\left|\frac{\partial}{\partial\sigma}l(m+\sigma z)\right| = |l'(m+\sigma z)z| = \left|\frac{z}{1+e^{m+\sigma z}}\right| \leq |z|$ and $\mathbb{E}_{Z\sim N(0,1)}[|Z|]<\infty$. Thus, by Lemma D.1,

$$
\frac{d}{d\sigma}f(\sigma) = \mathbb{E}_{Z\sim N(0,1)}\left[\frac{\partial}{\partial\sigma}l(m+\sigma Z)\right] = \mathbb{E}_{Z\sim N(0,1)}[l'(m+\sigma Z)Z]
$$

$$
\begin{aligned}
&= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{z}{1+e^{m+\sigma z}} e^{-z^2/2} dz \\
&= -\frac{1}{\sqrt{2\pi}} \left(\int_{0}^{\infty} \frac{z}{1+e^{m+\sigma z}} e^{-z^2/2} dz + \int_{-\infty}^{0} \frac{z}{1+e^{m-\sigma z}} e^{-z^2/2} dz \right) \\
&= -\frac{1}{\sqrt{2\pi}} \left(\int_{0}^{\infty} \frac{z}{1+e^{m+\sigma z}} e^{-z^2/2} dz - \int_{0}^{\infty} \frac{z}{1+e^{m-\sigma z}} e^{-z^2/2} dz \right) \\
&= -\frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \left(\frac{z}{1+e^{m+\sigma z}} - \frac{z}{1+e^{m-\sigma z}} \right) e^{-z^2/2} dz \\
&> 0.
\end{aligned}
$$

The last inequality holds since $\frac{z}{1+e^{m+\sigma z}} < \frac{z}{1+e^{m-\sigma z}}$ for each $z > 0$. Therefore, f is strictly increasing as we desired. □

A.3. Proof of Lemma 3.3

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ defined as $f(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$. Since $\nabla_{\boldsymbol{w}}^2 f(\boldsymbol{w}) = \boldsymbol{\Sigma}$ is positive definite, f is strictly convex. The strict convexity continues to hold even when we restrict the domain to $\{\boldsymbol{w} \mid \boldsymbol{w}^\top \boldsymbol{\mu} = C\}$, so $\min_{\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{w}^\top \boldsymbol{\mu} = C} f(\boldsymbol{w})$ has at most one minimizer. Let $\bar{\boldsymbol{w}} = \frac{C}{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$. Then, for any $\boldsymbol{w} \in \mathbb{R}^d$ such that $\boldsymbol{w}^\top \boldsymbol{\mu} = C$, we have

$$
f(\boldsymbol{w}) - f(\bar{\boldsymbol{w}}) \geq \nabla f(\bar{\boldsymbol{w}})^\top (\boldsymbol{w} - \bar{\boldsymbol{w}}) = (\boldsymbol{\Sigma} \bar{\boldsymbol{w}})^\top (\boldsymbol{w} - \bar{\boldsymbol{w}}) = \frac{C}{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} \boldsymbol{\mu}^\top \left(\boldsymbol{w} - \frac{C}{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right) = 0.
$$

Therefore, $\bar{\boldsymbol{w}} = \frac{C}{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$, a rescaling of $\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$, is the unique minimizer. □

A.4. Proof of Theorem 3.4

Since we consider sufficiently large κ, we may assume $n \geq d$ and let $R_\kappa := 2\|\boldsymbol{w}^*(\kappa)\|$. By Lemma 3.6, we know that $R_\kappa = \Theta(\kappa)$. Next, define a compact set $\mathcal{C}_\kappa := \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \leq R_\kappa\}$, which trivially contains $\boldsymbol{w}^*(\kappa)$. For any $\boldsymbol{w} \in \mathbb{R}^d$ and nonzero $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$
\boldsymbol{v}^\top \nabla_{\boldsymbol{w}}^2 \mathcal{L}_S(\boldsymbol{w}) \boldsymbol{v} = \frac{1}{n} \sum_{i=1}^{n} (y_i l''(\boldsymbol{w}^\top \boldsymbol{x}_i)(\boldsymbol{v}^\top \boldsymbol{x}_i)^2 + (1 - y_i) l''(-\boldsymbol{w}^\top \boldsymbol{x}_i)(\boldsymbol{v}^\top \boldsymbol{x}_i)^2) > 0,
$$

almost surely, since $\{\boldsymbol{x}_i\}_{i \in [n]}$ spans $\mathbb{R}^d$ almost surely. Therefore, $\mathcal{L}_S(\boldsymbol{w})$ is strictly convex on $\mathbb{R}^d$ almost surely and we conclude that $\mathcal{L}_S(\boldsymbol{w})$ has a unique minimizer $\hat{\boldsymbol{w}}_S^*$ on $\mathcal{C}_\kappa$ almost surely. Note that, if $\hat{\boldsymbol{w}}_S^*$ belongs to interior of $\mathcal{C}$, it is a unique minimizer of $\mathcal{L}_S(\boldsymbol{w})$ over the entire $\mathbb{R}^d$. We prove high probability convergence of $\hat{\boldsymbol{w}}_S^*$ to $\boldsymbol{w}^*(\kappa)$ using Lemma 3.5 and convert ℓ_2 convergence into directional convergence. For simplicity, we define

$$
f_i(\boldsymbol{w}) := y_i l(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1 - y_i) l(\boldsymbol{w}^\top \boldsymbol{x}_i),
$$

for each $i \in [n]$. We start with the following claim which is useful for estimating quantities described in assumptions of Lemma 3.5 for our setting.

Claim A.2. *For any $t > 0$, we have*

$$
\mathbb{E}_\kappa \left[e^{t\|\boldsymbol{x}_i\|} \right] \leq \left(2^{d/2} + e^{4\kappa^{-1} t^2 \|\boldsymbol{\Sigma}\|} \right) e^{t\|\boldsymbol{\mu}\|},
$$

for all $i \in [n]$.

Proof of Claim A.2. By applying triangular inequality and Lemma D.6, we have

$$
\begin{aligned}
\mathbb{E}_\kappa \left[e^{t\|\boldsymbol{x}_i\|} \right] &= \mathbb{E}_{\mathbf{x} \sim N(\boldsymbol{\mu}, \kappa^{-1}\boldsymbol{\Sigma})} \left[e^{t\|\mathbf{x}\|} \right] \\
&\leq \mathbb{E}_{\mathbf{x} \sim N(\boldsymbol{\mu}, \kappa^{-1}\boldsymbol{\Sigma})} \left[e^{t\|\mathbf{x} - \boldsymbol{\mu}\|} \right] e^{t\|\boldsymbol{\mu}\|} \\
&= \mathbb{E}_{\mathbf{z} \sim N(\mathbf{0}, \kappa^{-1} t^2 \boldsymbol{\Sigma})} \left[e^{\|\mathbf{z}\|} \right] e^{t\|\boldsymbol{\mu}\|} \\
&\leq \left(2^{d/2} + e^{4\kappa^{-1} t^2 \|\boldsymbol{\Sigma}\|} \right) e^{t\|\boldsymbol{\mu}\|},
\end{aligned}
$$

for each $i \in [n]$. □

Step 1: Estimate Upper Bound of $\mathbb{E}_\kappa\left[e^{|f_i(\boldsymbol{w})-\mathbb{E}_\kappa[f_i(\boldsymbol{w})]|}\right]$ on $\mathcal{C}_\kappa$

For any $\boldsymbol{w} \in \mathcal{C}_\kappa$ and $i \in [n]$,

$$|f_i(\boldsymbol{w})| = |y_i l(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i) l(-\boldsymbol{w}^\top \boldsymbol{x}_i)| \le l(-|\boldsymbol{w}^\top \boldsymbol{x}_i|) \le l(-R_\kappa \|\boldsymbol{x}_i\|).$$

Hence, we have $\mathbb{E}_\kappa\left[e^{|f_i(\boldsymbol{w})|}\right] \le \mathbb{E}_\kappa\left[1+e^{R_\kappa\|\boldsymbol{x}_i\|}\right]$. By applying Claim A.2 for $t = R_\kappa$, there exists M'_κ such that $\mathbb{E}_\kappa\left[e^{|f_i(\boldsymbol{w})|}\right] \le M'_\kappa$ and $M'_\kappa = \exp(\Theta(\kappa))$ since $R_\kappa = \Theta(\kappa)$ by Lemma 3.6. By triangular inequality and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}_\kappa\left[e^{|f_i(\boldsymbol{w})-\mathbb{E}_\kappa[f_i(\boldsymbol{w})]|}\right] &\le \mathbb{E}_\kappa\left[e^{|f_i(\boldsymbol{w})|+|\mathbb{E}_\kappa[f_i(\boldsymbol{w})]|}\right] \\
&\le \mathbb{E}_\kappa\left[e^{|f_i(\boldsymbol{w})|}\right]^2 \\
&\le {M'_\kappa}^2.
\end{aligned}$$

Defining $M_\kappa := {M'_\kappa}^2$, it follows that $M_\kappa = \exp(\Theta(\kappa))$ and $\mathbb{E}_\kappa\left[e^{|f_i(\boldsymbol{w})-\mathbb{E}_\kappa[f_i(\boldsymbol{w})]|}\right] \le M_\kappa$ for any $\boldsymbol{w} \in \mathcal{C}_\kappa$.

Step 2: Estimate Upper Bound of $\|\nabla_{\boldsymbol{w}} f_i(\boldsymbol{w})\|$ and $\|\nabla_{\boldsymbol{w}} \mathbb{E}_\kappa[f_i(\boldsymbol{w})]\|$

For each $\boldsymbol{w} \in \mathcal{C}_\kappa$ and $i \in [n]$,

$$\|\nabla_{\boldsymbol{w}} f_i(\boldsymbol{w})\| = \left\|\nabla_{\boldsymbol{w}}(y_i l(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i) l(-\boldsymbol{w}^\top \boldsymbol{x}_i))\right\| = \left\|y_i l'(\boldsymbol{w}^\top \boldsymbol{x}_i)\boldsymbol{x}_i - (1-y_i) l'(-\boldsymbol{w}^\top \boldsymbol{x}_i)\boldsymbol{x}_i\right\| \le \|\boldsymbol{x}_i\|.$$

The last inequality holds since $0 < l'(z) < 1$ for any $z \in \mathbb{R}$. In addition, by Lemma D.2,

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}} \mathbb{E}_\kappa[f_i(\boldsymbol{w})]\| &= \left\|\nabla_{\boldsymbol{w}} \mathbb{E}_\kappa[(y_i l(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i) l(-\boldsymbol{w}^\top \boldsymbol{x}_i))]\right\| \\
&= \left\|\mathbb{E}_\kappa[\nabla_{\boldsymbol{w}}(y_i l(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i) l(-\boldsymbol{w}^\top \boldsymbol{x}_i))]\right\| \\
&= \|\mathbb{E}_\kappa[(y_i l'(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i) l'(-\boldsymbol{w}^\top \boldsymbol{x}_i))\boldsymbol{x}_i]\| \\
&\le \mathbb{E}_\kappa[\|\boldsymbol{x}_i\|].
\end{aligned}$$

Also, applying Claim A.2 with $t = 1$, there exists L_κ such that $\mathbb{E}_\kappa\left[e^{\|\boldsymbol{x}_i\|}\right] < L_\kappa$ and $L_\kappa = \Theta(1)$.

Step 3: Estimate Strong Convexity Constant of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ on $\mathcal{C}_\kappa$

By Lemma D.2 and Lemma D.3, for any $\boldsymbol{w} \in \mathcal{C}_\kappa$ and unit vector $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\boldsymbol{v}^\top \nabla^2_{\boldsymbol{w}} \mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]\boldsymbol{v} = \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[l''(\boldsymbol{w}^\top \mathbf{x})(\boldsymbol{v}^\top \mathbf{x})^2] \ge \frac{1}{4}\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[e^{-(\boldsymbol{w}^\top \mathbf{x})^2/2}(\boldsymbol{v}^\top \mathbf{x})^2].$$

By Lemma D.5,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}[e^{-(\boldsymbol{w}^\top \mathbf{x})^2/2}(\boldsymbol{v}^\top \mathbf{x})^2] \\
&\ge \kappa^{d/2}\|\boldsymbol{\Sigma}\|^{-d/2}\left(\kappa\|\boldsymbol{\Sigma}^{-1}\| + \|\boldsymbol{w}\|^2\right)^{-(d+2)/2}\exp\left(-\|\boldsymbol{w}\|^2\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\mu}\|^2\right) \\
&\ge \kappa^{d/2}\|\boldsymbol{\Sigma}\|^{-d/2}\left(\kappa\|\boldsymbol{\Sigma}^{-1}\| + R_\kappa^2\right)^{-(d+2)/2}\exp\left(-R_\kappa^2\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\mu}\|^2\right),
\end{aligned}$$

and substituting $R_\kappa = \Theta(\kappa)$ to the RHS of the inequality above gives

$$\kappa^{d/2}\|\boldsymbol{\Sigma}\|^{-d/2}\left(\kappa\|\boldsymbol{\Sigma}^{-1}\| + R_\kappa^2\right)^{-(d+2)/2}\exp\left(-R_\kappa^2\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\mu}\|^2\right) = \frac{1}{\exp(\Theta(\kappa^2))}.$$

Since this hold for any unit vector $\boldsymbol{v} \in \mathbb{R}^d$, $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$ is α_κ-strongly convex with $\alpha_\kappa = \frac{1}{\exp(\Theta(\kappa^2))}$ on $\mathcal{C}_\kappa$.

Step 4: Sample Complexity for Directional Convergence

Since $\|\boldsymbol{w}^*(\kappa)\| = \Theta(\kappa)$ and $\alpha_\kappa = \frac{1}{\exp(\Theta(\kappa^2))}$, we assume κ is large enough so that $\alpha_\kappa \|\boldsymbol{w}^*(\kappa)\|^2 < 1$, which is quite easy to satisfy given the rate of decay in α_κ. Assume the unique existence of $\hat{\boldsymbol{w}}_S^*$ which occurs almost surely. By Lemma 3.5, for each $0 < \epsilon < 1$, if

$$n \geq \frac{C_1 M_\kappa}{\alpha_\kappa^2 \|\boldsymbol{w}^*(\kappa)\|^4 \epsilon^4} \log\left(\frac{3}{\delta} \max\left\{1, \left(\frac{2C_2 d^{1/2} R_\kappa L_\kappa}{\alpha_\kappa \|\boldsymbol{w}^*(\kappa)\|^2 \epsilon^2}\right)^d\right\}\right) = \frac{\exp(\Theta(\kappa^2))}{\epsilon^4}\left(1 + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right),$$

then we have $\|\boldsymbol{w}^*(\kappa) - \hat{\boldsymbol{w}}_S^*\| \leq \|\boldsymbol{w}^*(\kappa)\|\epsilon$ with probability at least $1-\delta$. Also, if $\|\boldsymbol{w}^*(\kappa) - \hat{\boldsymbol{w}}_S^*\| \leq \|\boldsymbol{w}^*(\kappa)\|\epsilon$, then $\hat{\boldsymbol{w}}_S^*$ belongs to interior of $\mathcal{C}_\kappa$. Hence, $\hat{\boldsymbol{w}}_S^*$ is a minimizer of $\mathcal{L}_S(\boldsymbol{w})$ over the entire $\mathbb{R}^d$. Also, we have

$$\begin{aligned}
\operatorname{sim}\left(\hat{\boldsymbol{w}}_S^*, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) = \operatorname{sim}\left(\hat{\boldsymbol{w}}_S^*, \boldsymbol{w}^*(\kappa)\right) &= \left(1 - \sin^2\left(\angle\left(\hat{\boldsymbol{w}}_S^*, \boldsymbol{w}^*(\kappa)\right)\right)\right)^{1/2} \\
&\geq 1 - \sin\left(\angle\left(\hat{\boldsymbol{w}}_S^*, \boldsymbol{w}^*(\kappa)\right)\right) = 1 - \frac{\|\boldsymbol{w}^*(\kappa) - \hat{\boldsymbol{w}}_S^*\|}{\|\boldsymbol{w}^*(\kappa)\|} \\
&\geq 1 - \epsilon.
\end{aligned}$$

Hence, we conclude that if $n = \frac{\exp(\Omega(\kappa^2))}{\epsilon^4}\left(1 + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$, then with probability at least $1-\delta$, the ERM loss $\mathcal{L}_S(\boldsymbol{w})$ has a unique minimizer $\hat{\boldsymbol{w}}_S^*$ and $\operatorname{sim}(\hat{\boldsymbol{w}}_S^*, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \geq 1 - \epsilon$. □

A.5. Proof of Lemma 3.5

By Lemma D.8, there exists a universal constant $C_1 > 0$ such that for any fixed $\boldsymbol{\theta} \in \mathcal{C}$ independent of the draws of $\boldsymbol{\eta}_i$'s,

$$\begin{aligned}
&\mathbb{P}\left[\left|\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\right| > \frac{\alpha\epsilon^2}{8}\right] \\
&= \mathbb{P}\left[\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{\theta}, \boldsymbol{\eta}_i) - \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\boldsymbol{\theta}, \boldsymbol{\eta})] > \frac{\alpha\epsilon^2}{8}\right] + \mathbb{P}\left[\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{\theta}, \boldsymbol{\eta}_i) - \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\boldsymbol{\theta}, \boldsymbol{\eta})] < -\frac{\alpha\epsilon^2}{8}\right] \\
&\leq 2\exp\left(-\frac{\alpha^2\epsilon^4}{C_1 M} N\right).
\end{aligned} \tag{5}$$

Notice that from the given condition and Jensen's inequality,

$$\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[g(\boldsymbol{\eta})] \leq \log \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}\left[e^{g(\boldsymbol{\eta})}\right] < \log L < L,$$

and from triangular inequality, we have

$$\begin{aligned}
\mathbb{P}\left[\sup_{\boldsymbol{\theta}\in\mathcal{C}}\left\|\nabla_{\boldsymbol{\theta}}\left(\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\right)\right\| > 3L\right] &\leq \mathbb{P}\left[\frac{1}{N}\sum_{i=1}^N \sup_{\boldsymbol{\theta}\in\mathcal{C}}\|\nabla_{\boldsymbol{\theta}}\left(f(\boldsymbol{\theta}, \boldsymbol{\eta}_i) - \mathbb{E}[f(\boldsymbol{\theta}, \boldsymbol{\eta}_i)]\right)\| > 3L\right] \\
&\leq \mathbb{P}\left[\frac{1}{N}\sum_{i=1}^N \left(g(\boldsymbol{\eta}_i) + \mathbb{E}[g(\boldsymbol{\eta}_i)]\right) > 3L\right] \\
&\leq \mathbb{P}\left[\frac{1}{N}\sum_{i=1}^N \left(g(\boldsymbol{\eta}_i) - \mathbb{E}[g(\boldsymbol{\eta}_i)]\right) > L\right].
\end{aligned} \tag{6}$$

Since we have $\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}\left[e^{|g(\boldsymbol{\eta}) - \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[g(\boldsymbol{\eta})]|}\right] \leq \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}\left[e^{g(\boldsymbol{\eta})}\right] \cdot e^{\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[g(\boldsymbol{\eta})]} < L^2$ by our assumption, we can apply Lemma D.8 to the RHS of Equation 6. Therefore, we have

$$\mathbb{P}\left[\sup_{\boldsymbol{\theta}\in\mathcal{C}}\left\|\nabla_{\boldsymbol{\theta}}\left(\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\right)\right\| > 3L\right] \leq \exp\left(-\frac{1}{C_1'}N\right), \tag{7}$$

where $C_1' > 0$ is a universal constant. Without loss of generality, we can choose $C_1 = C_1'$ that works for both Equation 5 and Equation 7.

We choose $\bar{\boldsymbol{\theta}}_1, \dots, \bar{\boldsymbol{\theta}}_m \in \mathcal{C}$ with $m \leq \max\left\{1, \left(\frac{C_2 k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\}$ where C_2 is a universal constant and satisfies following: For any $\boldsymbol{\theta} \in \mathcal{C}$, there exists $i_{\boldsymbol{\theta}} \in [m]$ such that $\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}\right\| < \frac{\alpha\epsilon^2}{24L}$. In other words, $\bar{\boldsymbol{\theta}}_1, \dots, \bar{\boldsymbol{\theta}}_m$ form an $\frac{\alpha\epsilon^2}{24L}$-cover of $\mathcal{C}$.

Suppose $|\hat{F}_N(\bar{\boldsymbol{\theta}}_k) - F(\bar{\boldsymbol{\theta}}_k)| < \frac{\alpha\epsilon^2}{8}$ for each $k \in [m]$ and $\sup_{\theta\in\mathcal{C}} \left\|\nabla_{\boldsymbol{\theta}}\left(\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\right)\right\| < 3L$ which implies $\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})$ is $3L$-Lipschitz. Then, for any $\boldsymbol{\theta} \in \mathcal{C}$, we have

$$
\begin{aligned}
&\left|\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\right| \\
&\leq \left|\hat{F}_N(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}) - F(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}})\right| + \left|\left(\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\right) - \left(\hat{F}_N(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}) - F(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}})\right)\right| \\
&\leq \frac{\alpha\epsilon^2}{8} + 3L\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}\| < \frac{\alpha\epsilon^2}{8} + \frac{\alpha\epsilon^2}{8} \\
&= \frac{\alpha\epsilon^2}{4}.
\end{aligned}
$$

By applying union bound, we conclude

$$
\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{\theta}\in\mathcal{C}} |\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})| > \frac{\alpha\epsilon^2}{4}\right] \\
&\leq 2\max\left\{1, \left(\frac{C_2 k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\} \exp\left(-\frac{\alpha^2\epsilon^4}{C_1 M} N\right) + \exp\left(-\frac{1}{C_1} N\right) \\
&\leq 3\max\left\{1, \left(\frac{C_2 k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\} \exp\left(-\min\left\{\frac{\alpha^2\epsilon^4}{C_1 M}, \frac{1}{C_1}\right\} N\right) \\
&= 3\max\left\{1, \left(\frac{C_2 k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\} \exp\left(-\frac{\alpha^2\epsilon^4}{C_1 M} N\right),
\end{aligned}
$$

where the last equality is due to $\alpha\epsilon^2 \leq 1$ and $M \geq 1$ which are implied by given conditions.

Suppose $\sup_{\boldsymbol{\theta}\in\mathcal{C}} |\hat{F}_N(\boldsymbol{\theta}) - F(\boldsymbol{\theta})| < \frac{\alpha\epsilon^2}{4}$ and $\left\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}^*\right\| > \epsilon$. Then, from the strong convexity of $F(\boldsymbol{\theta})$, we have

$$
\begin{aligned}
&\hat{F}_N(\hat{\boldsymbol{\theta}}_N^*) - \hat{F}_N(\boldsymbol{\theta}^*) \\
&= \left(\hat{F}_N(\hat{\boldsymbol{\theta}}_N^*) - F(\hat{\boldsymbol{\theta}}_N^*)\right) + \left(F(\boldsymbol{\theta}^*) - \hat{F}_N(\boldsymbol{\theta}^*)\right) + \left(F(\hat{\boldsymbol{\theta}}_N^*) - F(\boldsymbol{\theta}^*)\right) \\
&\geq -\frac{\alpha\epsilon^2}{2} + \frac{\alpha}{2}\left\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}^*\right\|^2 \\
&> 0.
\end{aligned}
$$

This is a contradiction to the fact that $\hat{\boldsymbol{\theta}}_N$ is a minimizer of $\hat{F}_N(\boldsymbol{\theta})$. Hence, we have

$$
\mathbb{P}\left[\left\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}^*\right\| > \epsilon\right] \leq 3\max\left\{1, \left(\frac{C_2 k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\} \exp\left(-\frac{\alpha^2\epsilon^4}{C_1 M} N\right),
$$

and equivalently, if

$$
N \geq \frac{C_1 M}{\alpha^2\epsilon^4} \log\left(\frac{3}{\delta}\max\left\{1, \left(\frac{C_2 k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\}\right),
$$

then $\left\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}^*\right\| < \epsilon$ with probability at least $1 - \delta$. □

A.6. Proof of Lemma 3.6

By Theorem 3.1, there exists $c_\kappa^* \geq 0$ (strict positivity of c_κ^* will be proved here) such that $\boldsymbol{w}^*(\kappa) = c_\kappa^* \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. For any $c \geq 0$, we have

$$\mathbb{E}_\kappa[\mathcal{L}_S(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] = \mathbb{E}_{X\sim N(\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}[l(cX)] = \mathbb{E}_{X\sim N(s,\kappa^{-1}s)}[l(cX)],$$

where we define $s := \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ for simplicity. By Lemma D.2, we have

$$\begin{aligned}
-\frac{\partial}{\partial c}\mathbb{E}_\kappa[\mathcal{L}_S(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] &= -\mathbb{E}_{X\sim N(s,\kappa^{-1}s)}\left[l'(cX)X\right] \\
&= \mathbb{E}_{X\sim N(s,\kappa^{-1}s)}\left[\frac{X}{1+e^{cX}}\right] \\
&= \mathbb{E}_{X\sim N(s,\kappa^{-1}s)}\left[\frac{X}{1+e^{cX}}\cdot\mathbf{1}_{X\geq 0}\right] + \mathbb{E}_{X\sim N(s,\kappa^{-1}s)}\left[\frac{X}{1+e^{cX}}\cdot\mathbf{1}_{X<0}\right].
\end{aligned}$$

For any $c \geq 0$, $1+e^{ct} \geq 2e^{ct/2}$ for all $t \in \mathbb{R}$ by the AM-GM inequality and $1+e^{ct} \leq 2$ if $t < 0$. Then, we have

$$\begin{aligned}
&-\frac{\partial}{\partial c}\mathbb{E}[\mathcal{L}_S(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] \\
&\leq \frac{1}{2}\left(\mathbb{E}_{X\sim N(s,\kappa^{-1}s)}[Xe^{-cX/2}\cdot\mathbf{1}_{X\geq 0}] + \mathbb{E}_{X\sim N(s,\kappa^{-1}s)}[X\cdot\mathbf{1}_{X<0}]\right) \\
&= \frac{1}{2}\int_0^\infty \frac{1}{\sqrt{2\pi}(\kappa^{-1}s)^{1/2}}x\exp\left(-\frac{(x-s)^2}{2\kappa^{-1}s} - \frac{cx}{2}\right)dx + \frac{1}{2}\int_{-\infty}^0 \frac{1}{\sqrt{2\pi}(\kappa^{-1}s)^{1/2}}x\exp\left(-\frac{(x-s)^2}{2\kappa^{-1}s}\right)dx \\
&= \frac{1}{2}\int_0^\infty \frac{1}{\sqrt{2\pi}(\kappa^{-1}s)^{1/2}}x\exp\left(-\frac{(x-s)^2}{2\kappa^{-1}s} - \frac{cx}{2}\right)dx - \frac{1}{2}\int_0^\infty \frac{1}{\sqrt{2\pi}(\kappa^{-1}s)^{1/2}}x\exp\left(-\frac{(x+s)^2}{2\kappa^{-1}s}\right)dx \\
&= \frac{1}{2}\int_0^\infty \frac{1}{\sqrt{2\pi}(\kappa^{-1}s)^{1/2}}x(e^{-cx/2} - e^{-2\kappa x})\exp\left(-\frac{(x-s)^2}{2\kappa^{-1}s}\right)dx.
\end{aligned}$$

If $c > 4\kappa$, then $e^{-cx/2} - e^{-2\kappa x} < 0$ for each $x > 0$ and $-\frac{\partial}{\partial c}\mathbb{E}_\kappa\left[\mathcal{L}_S(c\Sigma^{-1}\mu)\right] < 0$. Thus, $c_\kappa^* \leq 4\kappa$.

For $c \geq 0$, $1+e^{ct} \leq 2e^{ct}$ if $t \geq 0$ and $1+e^{ct} \geq 2e^{ct/2}$ for all $t \in \mathbb{R}$ by AM-GM inequality. Therefore, we have

$$\begin{aligned}
&-\frac{\partial}{\partial c}\mathbb{E}_\kappa[\mathcal{L}_S(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] \\
&\geq \frac{1}{2}\left(\mathbb{E}_{X\sim N(s,\kappa^{-1}s)}[Xe^{-cX}\cdot\mathbf{1}_{X\geq 0}] + \mathbb{E}_{X\sim N(s,\kappa^{-1}s)}[Xe^{-cX/2}\cdot\mathbf{1}_{X<0}]\right) \\
&= \frac{1}{2}\int_0^\infty \frac{1}{\sqrt{2\pi}\kappa^{-1}s}x\exp\left(-\frac{(x-s)^2}{2\kappa^{-1}s} - cx\right)dx + \frac{1}{2}\int_{-\infty}^0 \frac{1}{\sqrt{2\pi}\kappa^{-1}s}x\exp\left(-\frac{(x-s)^2}{2\kappa^{-1}s} - \frac{cx}{2}\right)dx \\
&= \frac{1}{2}\int_0^\infty \frac{1}{\sqrt{2\pi}\kappa^{-1}s}x\exp\left(-\frac{(x-s)^2}{2\kappa^{-1}s} - cx\right)dx - \frac{1}{2}\int_0^\infty \frac{1}{\sqrt{2\pi}\kappa^{-1}s}x\exp\left(-\frac{(x+s)^2}{2\kappa^{-1}s} + \frac{cx}{2}\right)dx \\
&= \frac{1}{2}\int_0^\infty \frac{1}{\sqrt{2\pi}\kappa^{-1}s}x(e^{-cx} - e^{-(2\kappa-c/2)x})\exp\left(-\frac{(x-s)^2}{2\kappa^{-1}s}\right)dx.
\end{aligned}$$

If $c < \frac{4}{3}\kappa$, $e^{-cx} - e^{-(2\kappa-c/2)x} < 0$ for each $x > 0$ and $\frac{\partial}{\partial c}\mathbb{E}_\kappa[\mathcal{L}_S(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] < 0$. Hence, $c_\kappa^* \geq \frac{4}{3}\kappa$ and we conclude $\|\boldsymbol{w}^*(\kappa)\| = \Theta(\kappa)$. □

A.7. Proof of Theorem 3.7

We first show that for sufficiently large $\kappa \in (0,\infty)$ and if n is not sufficiently large, S can be linearly separable because data points usually concentrated around $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$. Next, we characterize cosine similarity between ℓ_2 max-margin vector and the Bayes optimal solution. In our analysis, we use the following well-known lemma (See Hanson & Wright (1971); Lugosi & Mendelson (2019)).

Lemma A.1. *For positive definite matrix $\boldsymbol{M}$, let $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{M})$. For any $\delta \in (0,1)$*

$$\|\mathbf{x}\| \leq \mathrm{Tr}(\boldsymbol{M})^{1/2} + (2\|\boldsymbol{M}\| \log(1/\delta))^{1/2},$$

with probability at least $1-\delta$.

First, we introduce some technical quantities. Let $\delta = 0.99$ and $t = \left(\mathrm{Tr}(\boldsymbol{\Sigma})^{1/2} + (2\|\boldsymbol{\Sigma}\| \log(n/\delta))^{1/2}\right)/\|\boldsymbol{\mu}\|$ with

$$n \leq \delta \exp\left(\frac{1}{2}\|\boldsymbol{\Sigma}\|^{-1}\left(\kappa^{1/2}\|\boldsymbol{\mu}\| \cdot \frac{1-\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2} - \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2}\right)^2\right) = \exp(\Theta(\kappa)), \tag{8}$$

and we assume κ is large enough so that $\kappa^{1/2}\|\boldsymbol{\mu}\| \cdot \frac{1-\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2} - \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2} > 0$. Then, by substituting n in the definition of t by RHS of Equation 8 we have

$$\begin{aligned}
\|\boldsymbol{\mu}\| t &= \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2} + (2\|\boldsymbol{\Sigma}\| \log(n/\delta))^{1/2} \\
&\leq \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2} + \left[2\|\boldsymbol{\Sigma}\| \log\left(\exp\left(\frac{1}{2}\|\boldsymbol{\Sigma}\|^{-1}\left(\kappa^{1/2}\|\boldsymbol{\mu}\| \cdot \frac{1-\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2} - \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2}\right)^2\right)\right)\right]^{1/2} \\
&= \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2} + \left[2\|\boldsymbol{\Sigma}\| \left\{\frac{1}{2}\|\boldsymbol{\Sigma}\|^{-1}\left(\kappa^{1/2}\|\boldsymbol{\mu}\| \cdot \frac{1-\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2} - \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2}\right)^2\right\}\right]^{1/2} \\
&= \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2} + \left[\left(\kappa^{1/2}\|\boldsymbol{\mu}\| \cdot \frac{1-\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2} - \mathrm{Tr}(\boldsymbol{\Sigma})^{1/2}\right)^2\right]^{1/2} \\
&= \kappa^{1/2}\|\boldsymbol{\mu}\| \cdot \frac{1-\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2}.
\end{aligned}$$

Thus,

$$\kappa^{-1/2} t \leq \frac{1-\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2} < \sqrt{1-\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}. \tag{9}$$

Next, we investigate how much positive and negative data points are concentrated near their means $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$. By applying Lemma A.1 with $\boldsymbol{M} = \kappa^{-1}\boldsymbol{\Sigma}$, for each $i \in [n]$, $\|(2y_i - 1)\boldsymbol{x}_i - \boldsymbol{\mu}\| \leq \kappa^{-1/2} t \|\boldsymbol{\mu}\|$ with probability at least $1 - \delta/n$; to see why, recall the definition of t. Hence, by union bound, we have $\|(2y_i - 1)\boldsymbol{x}_i - \boldsymbol{\mu}\| \leq \kappa^{-1/2} t \|\boldsymbol{\mu}\|$ for all $i \in [n]$, with probability at least $1 - \delta$. We now condition that this event occurred and we prove that our conclusion holds. First, S is strictly linearly separable by $\boldsymbol{\mu}$ since

$$\begin{aligned}
(2y_i - 1)\boldsymbol{\mu}^\top \boldsymbol{x}_i &= \boldsymbol{\mu}^\top\left((2y_i - 1)\boldsymbol{x}_i - \boldsymbol{\mu}\right) + \|\boldsymbol{\mu}\|^2 \\
&\geq \left(-\|(2y_i - 1)\boldsymbol{x}_i - \boldsymbol{\mu}\| + \|\boldsymbol{\mu}\|\right)\|\boldsymbol{\mu}\| \\
&\geq (1 - \kappa^{-1/2} t)\|\boldsymbol{\mu}\|^2 \\
&\geq \frac{1+\mathrm{sim}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2} \cdot \|\boldsymbol{\mu}\|^2 \\
&> 0.
\end{aligned}$$

Hence, there exists ℓ_2 max-margin vector

$$\bar{\boldsymbol{w}}_S = \underset{\boldsymbol{w} \in \mathbb{R}^d}{\arg\min} \|\boldsymbol{w}\|^2 \quad \text{subject to} \quad (2y_i - 1)\boldsymbol{w}^\top \boldsymbol{x}_i \geq 1 \quad \forall i \in [n].$$

From the KKT condition of problem above, we have $\bar{\boldsymbol{w}}_S = \sum_{i=1}^n \alpha_i (2y_i - 1)\boldsymbol{x}_i$ where $\alpha_i \geq 0$ for all $i \in [n]$. By triangular inequality, we have

$$\mathrm{sim}(\bar{\boldsymbol{w}}_S, \boldsymbol{\mu}) = \frac{\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_S}{\|\boldsymbol{\mu}\|\|\bar{\boldsymbol{w}}_S\|} = \frac{\boldsymbol{\mu}^\top(\sum_{i=1}^n \alpha_i (2y_i - 1)\boldsymbol{x}_i)}{\|\boldsymbol{\mu}\|\|\sum_{i=1}^n \alpha_i (2y_i - 1)\boldsymbol{x}_i\|} \geq \frac{\sum_{i=1}^n \alpha_i \boldsymbol{\mu}^\top (2y_i - 1)\boldsymbol{x}_i}{\sum_{i=1}^n \alpha_i \|\boldsymbol{\mu}\|\|(2y_i - 1)\boldsymbol{x}_i\|}.$$

Also, for all $i \in [n]$,

$$\begin{aligned}\frac{\boldsymbol{\mu}^\top(2y_i-1)\boldsymbol{x}_i}{\|\boldsymbol{\mu}\|\|(2y_i-1)\boldsymbol{x}_i\|} &= \operatorname{sim}(\boldsymbol{\mu},(2y_i-1)\boldsymbol{x}_i) = \left(1-\sin^2\angle(\boldsymbol{\mu},(2y_i-1)\boldsymbol{x}_i)\right)^{1/2} \\ &\geq \left(1-\left(\frac{\|(2y_i-1)\boldsymbol{x}_i-\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}\|}\right)^2\right)^{1/2} \geq \sqrt{1-\kappa^{-1}t^2} \\ &\geq \operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}),\end{aligned}$$

where the last inequality used Equation 9. Hence, we have $\operatorname{sim}(\bar{\boldsymbol{w}}_S,\boldsymbol{\mu}) \geq \sqrt{1-\kappa^{-1}t^2} \geq \operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$. By triangular inequality for angle,

$$\angle(\bar{\boldsymbol{w}}_S,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \geq \angle(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) - \angle(\bar{\boldsymbol{w}}_S,\boldsymbol{\mu}) \geq \cos^{-1}(\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})) - \cos^{-1}\left(\sqrt{1-\kappa^{-1}t^2}\right) \geq 0,$$

and we have

$$\begin{aligned}&\operatorname{sim}(\bar{\boldsymbol{w}}_S,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \\ &\leq \cos\left(\cos^{-1}(\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})) - \cos^{-1}\left(\sqrt{1-\kappa^{-1}t^2}\right)\right) \\ &= \cos\left(\cos^{-1}(\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}))\right)\cdot\cos\left(\cos^{-1}\left(\sqrt{1-\kappa^{-1}t^2}\right)\right) \\ &\quad + \sin\left(\cos^{-1}(\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}))\right)\cdot\sin\left(\cos^{-1}\left(\sqrt{1-\kappa^{-1}t^2}\right)\right) \\ &= \operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\cdot\sqrt{1-\kappa^{-1}t^2} + \sqrt{1-\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}\cdot\kappa^{-1/2}t \\ &= \operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\cdot\left(\sqrt{1-\kappa^{-1}t^2}-1\right) + \sqrt{1-\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}\cdot\kappa^{-1/2}t + \operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}).\end{aligned} \tag{10}$$

It is clear that $s \mapsto s\left(\sqrt{1-\kappa^{-1}t^2}-1\right) + \sqrt{1-s^2}\cdot\kappa^{-1/2}t$ is a decreasing function on $[0,1]$ for each fixed $t \in [0,1]$. Therefore, by changing $s = \operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$ to $s = 0$, we have

$$\begin{aligned}&\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\cdot\left(\sqrt{1-\kappa^{-1}t^2}-1\right) + \sqrt{1-\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}\cdot\kappa^{-1/2}t + \operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \\ &\leq \kappa^{-1/2}t + \operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \\ &\leq \frac{1+\operatorname{sim}(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2},\end{aligned} \tag{11}$$

where the last inequality used Equation 9. Combining Equation 10 and Equation 11, we have our conclusion. □

B. Proofs for Section 4

B.1. Proof of Theorem 4.2

We first prove the existence and uniqueness of a minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ and characterize its direction in the next part.

Step 1: Existence and Uniqueness of a Minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$

Since $(\tilde{\boldsymbol{x}}_{i,i},\tilde{y}_{i,i}) = (\boldsymbol{x}_i, y_i)$ for each $i \in [n]$ and $l(\cdot)$ is non-negative, for each $\boldsymbol{w} \in \mathbb{R}^d$, we have

$$\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})] \geq \mathbb{E}_\kappa\left[\frac{1}{n^2}\sum_{i=1}^n y_i l(\boldsymbol{w}^\top\boldsymbol{x}_i) + (1-y_i)l(-\boldsymbol{w}^\top\boldsymbol{x}_i)\right] = \frac{1}{n}\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})].$$

As we discussed in Equation 3, if $\|\boldsymbol{w}\| \geq 2nm_\kappa^{-1}$, we have

$$\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})] \geq \frac{1}{n}\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})] \geq \frac{1}{n}\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[-\frac{1}{2}\boldsymbol{w}^\top\mathbf{x}\cdot\mathbf{1}_{\boldsymbol{w}^\top\mathbf{x}<0}\right]$$

$$
\begin{aligned}
&= \frac{1}{n}\|\boldsymbol{w}\|\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[-\frac{1}{2}\cdot\frac{\boldsymbol{w}^\top\mathbf{x}}{\|\boldsymbol{w}\|}\cdot\mathbf{1}_{\frac{\boldsymbol{w}^\top\mathbf{x}}{\|\boldsymbol{w}\|}<0}\right] \\
&\geq \frac{m_\kappa}{2n}\|\boldsymbol{w}\| > 1 \geq \log 2 \\
&= \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\mathbf{0})].
\end{aligned}
$$

Therefore, a minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ necessarily contained in a compact set $\{\boldsymbol{w}\in\mathbb{R}^d : \|\boldsymbol{w}\|\leq 2nm_\kappa^{-1}\}$. Since $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ is a continuous function of $\boldsymbol{w}$, there must exist a minimizer. The existence part is hence proved.

To show uniqueness, we prove strict convexity of $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$. From strict convexity of $l(\cdot)$, for any $t\in[0,1]$, $\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{R}^d$ with $\boldsymbol{w}_1\neq\boldsymbol{w}_2$, and $y\in[0,1]$, we have

$$
\begin{aligned}
&t[yl(\boldsymbol{w}_1^\top\boldsymbol{x})+(1-y)l(-\boldsymbol{w}_1^\top\boldsymbol{x})]+(1-t)[yl(\boldsymbol{w}_2^\top\boldsymbol{x})+(1-y)l(-\boldsymbol{w}_2^\top\boldsymbol{x})] \\
&> yl((t\boldsymbol{w}_1+(1-t)\boldsymbol{w}_2)^\top\boldsymbol{x})+(1-y)l(-(t\boldsymbol{w}_1+(1-t)\boldsymbol{w}_2)^\top\boldsymbol{x}),
\end{aligned}
$$

and any $\boldsymbol{x}\in\mathbb{R}^d$ except for a Lebesgue measure zero set (i.e., the set of points $\boldsymbol{x}\in\mathbb{R}^d$ satisfying $\boldsymbol{w}_1^\top\boldsymbol{x}=\boldsymbol{w}_2^\top\boldsymbol{x}$).

By taking expectations, we have

$$
t\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w}_1)]+(1-t)\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w}_2)] > \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(t\boldsymbol{w}_1+(1-t)\boldsymbol{w}_2)].
$$

We conclude $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ is strictly convex. Since a strictly convex function has at most one minimizer, we conclude that $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ has a unique minimizer for any given $\kappa\in(0,\infty)$.

Step 2: Direction of the Unique Minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$

We express expected losses $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ as the form

$$
\mathbb{E}_{Z\sim N(0,1)}\left[\sum_{i=1}^k a_i l\left(b_i\kappa^{-1/2}\left(\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right)^{1/2}Z+c_i\boldsymbol{w}^\top\boldsymbol{\mu}\right)\right], \tag{12}
$$

where a_i,b_i,c_i's are real valued random variables depending on Λ and a_i,b_i's are positive. Note that $(\tilde{\boldsymbol{x}}_{i,i},\tilde{y}_{i,i})=(\boldsymbol{x}_i,y_i)$ for each $i\in[n]$ and for each $i,j\in[n]$ with $i\neq j$,

$$
\tilde{y}_{i,j}\mid\lambda_{i,j}=\begin{cases}1 & \text{with probability } \frac{1}{4},\\ \lambda_{i,j} & \text{with probability } \frac{1}{4},\\ 1-\lambda_{i,j} & \text{with probability } \frac{1}{4},\\ 0 & \text{with probability } \frac{1}{4}.\end{cases} \tag{13}
$$

and the conditional probability distribution of the random variable $\tilde{\boldsymbol{x}}_{i,j}$ given $\tilde{y}_{i,j}$ and $\lambda_{i,j}$ can be formulated as the following four cases, depending on the outcome of $\tilde{y}_{i,j}$:

$$
\begin{cases}
\tilde{\boldsymbol{x}}_{i,j}\mid\tilde{y}_{i,j}=1,\lambda_{i,j} & \sim N\left(\boldsymbol{\mu},(g(\lambda_{i,j})^2+(1-g(\lambda_{i,j}))^2)\kappa^{-1}\boldsymbol{\Sigma}\right),\\
\tilde{\boldsymbol{x}}_{i,j}\mid\tilde{y}_{i,j}=\lambda_{i,j},\lambda_{i,j} & \sim N\left((2g(\lambda_{i,j})-1)\boldsymbol{\mu},(g(\lambda_{i,j})^2+(1-g(\lambda_{i,j}))^2)\kappa^{-1}\boldsymbol{\Sigma}\right),\\
\tilde{\boldsymbol{x}}_{i,j}\mid\tilde{y}_{i,j}=1-\lambda_{i,j},\lambda_{i,j} & \sim N\left(-(2g(\lambda_{i,j})-1)\boldsymbol{\mu},(g(\lambda_{i,j})^2+(1-g(\lambda_{i,j}))^2)\kappa^{-1}\boldsymbol{\Sigma}\right),\\
\tilde{\boldsymbol{x}}_{i,j}\mid\tilde{y}_{i,j}=0,\lambda_{i,j} & \sim N\left(-\boldsymbol{\mu},(g(\lambda_{i,j})^2+(1-g(\lambda_{i,j}))^2)\kappa^{-1}\boldsymbol{\Sigma}\right).
\end{cases} \tag{14}
$$

For simplicity, we denote $\boldsymbol{\Sigma}_\lambda=(g(\lambda)^2+(1-g(\lambda))^2)\boldsymbol{\Sigma}$ for each $\lambda\in[0,1]$. Then, we have

$$
\begin{aligned}
&\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})] \\
&= \frac{1}{n^2}\sum_{i,j=1}^n\mathbb{E}_\kappa[\tilde{y}_{i,j}l(\boldsymbol{w}^\top\tilde{\boldsymbol{x}}_{i,j})+(1-\tilde{y}_{i,j})l(-\boldsymbol{w}^\top\tilde{\boldsymbol{x}}_{i,j})] \\
&= \frac{1}{n^2}\left(\sum_{\substack{i,j=1\\ i\neq j}}^n\mathbb{E}_\kappa\left[\tilde{y}_{i,j}l(\boldsymbol{w}^\top\tilde{\boldsymbol{x}}_{i,j})+(1-\tilde{y}_{i,j})l(-\boldsymbol{w}^\top\tilde{\boldsymbol{x}}_{i,j})\right]+\sum_{i=1}^n\mathbb{E}_\kappa\left[y_il(\boldsymbol{w}^\top\boldsymbol{x}_i)+(1-y_i)l(-\boldsymbol{w}^\top\boldsymbol{x}_i)\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n-1}{2n}\mathbb{E}_{Z\sim N(0,1),\lambda\sim\Lambda}\left[l\left((\kappa^{-1}\boldsymbol{w}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w})^{1/2}Z+\boldsymbol{w}^\top\boldsymbol{\mu}\right)\right] \\
&\quad+\frac{n-1}{2n}\mathbb{E}_{Z\sim N(0,1),\lambda\sim\Lambda}\left[\lambda l\left((\kappa^{-1}\boldsymbol{w}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w})^{1/2}Z+(2g(\lambda)-1)\boldsymbol{w}^\top\boldsymbol{\mu}\right)\right] \\
&\quad+\frac{n-1}{2n}\mathbb{E}_{Z\sim N(0,1),\lambda\sim\Lambda}\left[(1-\lambda) l\left((\kappa^{-1}\boldsymbol{w}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w})^{1/2}Z-(2g(\lambda)-1)\boldsymbol{w}^\top\boldsymbol{\mu}\right)\right] \\
&\quad+\frac{1}{n}\mathbb{E}_{Z\sim N(0,1)}\left[l\left((\kappa^{-1}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w})^{1/2}Z+\boldsymbol{w}^\top\boldsymbol{\mu}\right)\right]. \qquad (15)
\end{aligned}
$$

This is the form in Equation 12. Let $C_\kappa = \boldsymbol{w}^*_{\text{mix}}(n,\kappa)^\top\boldsymbol{\mu}$, then by Lemma 3.2, $\boldsymbol{w}^*_{\text{mix}}(n,\kappa)$ have to be a solution of the problem $\min_{\boldsymbol{w}^\top\boldsymbol{\mu}=C_\kappa}\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}$, and by Lemma 3.3, there exists $c_{\text{mix},n,\kappa}$ such that $\boldsymbol{w}^*_{\text{mix}}(n,\kappa) := c_{\text{mix},n,\kappa}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ be the unique minimizer of $\mathbb{E}_\kappa[\mathcal{L}^{\text{mix}}_S(\boldsymbol{w})]$.

The only remaining part is showing $c_{\text{mix},n,\kappa} > 0$. For simplicity, we will omit κ and n in $\boldsymbol{w}^*_{\text{mix}}(n,\kappa)$. If $c_{\text{mix},n,\kappa} < 0$,

$$
\begin{aligned}
&\mathbb{E}_\kappa\left[\mathcal{L}^{\text{mix}}_S\left(\boldsymbol{w}^*_{\text{mix}}\right)\right] \\
&= \frac{n-1}{2n}\underbrace{\mathbb{E}_{\substack{Z\sim N(0,1)\\ \lambda\sim\Lambda}}\left[l\left(\left(\kappa^{-1}\boldsymbol{w}^*_{\text{mix}}{}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w}^*_{\text{mix}}\right)^{1/2}Z+c_{\text{mix},n,\kappa}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\mu\right)\right]}_{\text{(a)}} \\
&\quad+\frac{n-1}{2n}\underbrace{\mathbb{E}_{\substack{Z\sim N(0,1)\\ \lambda\sim\Lambda}}\left[\lambda l\left(\left(\kappa^{-1}\boldsymbol{w}^*_{\text{mix}}{}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w}^*_{\text{mix}}\right)^{1/2}Z+(2g(\lambda)-1)c_{\text{mix},n,\kappa}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]}_{\text{(b)}} \\
&\quad+\frac{n-1}{2n}\underbrace{\mathbb{E}_{\substack{Z\sim N(0,1)\\ \lambda\sim\Lambda}}\left[(1-\lambda) l\left(\left(\kappa^{-1}\boldsymbol{w}^*_{\text{mix}}{}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w}^*_{\text{mix}}\right)^{1/2}Z-(2g(\lambda)-1)c_{\text{mix},n,\kappa}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]}_{\text{(c)}} \\
&\quad+\frac{1}{n}\underbrace{\mathbb{E}_{Z\sim N(0,1)}\left[l\left(\left(\kappa^{-1}\boldsymbol{w}^*_{\text{mix}}{}^\top\boldsymbol{\Sigma}\boldsymbol{w}^*_{\text{mix}}\right)^{1/2}Z+c_{\text{mix},n,\kappa}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]}_{\text{(d)}} \\
&> \frac{n-1}{2n}\underbrace{\mathbb{E}_{\substack{Z\sim N(0,1)\\ \lambda\sim\Lambda}}\left[l\left(\left(\kappa^{-1}\boldsymbol{w}^*_{\text{mix}}{}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w}^*_{\text{mix}}\right)^{1/2}Z-c_{\text{mix},n,\kappa}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]}_{\text{(a)'}} \\
&\quad+\frac{n-1}{2n}\underbrace{\mathbb{E}_{\substack{Z\sim N(0,1)\\ \lambda\sim\Lambda}}\left[\lambda l\left(\left(\kappa^{-1}\boldsymbol{w}^*_{\text{mix}}{}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w}^*_{\text{mix}}\right)^{1/2}Z-(2g(\lambda)-1)c_{\text{mix},n,\kappa}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]}_{\text{(b)'}} \\
&\quad+\frac{n-1}{2n}\underbrace{\mathbb{E}_{\substack{Z\sim N(0,1)\\ \lambda\sim\Lambda}}\left[(1-\lambda) l\left(\left(\kappa^{-1}\boldsymbol{w}^*_{\text{mix}}{}^\top\boldsymbol{\Sigma}_\lambda\boldsymbol{w}^*_{\text{mix}}\right)^{1/2}Z+(2g(\lambda)-1)c_{\text{mix},n,\kappa}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]}_{\text{(c)'}} \\
&\quad+\frac{1}{n}\underbrace{\mathbb{E}_{Z\sim N(0,1)}\left[l\left(\left(\kappa^{-1}\boldsymbol{w}^*_{\text{mix}}{}^\top\boldsymbol{\Sigma}\boldsymbol{w}^*_{\text{mix}}\right)^{1/2}Z-c_{\text{mix},n,\kappa}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]}_{\text{(d)'}} \\
&= \mathbb{E}_\kappa[\mathcal{L}^{\text{mix}}_S(-\boldsymbol{w}^*_{\text{mix}})].
\end{aligned}
$$

The inequality holds by comparing (a), (b), (c), and (d) with (a)', (b)', (c)', and (d)' respectively. This contradicts the assumption that $\boldsymbol{w}^*_{\text{mix}}$ is a unique minimizer of $\mathbb{E}_\kappa[\mathcal{L}^{\text{mix}}_S(\boldsymbol{w})]$, letting us to conclude $c_{\text{mix},n,\kappa} \geq 0$. Showing $c_{\text{mix},n,\kappa}$ is strictly positive will be handled in the proof of Lemma 4.5 which can be found in Appendix B.3. □

B.2. Proof of Theorem 4.4

Since we consider sufficiently large κ, we may assume $n \geq d$ and $n \geq 2$. By Lemma 4.5, we can choose $R_\kappa = \Theta(1)$ such that $\|\boldsymbol{w}^*_{\text{mix}}(n,\kappa)\| < R_\kappa$ for *any* $n \in \mathbb{N}$. Let us define a compact set $\mathcal{C}_\kappa := \{\boldsymbol{w} \in \mathbb{R}^d \mid \|\boldsymbol{w}\| \leq R_\kappa\}$. For any vector

$\boldsymbol{w} \in \mathbb{R}^d$ and nonzero vector $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\begin{aligned}\boldsymbol{v}^\top \nabla^2_{\boldsymbol{w}} \mathcal{L}_S^{\text{mix}}(\boldsymbol{w})\boldsymbol{v} &= \frac{1}{n^2}\sum_{i,j=1}^{n} (\tilde{y}_{i,j} l''(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j})(\boldsymbol{v}^\top \tilde{\boldsymbol{x}}_{i,j})^2 + (1-\tilde{y}_{i,j}) l''(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j})(\boldsymbol{v}^\top \tilde{\boldsymbol{x}}_{i,j})^2) \\ &\geq \frac{1}{n^2}\sum_{i=1}^{n} (y_i l''(\boldsymbol{w}^\top \boldsymbol{x}_i)(\boldsymbol{v}^\top \boldsymbol{x}_i)^2 + (1-y_i) l''(-\boldsymbol{w}^\top \boldsymbol{x}_i)(\boldsymbol{v}^\top \boldsymbol{x}_i)^2) \\ &> 0,\end{aligned}$$

almost surely since $\{\boldsymbol{x}_i\}_{i=1}^n$ spans $\mathbb{R}^d$ almost surely. Therefore, $\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})$ is strictly convex almost surely and we conclude that $\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})$ has a unique minimizer $\hat{\boldsymbol{w}}^*_{\text{mix},S}$ on $\mathcal{C}$ almost surely. Also, if $\hat{\boldsymbol{w}}^*_{\text{mix},S}$ belongs to the interior of $\mathcal{C}$, then it is a unique minimizer of $\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})$ on $\mathbb{R}^d$. We prove high probability convergence of $\hat{\boldsymbol{w}}^*_{\text{mix},S}$ to $\boldsymbol{w}^*_{\text{mix}}(n,\kappa)$ using Lemma 4.3 and convert ℓ_2 convergence into directional convergence. For simplicity, we define

$$f_{i,j}(\boldsymbol{w}) := \tilde{y}_{i,j} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}) + (1-\tilde{y}_{i,j}) l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}),$$

for each $i, j \in [n]$. We start with the following claim which is useful for estimating quantities described in assumptions of Lemma 4.3 for our setting.

Claim B.1. *For any $t > 0$, we have*

$$\mathbb{E}_\kappa\left[e^{t\|\tilde{\boldsymbol{x}}_{i,j}\|}\right] \leq \left(2^{d/2} + e^{4\kappa^{-1}t^2\|\boldsymbol{\Sigma}\|}\right) e^{t\|\boldsymbol{\mu}\|},$$

for all $i, j \in [n]$.

Proof of Claim B.1. We first consider the case $i = j$. By applying triangular inequality and Lemma D.6, we have

$$\begin{aligned}\mathbb{E}_\kappa\left[e^{t\|\tilde{\boldsymbol{x}}_{i,i}\|}\right] &= \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[e^{t\|\mathbf{x}\|}\right] \\ &\leq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[e^{t\|\mathbf{x}-\boldsymbol{\mu}\|}\right] e^{t\|\boldsymbol{\mu}\|} \\ &= \mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\kappa^{-1}t^2\boldsymbol{\Sigma})}\left[e^{\|\mathbf{z}\|}\right] e^{t\|\boldsymbol{\mu}\|} \\ &\leq \left(2^{d/2} + e^{4\kappa^{-1}t^2\|\boldsymbol{\Sigma}\|}\right) e^{t\|\boldsymbol{\mu}\|},\end{aligned}$$

for each $i \in [n]$.

Next, we handle the case $i \neq j$. For simplicity, we denote $\boldsymbol{\Sigma}_\lambda = (g(\lambda)^2 + (1-g(\lambda))^2)\boldsymbol{\Sigma}$ for each $\lambda \in [0,1]$. Equation 13 and Equation 14 imply that for each $i, j \in [n]$ with $i \neq j$,

$$\begin{aligned}&\mathbb{E}_\kappa\left[e^{t\|\tilde{\boldsymbol{x}}_{i,j}\|}\right] \\ &= \mathbb{E}_{\lambda\sim\Lambda}\left[\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}\left[e^{t\|\mathbf{x}\|}\right] + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim N((2g(\lambda)-1)\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}\left[e^{t\|\mathbf{x}\|}\right]\right] \\ &\leq \frac{1}{2}\mathbb{E}_{\lambda\sim\Lambda}\left[\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}\left[e^{t\|\mathbf{x}-\boldsymbol{\mu}\|}\right] e^{t\|\boldsymbol{\mu}\|} + \mathbb{E}_{\mathbf{x}\sim N((2g(\lambda)-1)\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}\left[e^{t\|\mathbf{x}-(2g(\lambda)-1)\boldsymbol{\mu}\|}\right] e^{t\|(2g(\lambda)-1)\boldsymbol{\mu}\|}\right] \\ &\leq \mathbb{E}_{\lambda\sim\Lambda}\left[\mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}\left[e^{t\|\mathbf{z}\|}\right]\right] e^{t\|\boldsymbol{\mu}\|}.\end{aligned}$$

By Lemma D.6 for each $\lambda \in [0,1]$

$$\mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}\left[e^{t\|\mathbf{z}\|}\right] \leq 2^{d/2} + e^{4\kappa^{-1}t^2\|\boldsymbol{\Sigma}_\lambda\|} \leq 2^{d/2} + e^{4\kappa^{-1}t^2\|\boldsymbol{\Sigma}\|}.$$

and we have our conclusion. □

Step 1: Estimate Upper Bound of $\mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})-\mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]|}\right]$ on $\mathcal{C}_\kappa$

For any $\boldsymbol{w} \in \mathcal{C}$ and $i, j \in [n]$, we have

$$|f_{i,j}(\boldsymbol{w})| = |\tilde{y}_{i,j} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}) + (1-\tilde{y}_{i,j}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j})| \leq l(-|\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}|) \leq l(-R_\kappa \|\tilde{\boldsymbol{x}}_{i,j}\|) = \log\left(1 + e^{R_\kappa \|\tilde{\boldsymbol{x}}_{i,j}\|}\right).$$

By Lemma 4.5, $R_\kappa = \Theta(1)$ and by applying Claim B.1 for $t = R_\kappa$, we obtain M'_κ such that $\mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})|}\right] < M'_\kappa$ and $M'_\kappa = \Theta(1)$. Also, by triangular inequality and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})-\mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]|}\right] &\leq \mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})|+|\mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]|}\right] \\
&\leq \mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})|}\right]^2 \\
&\leq {M'_\kappa}^2.
\end{aligned}$$

Letting $M_\kappa = M'^2_\kappa$, we have

$$\mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})-\mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]|}\right] \leq M_\kappa,$$

for any $\boldsymbol{w} \in \mathcal{C}_\kappa$ and $M_\kappa = \Theta(1)$.

Step 2: Estimate Upper Bound of $\|\nabla_{\boldsymbol{w}} f_{i,j}(\boldsymbol{w})\|$ and $\|\nabla_{\boldsymbol{w}} \mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]\|$

For each $\boldsymbol{w} \in \mathcal{C}_\kappa$ and $i, j \in [n]$,

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}} f_{i,j}(\boldsymbol{w})\| &= \left\|\nabla_{\boldsymbol{w}}(\tilde{y}_{i,j} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}) + (1-\tilde{y}_{i,j}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}))\right\| \\
&= \left\|\tilde{y}_{i,j} l'(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j})\tilde{\boldsymbol{x}}_{i,j} - (1-\tilde{y}_{i,j}) l'(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j})\tilde{\boldsymbol{x}}_{i,j}\right\| \\
&\leq \|\tilde{\boldsymbol{x}}_{i,j}\|.
\end{aligned}$$

In addition, by Lemma D.2,

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}} \mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]\| &= \left\|\nabla_{\boldsymbol{w}} \mathbb{E}_\kappa[(\tilde{y}_{i,j} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}) + (1-\tilde{y}_{i,j}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}))]\right\| \\
&= \left\|\mathbb{E}_\kappa[\nabla_{\boldsymbol{w}}(\tilde{y}_{i,j} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}) + (1-\tilde{y}_{i,j}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}))]\right\| \\
&= \left\|\mathbb{E}_\kappa[\tilde{y}_{i,j} l'(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j})\tilde{\boldsymbol{x}}_{i,j} - (1-\tilde{y}_{i,j}) l'(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j})\tilde{\boldsymbol{x}}_{i,j}]\right\| \\
&\leq \mathbb{E}_\kappa[\|\tilde{\boldsymbol{x}}_{i,j}\|].
\end{aligned}$$

Also, applying Claim B.1 with $t = 1$, there exists L_κ such that $\mathbb{E}_\kappa\left[e^{\|\tilde{\boldsymbol{x}}_{i,j}\|}\right] < L_\kappa$ and $L_\kappa = \Theta(1)$.

Step 3: Estimate Strong convexity Constant of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ on $\mathcal{C}_\kappa$

For each $\lambda \in [0, 1]$, denote $\boldsymbol{\Sigma}_\lambda := (g(\lambda)^2 + (1-g(\lambda))^2)\boldsymbol{\Sigma}$, for simplicity. By Lemma D.2, Lemma D.3 and Equation 15, we have

$$\begin{aligned}
\boldsymbol{v}^\top \nabla^2_{\boldsymbol{w}} \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]\boldsymbol{v} &\geq \frac{n-1}{2n} \mathbb{E}_{\lambda\sim\Lambda}[\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}[l''(\boldsymbol{w}^\top \mathbf{x})(\boldsymbol{v}^\top \mathbf{x})^2]] \\
&\geq \frac{1}{16} \mathbb{E}_{\lambda\sim\Lambda}[\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}[e^{-(\boldsymbol{w}^\top \mathbf{x})^2/2}(\boldsymbol{v}^\top \mathbf{x})^2]],
\end{aligned}$$

for any vector $\boldsymbol{w} \in \mathcal{C}$ and unit vector $\boldsymbol{v} \in \mathbb{R}^d$. By Lemma D.5 and Lemma 4.5, for each $\lambda \in [0, 1]$,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}_\lambda)}[e^{-(\boldsymbol{w}^\top \mathbf{x})^2/2}(\boldsymbol{v}^\top \mathbf{x})^2] \\
&\geq \kappa^{d/2}\|\boldsymbol{\Sigma}_\lambda\|^{-d/2}(\kappa\|\boldsymbol{\Sigma}_\lambda^{-1}\| + R_\kappa^2)^{-(d+2)/2}\exp(-R_\kappa^2\|\boldsymbol{\Sigma}_\lambda\|\|\boldsymbol{\Sigma}_\lambda^{-1}\|\|\boldsymbol{\mu}\|^2) \\
&= \Theta(\kappa^{-1}).
\end{aligned}$$

Hence, $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ is α_κ-strongly convex with $\alpha_\kappa = \Theta(\kappa^{-1})$.

Step 4: Sample Complexity for Directional Convergence

By Lemma 4.5, we can choose $r_\kappa = \Theta(1)$ such that $r_\kappa \leq \|\boldsymbol{w}^*_{\text{mix}}(n,\kappa)\|$ for any n. Since $r_\kappa = \Theta(1)$, we assume κ is large enough so that $r_\kappa \epsilon < r_\kappa < \alpha_\kappa^{-1/2}$. Assume the unique existence of $\hat{\boldsymbol{w}}^*_{\text{mix},S}$ which occurs almost surely. By Lemma 4.3, if

$$n \geq \frac{C_1' M_\kappa}{\alpha_\kappa^2 r_\kappa^4 \epsilon^4} \log\left(\frac{3}{\delta}\max\left\{1, \left(\frac{2C_2' d^{1/2} R_\kappa L_\kappa}{\alpha_\kappa r_\kappa^2 \epsilon^2}\right)^d\right\}\right) = \frac{\Theta(\kappa^2)}{\epsilon^4}\left(1 + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right),$$

then we have $\|\boldsymbol{w}^*_{\text{mix}}(n,\kappa) - \hat{\boldsymbol{w}}^*_{\text{mix},S}\| \leq r_\kappa \epsilon \leq \|\boldsymbol{w}^*_{\text{mix}}(n,\kappa)\|\epsilon$ with probability at least $1-\delta$. Furthermore, if $\|\boldsymbol{w}^*_{\text{mix}}(n,\kappa) - \hat{\boldsymbol{w}}^*_{\text{mix},S}\| \leq \|\boldsymbol{w}^*_{\text{mix}}(n,\kappa)\|\epsilon$, then $\hat{\boldsymbol{w}}^*_{\text{mix},S}$ belongs to interior of $\mathcal{C}_\kappa$. Hence, $\hat{\boldsymbol{w}}^*_{\text{mix},S}$ is a minimizer of $\mathcal{L}^{\text{mix}}_S(\boldsymbol{w})$ over the entire $\mathbb{R}^d$. Also, we have

$$\begin{aligned}
\text{sim}(\hat{\boldsymbol{w}}^*_{\text{mix},S}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) &= \text{sim}(\hat{\boldsymbol{w}}^*_{\text{mix},S}, \boldsymbol{w}^*_{\text{mix}}(n,\kappa)) \\
&= \left(1 - \sin^2\left(\angle\big(\hat{\boldsymbol{w}}^*_{\text{mix},S}, \boldsymbol{w}^*_{\text{mix}}(n,\kappa)\big)\right)\right)^{1/2} \\
&\geq 1 - \sin\left(\angle\big(\hat{\boldsymbol{w}}^*_{\text{mix},S}, \boldsymbol{w}^*_{\text{mix}}(n,\kappa)\big)\right) \\
&\geq 1 - \epsilon.
\end{aligned}$$

Hence, we conclude that if $n = \frac{\Omega(\kappa^2)}{\epsilon^4}\left(1 + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$, then with probability at least $1-\delta$, the Mixup loss $\mathcal{L}^{\text{mix}}_S(w)$ has a unique minimizer $\hat{\boldsymbol{w}}^*_{\text{mix},S}$ and $\text{sim}(\hat{\boldsymbol{w}}^*_{\text{mix},S}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \geq 1 - \epsilon$. □

B.3. Proof of Lemma 4.5

By Theorem 4.2, let $\boldsymbol{w}^*_{\text{mix}}(n,\kappa) = c^*_{\text{mix},n,\kappa}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ where $c^*_{\text{mix},n,\kappa} \geq 0$. For any $c \in \mathbb{R}$, from Equation 15, we have

$$\begin{aligned}
\mathbb{E}_\kappa[\mathcal{L}^{\text{mix}}_S(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] = &\frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}[\mathbb{E}_{X\sim N(s,\kappa^{-1}s_\lambda)}[l(cX)]] \\
&+ \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}[\lambda\cdot\mathbb{E}_{X\sim N((2g(\lambda)-1)s,\kappa^{-1}s_\lambda)}[l(cX)]] \\
&+ \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}[(1-\lambda)\cdot\mathbb{E}_{\mathbf{x}\sim N(-(2g(\lambda)-1)s,\kappa^{-1}s_\lambda)}[l(cX)]] \\
&+ \frac{1}{n}\mathbb{E}_{X\sim N(s,\kappa^{-1}s)}\left[l(cX)\right].
\end{aligned}$$

where we denote $s_\lambda = (g(\lambda)^2 + (1-g(\lambda))^2)\boldsymbol{\mu}^\top\boldsymbol{\Sigma}\boldsymbol{\mu}$ for each $\lambda \in [0,1]$ and $s = \boldsymbol{\mu}^\top\boldsymbol{\Sigma}\boldsymbol{\mu}$, for simplicity.

By Lemma D.2, we have

$$\begin{aligned}
&-\frac{\partial}{\partial c}\mathbb{E}_\kappa[\mathcal{L}^{\text{mix}}_S(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] \\
=& -\frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}[\mathbb{E}_{X\sim N(s,\kappa^{-1}s_\lambda)}[l'(cX)X]] \\
&- \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}[\lambda\cdot\mathbb{E}_{X\sim N((2g(\lambda)-1)s,\kappa^{-1}s_\lambda)}[l'(cX)X]] \\
&- \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}[(1-\lambda)\cdot\mathbb{E}_{X\sim N(-(2g(\lambda)-1)s,\kappa^{-1}s_\lambda)}[l'(cX)X]] \\
&- \frac{1}{n}\mathbb{E}_{X\sim N(s,\kappa^{-1}s)}\left[l'(cX)X\right] \\
=& \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}\left[\mathbb{E}_{X\sim N(s,\kappa^{-1}s_\lambda)}\left[\frac{X}{1+e^{cX}}\right]\right] \\
&+ \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}\left[\lambda\cdot\mathbb{E}_{X\sim N((2g(\lambda)-1)s,\kappa^{-1}s_\lambda)}\left[\frac{X}{1+e^{cX}}\right]\right] \\
&+ \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}\left[(1-\lambda)\cdot\mathbb{E}_{X\sim N(-(2g(\lambda)-1)s,\kappa^{-1}s_\lambda)}\left[\frac{X}{1+e^{cX}}\right]\right]
\end{aligned}$$

$$+ \frac{1}{n}\mathbb{E}_{X\sim N(s,\kappa^{-1}s)}\left[\frac{X}{1+e^{cX}}\right],$$

and if $c \geq 0$, by applying Lemma D.4 with $z = cX$, we have

$$\begin{aligned}
&- \frac{\partial}{\partial c}\mathbb{E}[\mathcal{L}_S^{\text{mix}}(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] \\
&\geq \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}\left[\mathbb{E}_{X\sim N(s,\kappa^{-1}s_\lambda)}\left[\frac{1}{2}X - \frac{1}{4}cX^2\right]\right] \\
&\quad + \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}\left[\lambda\cdot\mathbb{E}_{X\sim N((2g(\lambda)-1)s,\kappa^{-1}s_\lambda)}\left[\frac{1}{2}X - \frac{1}{4}cX^2\right]\right] \\
&\quad + \frac{n-1}{2n}\mathbb{E}_{\lambda\sim\Lambda}\left[(1-\lambda)\cdot\mathbb{E}_{X\sim N(-(2g(\lambda)-1)s,\kappa^{-1}s_\lambda)}\left[\frac{1}{2}X - \frac{1}{4}cX^2\right]\right] \\
&\quad + \frac{1}{n}\mathbb{E}_{X\sim N(s,\kappa^{-1}s)}\left[\frac{1}{2}X - \frac{1}{4}cX^2\right] \\
&= \frac{s}{4n}\cdot\Big((n-1)\mathbb{E}_{\lambda\sim\Lambda}[1+(2\lambda-1)(2g(\lambda)-1)]+2\Big) \\
&\quad - \frac{c}{8n}\cdot\left(\kappa^{-1}\left(2(n-1)\mathbb{E}_{\lambda\sim\Lambda}[s_\lambda]+2s\right)+s^2\left((n-1)\mathbb{E}_{\lambda\sim\Lambda}[(2g(\lambda)-1)^2+1]+2\right)\right) \\
&\geq \frac{s}{8}\cdot\left(\mathbb{E}_{\lambda\sim\Lambda}[(2\lambda-1)(2g(\lambda)-1)] - c\Big(\kappa^{-1}\left(2\mathbb{E}_{\lambda\sim\Lambda}\left[g(\lambda)^2+(1-g(\lambda))^2\right]+1\right)+s\left(\mathbb{E}_{\lambda\sim\Lambda}[(2g(\lambda)-1)^2+2]\right)\Big)\right).
\end{aligned}$$

Thus, if $0 \leq c < \frac{\mathbb{E}_{\lambda\sim\Lambda}[(2\lambda-1)(2g(\lambda)-1)]}{\kappa^{-1}(2\mathbb{E}_{\lambda\sim\Lambda}[g(\lambda)^2+(1-g(\lambda))^2]+1)+s(\mathbb{E}_{\lambda\sim\Lambda}[(2g(\lambda)-1)^2+2])}$, then $\mathbb{E}[\mathcal{L}_S^{\text{mix}}(c\Sigma^{-1}\mu)]$ is decreasing as a function of c and we conclude $c_{\text{mix},n,\kappa} \geq \frac{\mathbb{E}_{\lambda\sim\Lambda}[(2\lambda-1)(2g(\lambda)-1)]}{\kappa^{-1}(2\mathbb{E}_{\lambda\sim\Lambda}[g(\lambda)^2+(1-g(\lambda))^2]+1)+s(\mathbb{E}_{\lambda\sim\Lambda}[(2g(\lambda)-1)^2+2])}$ which implies $\|\boldsymbol{w}^*_{\text{mix}}(n,\kappa)\| = \Omega(1)$, and one can note that this lower bound is independent of n.

In order to get the upper bound, we use the following inequality: For each $z \in \mathbb{R}$, we have

$$l(z) + l(-z) = \log(1+e^{-z}) + \log(1+e^{z}) = \log(2+e^{z}+e^{-z}) > |z|.$$

For each $c \geq 0$, we have

$$\begin{aligned}
\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] &= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}_\kappa\left[\tilde{y}_{i,j}l(c\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_{i,j}) + (1-\tilde{y}_{i,j})l(-c\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_{i,j})\right] \\
&> \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbb{E}_\kappa\left[\tilde{y}_{i,j}l(c\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_{i,j}) + (1-\tilde{y}_{i,j})l(-c\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_{i,j})\right] \\
&\geq \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbb{E}_\kappa\left[\min\{\tilde{y}_{i,j},1-\tilde{y}_{i,j}\}(l(c\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_{i,j}) + l(-c\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_{i,j}))\right] \\
&\geq \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbb{E}_\kappa\left[\min\{\tilde{y}_{i,j},1-\tilde{y}_{i,j}\}|c\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_{i,j}|\right] \\
&= \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbb{E}_\kappa\left[|\min\{\tilde{y}_{i,j},1-\tilde{y}_{i,j}\}c\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_{i,j}|\right] \\
&= \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbb{E}_\kappa\left[|\min\{\tilde{y}_{i,j},1-\tilde{y}_{i,j}\}(2g(\tilde{y}_{i,j})-1)s|\right]c \\
&= \frac{n-1}{n}\mathbb{E}_{\lambda\sim\Lambda}[\min\{\lambda,1-\lambda\}|2g(\lambda)-1|]sc
\end{aligned}$$

$$\geq \frac{1}{4}\mathbb{E}_{\lambda\sim\Lambda}[\min\{\lambda, 1-\lambda\}|2g(\lambda)-1|]sc.$$

From our assumption $\mathbb{P}_{\lambda\sim\Lambda}\left[\lambda \notin \{0,1\} \wedge g(\lambda) \neq \frac{1}{2}\right] > 0$, we have $\mathbb{E}_{\lambda\sim\Lambda}[\min(\lambda, 1-\lambda)]|2g(\lambda)-1| > 0$. Thus, if $c \geq \frac{4\log 2}{\mathbb{E}_{\lambda\sim\Lambda}[\min(\lambda,1-\lambda)|2g(\lambda)-1|]s}$, then $\mathbb{E}_\lambda[\mathcal{L}_S^{\text{mix}}(c\Sigma^{-1}\mu)] > \log 2 = \mathbb{E}_\lambda[\mathcal{L}_S^{\text{mix}}(0)]$, so $c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ cannot be a minimizer. Hence, $c^*_{\text{mix},n,\kappa} < \frac{4\log 2}{\mathbb{E}_{\lambda\sim\Lambda}[\min(\lambda,1-\lambda)|2g(\lambda)-1|]s} = \mathcal{O}(1)$, which is also independent of n. Combining the lower and upper bounds, we have our conclusion that $c^*_{\text{mix},n,\kappa} = \Theta(1)$. □

B.4. Proof of Lemma 4.3

The challenging part is that we cannot directly apply Lemma D.8 to show pointwise convergence of $\hat{F}_N(\boldsymbol{\theta})$ to $F_N(\boldsymbol{\theta})$ on $\mathcal{C}$ since $\boldsymbol{\eta}_{i,j}$'s are not independent and not identically distributed. We can overcome this problem by using Lemma D.11 which is the alternative version of Lemma D.8. We prove Lemma D.11 in Appendix D.4.

By Lemma D.11, there exists a universal constant $C_1' > 0$ such that for any $\boldsymbol{\theta} \in \mathcal{C}$, we have

$$\mathbb{P}\left[\left|\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})\right| > \frac{\alpha\epsilon^2}{8}\right] \leq 2\exp\left(-\frac{\alpha^2\epsilon^4}{C_1' M}N\right).$$

Notice that from the given condition and Jensen's inequality, for each $i, j \in [N]$,

$$\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}[g(\boldsymbol{\eta})] \leq \log \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}\left[e^{g(\boldsymbol{\eta})}\right] < \log L < L,$$

and from triangular inequality, we have

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{\theta}\in\mathcal{C}}\left\|\nabla_{\boldsymbol{\theta}}\left(\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})\right)\right\| > 3L\right] \\
&\leq \mathbb{P}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}\sup_{\boldsymbol{\theta}\in\mathcal{C}}\|\nabla_{\boldsymbol{\theta}}\left(f(\boldsymbol{\theta}, \boldsymbol{\eta}_{i,j}) - \mathbb{E}[f(\boldsymbol{\theta}, \boldsymbol{\eta}_{i,j})]\right)\| > 3L\right] \\
&\leq \mathbb{P}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}\left(g(\boldsymbol{\eta}_{i,j}) + \mathbb{E}[g(\boldsymbol{\eta}_{i,j})]\right) > 3L\right] \\
&\leq \mathbb{P}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}\left(g(\boldsymbol{\eta}_{i,j}) - \mathbb{E}[g(\boldsymbol{\eta}_{i,j})]\right) > L\right].
\end{aligned} \tag{16}$$

Since we have $\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}\left[e^{|g(\boldsymbol{\eta})-\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}[g(\boldsymbol{\eta})]|}\right] \leq \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}\left[e^{g(\boldsymbol{\eta})}\right] \cdot e^{\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}_{i,j}}[g(\boldsymbol{\eta})]} < L^2$ by our assumption, we can apply Lemma D.11 to the RHS of Equation 16. Therefore, we have

$$\mathbb{P}\left[\sup_{\boldsymbol{\theta}\in\mathcal{C}}\left\|\nabla_{\boldsymbol{\theta}}\left(\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})\right)\right\| > 3L\right] \leq \exp\left(-\frac{1}{C_1''}N\right), \tag{17}$$

where $C_1'' > 0$ is a universal constant. Without loss of generality, we can choose $C_1'' = C_1'$ that works for both Equation 16 and Equation 17.

We choose $\bar{\boldsymbol{\theta}}_1, \ldots, \bar{\boldsymbol{\theta}}_m \in \mathcal{C}$ with $m \leq \max\left\{1, \left(\frac{C_2' k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\}$ where $C_2' > 0$ is a universal constant and satisfies following: For any $\boldsymbol{\theta} \in \mathcal{C}$, there exists $i_{\boldsymbol{\theta}} \in [m]$ such that $\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}\right\| < \frac{\alpha\epsilon^2}{24L}$. In other words, $\bar{\boldsymbol{\theta}}_1, \ldots, \bar{\boldsymbol{\theta}}_m$ form an $\frac{\alpha\epsilon^2}{24L}$-cover of $\mathcal{C}$.

Suppose $|\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})| < \frac{\alpha\epsilon^2}{8}$ for each $k \in [m]$ and $\sup_{\theta\in\mathcal{C}}\left\|\nabla_{\boldsymbol{\theta}}\left(\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})\right)\right\| < 3L$ which implies $\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})$ is $3L$-Lipschitz. Then, for any $\boldsymbol{\theta} \in \mathcal{C}$, we have

$$\left|\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})\right|$$

$$\leq \left|\hat{F}_N(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}) - F_N(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}})\right| + \left|\left(\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})\right) - \left(\hat{F}_N(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}) - F_N(\bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}})\right)\right|$$
$$\leq \frac{\alpha\epsilon^2}{8} + 3L\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_{i_{\boldsymbol{\theta}}}\| < \frac{\alpha\epsilon^2}{8} + \frac{\alpha\epsilon^2}{8} = \frac{\alpha\epsilon^2}{4}.$$

By applying union bound, we conclude

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{\theta}\in\mathcal{C}} |\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})| > \frac{\alpha\epsilon^2}{4}\right] \\
&\leq 2\max\left\{1, \left(\frac{C_2' k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\} \exp\left(-\frac{\alpha^2\epsilon^4}{C_1' M} N\right) + \exp\left(-\frac{1}{C_1'} N\right) \\
&\leq 3\max\left\{1, \left(\frac{C_2' k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\} \exp\left(-\min\left\{\frac{\alpha^2\epsilon^4}{C_1' M}, \frac{1}{C_1'}\right\} N\right) \\
&= 3\max\left\{1, \left(\frac{C_2' k^{1/2} DL}{\alpha\epsilon^2}\right)^k\right\} \exp\left(-\frac{\alpha^2\epsilon^4}{C_1' M} N\right),
\end{aligned}$$

where the last equality equality is due to $\alpha\epsilon^2 \leq 1$ and $M \geq 1$ which are implied by given conditions.

Suppose $\sup_{\boldsymbol{\theta}\in\mathcal{C}} |\hat{F}_N(\boldsymbol{\theta}) - F_N(\boldsymbol{\theta})| < \frac{\alpha\epsilon^2}{4}$ and $\left\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_N^*\right\| > \epsilon$. Then, from the strong convexity of $F_N(\boldsymbol{\theta})$, we have

$$\begin{aligned}
&\hat{F}_N(\hat{\boldsymbol{\theta}}_N^*) - \hat{F}_N(\boldsymbol{\theta}_N^*) \\
&= \left(\hat{F}_N(\hat{\boldsymbol{\theta}}_N^*) - F_N(\hat{\boldsymbol{\theta}}_N^*)\right) + \left(F_N(\boldsymbol{\theta}_N^*) - \hat{F}_N(\boldsymbol{\theta}_N^*)\right) + \left(F_N(\hat{\boldsymbol{\theta}}_N^*) - F_N(\boldsymbol{\theta}_N^*))\right) \\
&\geq -\frac{\alpha\epsilon^2}{2} + \frac{\alpha}{2}\left\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_N^*\right\|^2 > 0.
\end{aligned}$$

This is a contradiction to the fact that $\hat{\boldsymbol{\theta}}_N$ is a minimizer of $\hat{F}_N(\boldsymbol{\theta})$. Hence, we have

$$\mathbb{P}\left[\left\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_N^*\right\| > \epsilon\right] \leq 3\max\left\{1, \left(\frac{C_2' k^{1/2} LD}{\alpha\epsilon^2}\right)^k\right\} \exp\left(-\frac{\alpha^2\epsilon^4}{C_1' M} N\right),$$

and equivalently, if

$$N \geq \frac{C_1' M}{\alpha^2\epsilon^4} \log\left(\frac{3}{\delta}\max\left\{1, \left(\frac{C_2' k^{1/2} LD}{\alpha\epsilon^2}\right)^k\right\}\right),$$

then $\left\|\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_N^*\right\| < \epsilon$ with probability at least $1 - \delta$. □

C. Proofs for Section 5

Before moving on to proof of the main results of Section 5, We would like to represent $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ in a simpler form. Note that $(\tilde{\boldsymbol{x}}_{i,i}^{\text{mask}}, \tilde{y}_{i,i}^{\text{mask}}) = (\boldsymbol{x}_i, y_i)$ for each $i \in [n]$. For $i, j \in [n]$ with $i \neq j$, conditioning on $(\boldsymbol{M}_{i,j}, \lambda_{i,j}) \sim \mathcal{M}$, we have

$$\begin{cases}
\tilde{y}_{i,j} = 1, & \tilde{x}_{i,j} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} \odot \left(\boldsymbol{M}_{i,j}\boldsymbol{M}_{i,j}^\top + (\mathbf{1} - \boldsymbol{M}_{i,j})(\mathbf{1} - \boldsymbol{M}_{i,j})^\top\right)) & \text{with probability } \frac{1}{4} \\
\tilde{y}_{i,j} = \lambda_{i,j}, & \tilde{x}_{i,j} \sim N((2\boldsymbol{M}_{i,j} - 1) \odot \boldsymbol{\mu}, \boldsymbol{\Sigma} \odot \left(\boldsymbol{M}_{i,j}\boldsymbol{M}_{i,j}^\top + (\mathbf{1} - \boldsymbol{M}_{i,j})(\mathbf{1} - \boldsymbol{M}_{i,j})^\top\right)) & \text{with probability } \frac{1}{4} \\
\tilde{y}_{i,j} = 1 - \lambda_{i,j}, & \tilde{x}_{i,j} \sim N(-(2\boldsymbol{M}_{i,j} - 1)\boldsymbol{\mu}, \boldsymbol{\Sigma} \odot \left(\boldsymbol{M}_{i,j}\boldsymbol{M}_{i,j}^\top + (\mathbf{1} - \boldsymbol{M}_{i,j})(\mathbf{1} - \boldsymbol{M}_{i,j})^\top\right)) & \text{with probability } \frac{1}{4} \\
\tilde{y}_{i,j} = 0, & \tilde{x}_{i,j} \sim N(-\boldsymbol{\mu}, \boldsymbol{\Sigma} \odot \left(\boldsymbol{M}_{i,j}\boldsymbol{M}_{i,j}^\top + (\mathbf{1} - \boldsymbol{M}_{i,j})(\mathbf{1} - \boldsymbol{M}_{i,j})^\top\right)) & \text{with probability } \frac{1}{4}
\end{cases}.$$

Let $\text{support}(\boldsymbol{M}) = \left\{\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(p)}\right\}$ where $(\boldsymbol{M}, \lambda) \sim \mathcal{M}$. For each $k \in [p]$, define

$$\boldsymbol{\mu}^{(k)} = \boldsymbol{\mu} \odot (2\boldsymbol{M}^{(k)} - 1), \boldsymbol{\Sigma}^{(k)} = \boldsymbol{\Sigma} \odot \left(\boldsymbol{M}^{(k)}\boldsymbol{M}^{(k)\top} + (\mathbf{1} - \boldsymbol{M}^{(k)})(\mathbf{1} - \boldsymbol{M}^{(k)})^\top\right),$$

and

$$a_k = \frac{1}{2}\mathbb{E}_{(\boldsymbol{M},\lambda)\sim\mathcal{M}}\left[\lambda \mathbf{1}_{\boldsymbol{M}=\boldsymbol{M}^{(k)}}\right], b_k = \frac{1}{2}\mathbb{E}_{(\boldsymbol{M},\lambda)\sim\mathcal{M}}\left[(1-\lambda)\mathbf{1}_{\boldsymbol{M}=\boldsymbol{M}^{(k)}}\right], c_k = \frac{1}{2}\mathbb{P}_{(\boldsymbol{M},\lambda)\sim\mathcal{M}}\left[\boldsymbol{M}=\boldsymbol{M}^{(k)}\right].$$

By Assumption 5.2, $a_1,\dots,a_p,b_1,\dots,b_p,c_1,\dots,c_p > 0$. One can check that $a_k + b_k = c_k$ for each $k \in [p]$ and $\sum_{k=1}^{p}(a_k+b_k+c_k) = 1$. Then, for each $\kappa \in (0,\infty)$, we can rewrite $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ as the following form:

$$\begin{aligned}
&\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_\kappa\left[y_i l(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i)l(-\boldsymbol{w}^\top \boldsymbol{x}_i)\right] + \sum_{\substack{i,j\in[n]\\ i\neq j}}\mathbb{E}_\kappa\left[\tilde{y}_{i,j}^{\text{mask}} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) + (1-\tilde{y}_{i,j}^{\text{mask}})l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}})\right]\right) \\
&= \frac{1}{n}\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[l(\boldsymbol{w}^\top \mathbf{x})\right] \\
&\quad + \frac{n-1}{n}\left(\sum_{k=1}^{p}\mathbb{E}_{\mathbf{x}^{(k)}\sim N(\boldsymbol{\mu}^{(k)},\kappa^{-1}\boldsymbol{\Sigma}^{(k)})}\left[a_k l\left(\boldsymbol{w}^\top \mathbf{x}^{(k)}\right) + b_k l\left(-\boldsymbol{w}^\top \mathbf{x}^{(k)}\right)\right] + c_k \mathbb{E}_{\mathbf{z}^{(k)}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}^{(k)})}\left[l\left(\boldsymbol{w}^\top \mathbf{z}^{(k)}\right)\right]\right),
\end{aligned} \tag{18}$$

and

$$\mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] = \frac{1}{n}l(\boldsymbol{w}^\top\boldsymbol{\mu}) + \frac{n-1}{n}\sum_{k=1}^{p}\left(a_k l\left(\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)}\right) + b_k l\left(-\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)}\right) + c_k l\left(\boldsymbol{w}^\top\boldsymbol{\mu}\right)\right). \tag{19}$$

In addition, notice that for each $k \in [p]$, we have

$$\begin{cases}
\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}} \sim N\left(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right) & \text{with probability } \frac{c_k}{2} \\
\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}} \sim N\left(-\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right) & \text{with probability } \frac{c_k}{2} \\
\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}} \sim N\left(\boldsymbol{\mu}^{(k)},\kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right) & \text{with probability } \frac{c_k}{2} \\
\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}} \sim N\left(-\boldsymbol{\mu}^{(k)},\kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right) & \text{with probability } \frac{c_k}{2}
\end{cases}. \tag{20}$$

C.1. Proof of Theorem 5.3

We first prove the uniqueness of a minimizer $\boldsymbol{w}_{\text{mask}}^*(n,\kappa)$ of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ for all $\kappa \in (0,\infty]$ and show that they are bounded. Next, we prove that $\boldsymbol{w}_{\text{mask}}^*(n,\kappa)$ converges to $\boldsymbol{w}_{\text{mask}}^*(n,\infty)$ as $\kappa \to \infty$.

Step 1: $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ has a Unique Minimizer $\boldsymbol{w}_{\text{mask}}^*(n,\kappa)$ and $\{\boldsymbol{w}_{\text{mask}}^*(n,\kappa) : \kappa \in (0,\infty]\}$ is Bounded

We prove the strict convexity of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ for each $\kappa \in (0,\infty]$. Consider the case $\kappa \neq \infty$ first. For $t \in [0,1]$, and $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$ with $\boldsymbol{w}_1 \neq \boldsymbol{w}_2$,

$$tl\left(\boldsymbol{w}_1^\top \boldsymbol{x}\right) + (1-t)l\left(\boldsymbol{w}_2^\top \boldsymbol{x}\right) > l\left((t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)^\top \boldsymbol{x}\right),$$

and

$$tl\left(-\boldsymbol{w}_1^\top \boldsymbol{x}\right) + (1-t)l\left(-\boldsymbol{w}_2^\top \boldsymbol{x}\right) > l\left(-(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)^\top \boldsymbol{x}\right),$$

for any $\boldsymbol{x} \in \mathbb{R}^d$ except a Lebesgue measure zero set (hyperplane orthogonal to $\boldsymbol{w}_1 - \boldsymbol{w}_2$).

By Equation 18 and taking expectation, we have for any $t \in [0,1]$, and $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$,

$$t\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}_1)] + (1-t)\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}_2)] > \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)].$$

For the case $\kappa = \infty$, by Assumption 5.2, there exist $i_1,\dots,i_d \in [m]$ such that $\left\{\boldsymbol{\mu}^{(i_1)},\dots,\boldsymbol{\mu}^{(i_d)}\right\}$ spans $\mathbb{R}^d$ and for any $t \in [0,1]$ and $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$ with $\boldsymbol{w}_1 \neq \boldsymbol{w}_2$, at least one of $k \in [d]$ satisfies

$$tl\left(\boldsymbol{w}_1^\top \boldsymbol{\mu}^{(i_k)}\right) + (1-t)l\left(\boldsymbol{w}_2^\top \boldsymbol{\mu}^{(i_k)}\right) > l\left((t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)^\top \boldsymbol{\mu}^{(i_k)}\right),$$

and

$$tl\left(-\boldsymbol{w}_1^\top \boldsymbol{\mu}^{(i_k)}\right) + (1-t)l\left(-\boldsymbol{w}_2^\top \boldsymbol{\mu}^{(i_k)}\right) > l\left(-(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)^\top \boldsymbol{\mu}^{(i_k)}\right).$$

From Equation 19, we can conclude the strict convexity of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$.

In order to complete this step, we prove the existence of a ball containing all minimizers of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$. For the case $\kappa \neq \infty$, from Equation 18, we have

$$\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] \geq \frac{1}{2}\sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \mathbb{E}_{\mathbf{x}^{(k)} \sim N(\boldsymbol{\mu}^{(i_k)}, \kappa^{-1}\boldsymbol{\Sigma}^{(i_k)})} \left[l\left(\boldsymbol{w}^\top \mathbf{x}^{(k)}\right) + l\left(-\boldsymbol{w}^\top \mathbf{x}^{(k)}\right)\right].$$

Since $l(z) + l(-z) \geq |z|$ for each $z \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] &\geq \frac{1}{2}\sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \mathbb{E}_{\mathbf{x}^{(k)} \sim N(\boldsymbol{\mu}^{(i_k)}, \kappa^{-1}\boldsymbol{\Sigma}^{(i_k)})} \left[\left|\boldsymbol{w}^\top \mathbf{x}^{(k)}\right|\right] \\
&\geq \frac{1}{2}\sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \left|\mathbb{E}_{\mathbf{x}^{(k)} \sim N(\boldsymbol{\mu}^{(i_k)}, \kappa^{-1}\boldsymbol{\Sigma}^{(i_k)})} \left[\boldsymbol{w}^\top \mathbf{x}^{(k)}\right]\right| \\
&= \frac{1}{2}\sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \left|\boldsymbol{w}^\top \boldsymbol{\mu}^{(i_k)}\right|.
\end{aligned}$$

Also, for the case $\kappa = \infty$, from Equation 19 and using similar argument, we have

$$\begin{aligned}
\mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] &\geq \frac{1}{2}\sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \left(l\left(\boldsymbol{w}^\top \boldsymbol{\mu}^{(k)}\right) + l\left(-\boldsymbol{w}^\top \boldsymbol{\mu}^{(k)}\right)\right) \\
&\geq \frac{1}{2}\sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \left|\boldsymbol{w}^\top \boldsymbol{\mu}^{(i_k)}\right|.
\end{aligned}$$

Since $\left\{\mu^{(i_1)}, \ldots, \mu^{(i_d)}\right\}$ spans $\mathbb{R}^d$, for any unit vector $\boldsymbol{u} \in \mathbb{R}^d$, $\sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \left|\boldsymbol{u}^\top \boldsymbol{\mu}^{(i_k)}\right| > 0$ and $\boldsymbol{u} \mapsto \sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \left|\boldsymbol{u}^\top \boldsymbol{\mu}^{(i_k)}\right|$ has the minimum value $m > 0$ since $\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\| = 1\}$ is compact and the mapping is continuous. If $\|\boldsymbol{w}\| > \frac{2\log 2}{m}$, then we have

$$\begin{aligned}
\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] &\geq \frac{1}{2}\sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \left|\boldsymbol{w}^\top \boldsymbol{\mu}^{(i_k)}\right| \\
&= \|\boldsymbol{w}\| \sum_{k=1}^{d} \min\{a_{i_k}, b_{i_k}\} \left|\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)^\top \boldsymbol{\mu}^{(i_k)}\right| \\
&\geq \frac{1}{2}\|\boldsymbol{w}\| m \geq \log 2 \\
&= \mathbb{E}_\kappa\left[\mathcal{L}_S^{\text{mask}}(\mathbf{0})\right].
\end{aligned}$$

Hence, for any $\kappa \in (0, \infty]$, the minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ is contained in the ball centered at origin with radius $R := \frac{2\log 2}{m}$. Together with the strict convexity of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$, we have our conclusion. □

Step 2: $\boldsymbol{w}^*_{\text{mask}}(n, \kappa)$ Converges to $\boldsymbol{w}^*_{\text{mask}}(n, \infty)$ as $\kappa \to \infty$

For each $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\| \leq R$ and unit vector $\boldsymbol{v} \in \mathbb{R}^d$, Equation 19 implies

$$\begin{aligned}
&\boldsymbol{v}^\top \nabla_{\boldsymbol{w}}^2 \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]\boldsymbol{v} \\
&= \frac{1}{n} l''(\boldsymbol{w}^\top \boldsymbol{\mu})(\boldsymbol{v}^\top \boldsymbol{\mu})^2 + \frac{n-1}{n}\sum_{k=1}^{p}\left[\left(a_k l''\left(\boldsymbol{w}^\top \boldsymbol{\mu}^{(k)}\right) + b_k l''\left(-\boldsymbol{w}^\top \boldsymbol{\mu}^{(k)}\right)\right)\left(\boldsymbol{v}^\top \boldsymbol{\mu}^{(k)}\right)^2 + c_k l''\left(\boldsymbol{w}^\top \boldsymbol{\mu}\right)\left(\boldsymbol{v}^\top \boldsymbol{\mu}\right)^2\right]
\end{aligned}$$

$$\geq \frac{1}{2}\sum_{k=1}^{p}(a_k+b_k)l''\left(R\|\boldsymbol{\mu}^{(k)}\|\right)\left(\boldsymbol{v}^\top\boldsymbol{\mu}^{(k)}\right)^2.$$

By Assumption 5.2, at least one of $k \in [p]$ satisfies $\boldsymbol{v}^\top\boldsymbol{\mu}^{(k)} \neq 0$ and thus, $\frac{1}{2}\sum_{k=1}^{p}(a_k+b_k)l''\left(R\|\boldsymbol{\mu}^{(k)}\|\right)\left(\boldsymbol{v}^\top\boldsymbol{\mu}^{(k)}\right)^2 > 0$. Since $\boldsymbol{v} \mapsto \frac{1}{2}\sum_{k=1}^{p}(a_k+b_k)l''\left(R\|\boldsymbol{\mu}^{(k)}\|\right)\left(\boldsymbol{v}^\top\boldsymbol{\mu}^{(k)}\right)^2$ is continuous and $\{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\| = 1\}$ is compact, it has minimum $\alpha > 0$ on this set. Hence, $\mathbb{E}_\infty\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\right]$ is α-strongly convex on $\left\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \leq R\right\}$ and since $\boldsymbol{w}^*_{\text{mask}}(n,\kappa) \in \left\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \leq R\right\}$ for each $\kappa \in (0,\infty)$, we have

$$\begin{aligned}
&\frac{\alpha}{2}\|\boldsymbol{w}^*_{\text{mask}}(n,\kappa) - \boldsymbol{w}^*_{\text{mask}}(n,\infty)\|^2 \\
&\leq \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\kappa))] - \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\infty))] \\
&\leq \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\kappa))] - \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\kappa))] \\
&\quad + \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\kappa))] - \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\infty))] \\
&\quad + \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\infty))] - \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\infty))] \\
&\leq \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\kappa))] - \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\kappa))] + \mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\infty))] - \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}^*_{\text{mask}}(n,\infty))].
\end{aligned}$$

The last inequality holds since $\boldsymbol{w}^*_{\text{mask}}(n,\kappa)$ is a minimizer of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$. For any $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\| \leq R$, by Lemma D.7, we have

$$\begin{aligned}
&\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] - \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] \\
&= \frac{1}{n}\mathbb{E}_{X\sim N(\boldsymbol{w}^\top\boldsymbol{\mu},\kappa^{-1}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w})}[l(X) - l(\boldsymbol{w}^\top\boldsymbol{\mu})] \\
&\quad + \frac{n-1}{n}\Bigg(\sum_{k=1}^{p}\mathbb{E}_{X^{(k)}\sim N(\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)},\kappa^{-1}\boldsymbol{w}^\top\boldsymbol{\Sigma}^{(k)}\boldsymbol{w})}\left[a_k\left(l\left(X^{(k)}\right) - l\left(\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)}\right)\right) + b_k\left(l\left(-X^{(k)}\right) - l\left(-\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)}\right)\right)\right] \\
&\qquad\qquad + c_k\mathbb{E}_{X^{(k)}\sim N(\boldsymbol{w}^\top\boldsymbol{\mu},\kappa^{-1}\boldsymbol{w}^\top\boldsymbol{\Sigma}^{(k)}\boldsymbol{w})}\left[l\left(X^{(k)}\right) - l(\boldsymbol{w}^\top\boldsymbol{\mu})\right]\Bigg) \\
&\leq \kappa^{-1/2}\left(\frac{1}{n}(\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w})^{1/2} + \frac{n-1}{n}\sum_{k=1}^{p}(a_k+b_k+c_k)\left(\boldsymbol{w}^\top\boldsymbol{\Sigma}^{(k)}\boldsymbol{w}\right)^{1/2}\right) \\
&\leq \kappa^{-1/2}\|\boldsymbol{w}\|\left(\frac{\|\boldsymbol{\Sigma}\|^{1/2}}{n} + \frac{n-1}{n}\sum_{k=1}^{p}(a_k+b_k+c_k)\left\|\boldsymbol{\Sigma}^{(k)}\right\|^{1/2}\right) \\
&\leq R\kappa^{-1/2}\left(\frac{\|\boldsymbol{\Sigma}\|^{1/2}}{n} + \frac{n-1}{n}\sum_{k=1}^{p}(a_k+b_k+c_k)\left\|\boldsymbol{\Sigma}^{(k)}\right\|^{1/2}\right), \\
&\leq R\kappa^{-1/2}\left(\|\boldsymbol{\Sigma}\|^{1/2} + \sum_{k=1}^{p}(a_k+b_k+c_k)\left\|\boldsymbol{\Sigma}^{(k)}\right\|^{1/2}\right).
\end{aligned}$$

Therefore,

$$\frac{\alpha}{2}\|\boldsymbol{w}^*_{\text{mask}}(n,\kappa) - \boldsymbol{w}^*_{\text{mask}}(n,\infty)\|^2 \leq 2R\kappa^{-1/2}\left(\|\boldsymbol{\Sigma}\| + \sum_{k=1}^{p}(a_k+b_k+c_k)\left\|\boldsymbol{\Sigma}^{(k)}\right\|^{1/2}\right), \tag{21}$$

and we conclude $\lim_{\kappa\to\infty}\boldsymbol{w}^*_{\text{mask}}(n,\kappa) = \boldsymbol{w}^*_{\text{mask}}(n,\infty)$. □

C.2. Proof of Theorem 5.4

Since our sample complexity bound contains $\Omega(1)$ which can hide d, we may assume $n \geq d$ and $n \geq 2$. Let R be the upper bound on $\{\|\boldsymbol{w}^*_{\text{mask}}(n,\kappa)\| : \kappa \in (0,\infty]\}$ which we defined in Appendix C.1. Next, define a compact set $\mathcal{C} := \{\boldsymbol{w} \in \mathbb{R}^d \mid \|\boldsymbol{w}\| \leq 2R\}$, which trivially contains $\boldsymbol{w}^*_{\text{mask}}(n,\kappa)$ for all $\kappa \in (0,\infty]$. For any $\boldsymbol{w} \in \mathbb{R}^d$ and nonzero $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\boldsymbol{v}\nabla^2_{\boldsymbol{w}}\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\boldsymbol{v} = \frac{1}{n^2}\sum_{i,j=1}^{n}\left(\tilde{y}_{i,j}^{\text{mask}}l''\left(w^\top\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right)\left(\boldsymbol{v}^\top\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right)^2 + \left(1-\tilde{y}_{i,j}^{\text{mask}}\right)l''\left(-\boldsymbol{w}^\top\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right)\left(\boldsymbol{v}^\top\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right)^2\right)$$

$$\geq \frac{1}{n^2}\sum_{i=1}^{n}(y_i l''(\boldsymbol{w}^\top \boldsymbol{x}_i)(\boldsymbol{v}^\top \boldsymbol{x}_i)^2 + (1-y_i)l''(-\boldsymbol{w}^\top \boldsymbol{x}_i)(\boldsymbol{v}^\top \boldsymbol{x}_i)^2) > 0,$$

almost surely, since $\{\boldsymbol{x}_i\}_{i\in[n]}$ spans $\mathbb{R}^d$ almost surely. Therefore, $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})$ is strictly convex almost surely and we conclude that $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})$ has a unique minimizer $\hat{\boldsymbol{w}}^*_{\text{mask},S}$ on $\mathcal{C}$ almost surely. Also, if $\hat{\boldsymbol{w}}^*_{\text{mask},S}$ belong to interior of $\mathcal{C}$, it is minimizer of $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})$ on $\mathbb{R}^d$. We will prove high probability convergence of $\hat{\boldsymbol{w}}^*_{\text{mask},S}$ to $\boldsymbol{w}^*_{\text{mask}}(n,\kappa)$ using Lemma 4.3 and convert ℓ_2 convergence into directional convergence. For simplicity, we define

$$f_{i,j}(\boldsymbol{w}) := \tilde{y}_{i,j}^{\text{mask}} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) + (1-\tilde{y}_{i,j}^{\text{mask}}) l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}),$$

for each $i,j\in[n]$. We start with the following claim which is useful for estimating quantities described in assumptions of Lemma 4.3 for our setting.

Claim C.1. *For any $t>0$, we have*

$$\mathbb{E}_\kappa\left[e^{t\left\|\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right\|}\right] \leq \max\left\{2^{d/2}+e^{4\kappa^{-1}t^2\|\boldsymbol{\Sigma}\|}, \sum_{k=1}^{p} 2c_k\left(2^{d/2}+e^{4\kappa^{-1}t^2\left\|\boldsymbol{\Sigma}^{(k)}\right\|}\right)\right\} e^{t\|\boldsymbol{\mu}\|},$$

for all $i,j\in[n]$.

Proof of Claim C.1. We first consider the case $i=j$. By invoking triangular inequality and Lemma D.6, we have

$$\begin{aligned}
\mathbb{E}_\kappa\left[e^{t\left\|\tilde{\boldsymbol{x}}_{i,i}^{\text{mask}}\right\|}\right] &= \mathbb{E}_\kappa\left[e^{t\|\boldsymbol{x}\|}\right] = \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[e^{t\|\mathbf{x}\|}\right] \\
&= \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[e^{t\|\mathbf{x}\|}\right] \leq \mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[e^{t\|\mathbf{x}-\boldsymbol{\mu}\|}\right] e^{t\|\boldsymbol{\mu}\|} \\
&= \mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\kappa^{-1}t^2\boldsymbol{\Sigma})}\left[e^{\|\mathbf{z}\|}\right] e^{t\|\boldsymbol{\mu}\|} \\
&= \left(2^{d/2}+e^{4\kappa^{-1}t^2\|\boldsymbol{\Sigma}\|}\right) e^{t\|\boldsymbol{\mu}\|}.
\end{aligned}$$

for each $i\in[n]$. Next, we handle the case $i\neq j$. From Equation 20, for any $i,j\in[n]$ with $i\neq j$, we have

$$\begin{aligned}
&\mathbb{E}_\kappa\left[e^{t\left\|\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right\|}\right] \\
&= \sum_{k=1}^{p} c_k\left(\mathbb{E}_{\mathbf{x}\sim N\left(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right)}\left[e^{t\|\mathbf{x}\|}\right] + \mathbb{E}_{\mathbf{x}\sim N\left(\boldsymbol{\mu}^{(k)},\kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right)}\left[e^{t\|\mathbf{x}\|}\right]\right) \\
&\leq \sum_{k=1}^{p} c_k\left(\mathbb{E}_{\mathbf{x}\sim N\left(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right)}\left[e^{t\|\mathbf{x}-\boldsymbol{\mu}\|}\right] e^{t\|\boldsymbol{\mu}\|} + \mathbb{E}_{\mathbf{x}\sim N\left(\boldsymbol{\mu}^{(k)},\kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right)}\left[e^{t\|\mathbf{x}-\boldsymbol{\mu}^{(k)}\|}\right] e^{t\|\boldsymbol{\mu}^{(k)}\|}\right) \\
&= \sum_{k=1}^{p} 2c_k \mathbb{E}_{\mathbf{z}\sim N\left(\mathbf{0},\kappa^{-1}t^2\boldsymbol{\Sigma}^{(k)}\right)}\left[e^{\|\mathbf{z}\|}\right] e^{t\|\boldsymbol{\mu}\|}.
\end{aligned}$$

By applying Lemma D.6,

$$\mathbb{E}_{\mathbf{z}\sim N\left(\mathbf{0},\kappa^{-1}t^2\boldsymbol{\Sigma}^{(k)}\right)}\left[e^{\|\mathbf{z}\|}\right] \leq 2^{d/2}+e^{4\kappa^{-1}t^2\left\|\boldsymbol{\Sigma}^{(k)}\right\|},$$

and we have our conclusion. □

Step 1: Estimate Upper Bound of $\mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})-\mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]|}\right]$ on $\mathcal{C}$

For any $\boldsymbol{w}\in\mathcal{C}$ and $i,j\in[n]$, we have

$$\begin{aligned}
|f_{i,j}(\boldsymbol{w})| &= |\tilde{y}_{i,j}^{\text{mask}} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) + (1-\tilde{y}_{i,j}^{\text{mask}}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}})| \\
&\leq l(-|\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}|) \\
&\leq l(-2R\|\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\|)
\end{aligned}$$

$$= \log\left(1 + e^{2R\|\tilde{x}_{i,j}^{\text{mask}}\|}\right).$$

By applying Claim C.1 for $t = 2R$, there exists M'_κ such that $M'_\kappa = \Theta(1)$ and

$$\mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})|}\right] \leq \mathbb{E}_\kappa\left[1 + e^{2R\|\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\|}\right] < M'_\kappa.$$

From triangular inequality and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w}) - \mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]|}\right] &\leq \mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})| + |\mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]|}\right] \\
&\leq \mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w})|}\right]^2 \\
&\leq {M'_\kappa}^2.
\end{aligned}$$

Letting $M_\kappa = {M'_\kappa}^2$, it follows that $M_\kappa = \Theta(1)$ and $\mathbb{E}_\kappa\left[e^{|f_{i,j}(\boldsymbol{w}) - \mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]|}\right] < M_\kappa$ for all $\boldsymbol{w} \in \mathcal{C}$.

Step 2: Estimate Upper Bound of $\|\nabla_{\boldsymbol{w}} f_{i,j}(\boldsymbol{w})\|$ and $\|\nabla_{\boldsymbol{w}} \mathbb{E}_\kappa[f_{i,j}(\boldsymbol{w})]\|$

For each $\boldsymbol{w} \in \mathcal{C}$ and $i, j \in [n]$,

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}} f_{i,j}(\boldsymbol{w})\| &= \left\|\nabla_{\boldsymbol{w}}(\tilde{y}_{i,j}^{\text{mask}} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) + (1 - \tilde{y}_{i,j}^{\text{mask}}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}))\right\| \\
&= \left\|\tilde{y}_{i,j}^{\text{mask}} l'(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}} - (1 - \tilde{y}_{i,j}^{\text{mask}}) l'(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right\| \\
&\leq \left\|\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right\|.
\end{aligned}$$

In addition, by Lemma D.2,

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}} \mathbb{E}[f_{i,j}(\boldsymbol{w})]\| &= \left\|\nabla_{\boldsymbol{w}} \mathbb{E}\left[(\tilde{y}_{i,j}^{\text{mask}} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) + (1 - \tilde{y}_{i,j}^{\text{mask}}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}))\right]\right\| \\
&= \left\|\mathbb{E}\left[\nabla_{\boldsymbol{w}}(\tilde{y}_{i,j}^{\text{mask}} l(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) + (1 - \tilde{y}_{i,j}^{\text{mask}}) l(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}))\right]\right\| \\
&= \left\|\mathbb{E}\left[\tilde{y}_{i,j}^{\text{mask}} l'(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}} - (1 - \tilde{y}_{i,j}^{\text{mask}}) l'(-\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}) \tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right]\right\| \\
&\leq \mathbb{E}\left[\left\|\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\right\|\right].
\end{aligned}$$

Also, by applying Claim C.1 with $t = 1$, there exists L_κ such that $\mathbb{E}_\kappa\left[e^{\|\tilde{\boldsymbol{x}}_{i,j}^{\text{mask}}\|}\right] < L_\kappa$ and $L_\kappa = \Theta(1)$.

Step 3: Estimate Strong Convexity Constant of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ on $\mathcal{C}$

By Lemma D.2, Lemma D.3 and Equation 18, we have

$$\begin{aligned}
\boldsymbol{v} \nabla_{\boldsymbol{w}}^2 \mathbb{E}_\kappa\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})\right] \boldsymbol{v} &\geq \frac{n-1}{n} \sum_{k=1}^{p} a_k \mathbb{E}_{\mathbf{x}^{(k)} \sim N\left(\boldsymbol{\mu}^{(k)}, \kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right)}\left[l''\left(\boldsymbol{w}^\top \mathbf{x}^{(k)}\right)\left(\boldsymbol{v}^\top \mathbf{x}^{(k)}\right)^2\right] \\
&\geq \frac{1}{8} \min_{k \in [p]}\{a_k\} \sum_{k=1}^{p} \mathbb{E}_{\mathbf{x}^{(k)} \sim N\left(\boldsymbol{\mu}^{(k)}, \kappa^{-1}\boldsymbol{\Sigma}^{(k)}\right)}\left[e^{-\left(\boldsymbol{w}^\top \mathbf{x}^{(k)}\right)^2/2}\left(\boldsymbol{v}^\top \mathbf{x}^{(k)}\right)^2\right],
\end{aligned}$$

for any $\boldsymbol{w} \in \mathcal{C}$ and unit vector $\boldsymbol{v} \in \mathbb{R}^d$.

For each $k \in [p]$,

$$\begin{aligned}
&\boldsymbol{v}^\top \boldsymbol{\Sigma}^{(k)} \boldsymbol{v} \\
&= \left(\boldsymbol{M}^{(k)} \odot \boldsymbol{v}\right)^\top \boldsymbol{\Sigma}\left(\boldsymbol{M}^{(k)} \odot \boldsymbol{v}\right) + \left(\left(\mathbf{1} - \boldsymbol{M}^{(k)}\right) \odot \boldsymbol{v}\right)^\top \boldsymbol{\Sigma}\left(\left(\mathbf{1} - \boldsymbol{M}^{(k)}\right) \odot \boldsymbol{v}\right) \\
&\geq \left\|\boldsymbol{\Sigma}^{-1}\right\|^{-1} \left\|\boldsymbol{M}^{(k)} \odot \boldsymbol{v}\right\|^2 + \left\|\boldsymbol{\Sigma}^{-1}\right\|^{-1} \left\|\left(\mathbf{1} - \boldsymbol{M}^{(k)}\right) \odot \boldsymbol{v}\right\|^2 \\
&= \left\|\boldsymbol{\Sigma}^{-1}\right\|^{-1} \|\boldsymbol{v}\|^2.
\end{aligned}$$

Hence, $\mathbf{\Sigma}^{(k)}$ is positive definite for all $k \in [p]$ and by Lemma D.5,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}^{(k)} \sim N\left(\boldsymbol{\mu}^{(k)}, \kappa^{-1}\mathbf{\Sigma}^{(k)}\right)} \left[e^{-\left(\boldsymbol{w}^\top \mathbf{x}^{(k)}\right)^2/2} \left(\boldsymbol{v}^\top \mathbf{x}^{(k)}\right)^2 \right] \\
&\geq \left(\frac{1}{2} \left(\boldsymbol{v}^\top \boldsymbol{\mu}^{(k)}\right)^2 - \kappa^{-2} \left\|\mathbf{\Sigma}^{(k)}\right\|^2 \|\boldsymbol{w}\|^4 \left\|\boldsymbol{\mu}^{(k)}\right\|^2 \right) \\
&\quad \cdot \left(\kappa^{-1} \left\|\mathbf{\Sigma}^{(k)}\right\| \left(\kappa \left\| \left(\mathbf{\Sigma}^{(k)}\right)^{-1} \right\| + \|\boldsymbol{w}\|^2 \right) \right)^{-d/2} \exp\left(-\|\boldsymbol{w}\|^2 \left\| \left(\mathbf{\Sigma}^{(k)}\right)^{-1} \right\| \left\|\mathbf{\Sigma}^{(k)}\right\| \left\|\boldsymbol{\mu}^{(k)}\right\|^2 \right) \\
&\geq \left(\frac{1}{2} \left(\boldsymbol{v}^\top \boldsymbol{\mu}^{(k)}\right)^2 - 16\kappa^{-2} \left\|\mathbf{\Sigma}^{(k)}\right\|^2 R^4 \left\|\boldsymbol{\mu}^{(k)}\right\|^2 \right) \\
&\quad \cdot \left(\kappa^{-1} \left\|\mathbf{\Sigma}^{(k)}\right\| \left(\kappa \left\| \left(\mathbf{\Sigma}^{(k)}\right)^{-1} \right\| + 4R^2 \right) \right)^{-d/2} \exp\left(-4R^2 \left\| \left(\mathbf{\Sigma}^{(k)}\right)^{-1} \right\| \left\|\mathbf{\Sigma}^{(k)}\right\| \left\|\boldsymbol{\mu}^{(k)}\right\|^2 \right) \\
&= \left(\boldsymbol{v}^\top \boldsymbol{\mu}^{(k)}\right)^2 \Theta(1).
\end{aligned}$$

By Assumption 5.2, $\sum_{k=1}^p \left(\boldsymbol{v}^\top \boldsymbol{\mu}^{(k)}\right)^2 > 0$ for each unit vector $\boldsymbol{v} \in \mathbb{R}^d$ and since $\boldsymbol{v} \mapsto \sum_{k=1}^p \left(\boldsymbol{v}^\top \boldsymbol{\mu}^{(k)}\right)^2$ is continuous, we conclude that $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ is α_k-strongly convex on $\mathcal{C}$ where $\alpha_\kappa = \Theta(1)$. In addition, we can choose α_κ small enough so that for sufficiently large κ, $\|\boldsymbol{w}^*_{\text{mask}}(n, \kappa)\| \leq R < \alpha_\kappa^{-1/2}$ since $\|\boldsymbol{w}^*_{\text{mask}}(n, \kappa)\| = \Theta(1)$. This choice of α_κ makes it possible to apply Lemma 4.3.

Step 4: Lower Bounds on $\|\boldsymbol{w}^*_{\text{mask}}(n, \kappa)\|$ for Sufficiently Large n and κ

We need lower bounds on $\|\boldsymbol{w}^*_{\text{mask}}(n, \kappa)\|$ that are independent of n when we apply Lemma 4.3 in our final step. However, finding such lower bounds is challenging since we do not know the exact direction of $\boldsymbol{w}^*_{\text{mask}}(n, \kappa)$ unlike $\boldsymbol{w}^*(\kappa)$ and $\boldsymbol{w}^*_{\text{mix}}(n, \kappa)$. In addition, the fact that $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})$ is dependent on n also makes it hard. Instead, we will focus on sufficiently large n and κ and look for lower bounds independent of n, which is sufficient for our analysis.

We introduce a function $\mathcal{L}_\infty^{\text{mask}} : \mathbb{R}^d \to \mathbb{R}$ which corresponds to the limit case of $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ as $n \to \infty$ and defined as follows (i.e., the limit when both n and κ approach ∞):

$$\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w}) := \lim_{n\to\infty} \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})] = \sum_{k=1}^p \left(a_k l\left(\boldsymbol{w}^\top \boldsymbol{\mu}^{(k)}\right) + b_k l\left(-\boldsymbol{w}^\top \boldsymbol{\mu}^{(k)}\right) + c_k l\left(\boldsymbol{w}^\top \boldsymbol{\mu}\right) \right). \tag{22}$$

Analyzing a minimizer of $\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w})$ is helpful because it is independent of n and approximates $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})$ for large enough n and κ.

Recall that we choose $i_1, \ldots, i_d \in [m]$ such that $\left\{\boldsymbol{\mu}^{(i_1)}, \ldots, \boldsymbol{\mu}^{(i_d)}\right\}$ spans $\mathbb{R}^d$ in Appendix C.1. For any $t \in [0, 1]$ and $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$ with $\boldsymbol{w}_1 \neq \boldsymbol{w}_2$, at least one of $k \in [d]$ satisfies

$$tl\left(\boldsymbol{w}_1^\top \boldsymbol{\mu}^{(i_k)}\right) + (1-t)l\left(\boldsymbol{w}_2^\top \boldsymbol{\mu}^{(i_k)}\right) > l\left((t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)^\top \boldsymbol{\mu}^{(i_k)}\right),$$

and

$$tl\left(-\boldsymbol{w}_1^\top \boldsymbol{\mu}^{(i_k)}\right) + (1-t)l\left(-\boldsymbol{w}_2^\top \boldsymbol{\mu}^{(i_k)}\right) > l\left(-(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)^\top \boldsymbol{\mu}^{(i_k)}\right).$$

We can conclude the strict convexity of $\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w})$. Also, since $l(z) + l(-z) \geq |z|$ for each $z \in \mathbb{R}$, we have

$$\begin{aligned}
\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w}) &\geq \frac{1}{2} \sum_{k=1}^d \min\{a_{i_k}, b_{i_k}\} \left(l\left(\boldsymbol{w}^\top \boldsymbol{\mu}^{(k)}\right) + l\left(-\boldsymbol{w}^\top \boldsymbol{\mu}^{(k)}\right) \right) \\
&\geq \frac{1}{2} \sum_{k=1}^d \min\{a_{i_k}, b_{i_k}\} \left|\boldsymbol{w}^\top \boldsymbol{\mu}^{(i_k)}\right|.
\end{aligned}$$

In Appendix C.1, we have shown that for any unit vector $\boldsymbol{u} \in \mathbb{R}^d$, $\sum_{k=1}^d \min\{a_{i_k}, b_{i_k}\} \left|\boldsymbol{u}^\top \boldsymbol{\mu}^{(i_k)}\right| > 0$ and $\boldsymbol{u} \mapsto \sum_{k=1}^d \min\{a_{i_k}, b_{i_k}\} \left|\boldsymbol{u}^\top \boldsymbol{\mu}^{(i_k)}\right|$ has the minimum value $m > 0$. If $\|\boldsymbol{w}\| > R$, where we previously defined $R := \frac{2\log 2}{m}$,

then we have

$$\begin{aligned}\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w}) &\geq \frac{1}{2}\sum_{k=1}^{d}\min\{a_{i_k}, b_{i_k}\}\left|\boldsymbol{w}^\top\boldsymbol{\mu}^{(i_k)}\right| \\ &= \|\boldsymbol{w}\|\sum_{k=1}^{d}\min\{a_{i_k}, b_{i_k}\}\left|\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)^\top\boldsymbol{\mu}^{(i_k)}\right| \\ &\geq \frac{1}{2}\|\boldsymbol{w}\|m \geq \log 2 \\ &= \mathcal{L}_\infty^{\text{mask}}(\mathbf{0}).\end{aligned}$$

Hence, a minimizer of $\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w})$ contained in the ball centered origin with radius R. Together with the strict convexity, we can conclude that $\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w})$ has the unique minimizer $\boldsymbol{w}_{\text{mask}}^*$ and it satisfies $\|\boldsymbol{w}_{\text{mask}}^*\| \leq R$.

In addition, we would like to prove that $\boldsymbol{w}_{\text{mask}}^*$ is nonzero. This will make our lower bounds on $\|\boldsymbol{w}_{\text{mask}}^*(n,\kappa)\|$ positive. Since $\boldsymbol{\mu}$ is nonzero, without loss of generality, we assume 1st coordinate of $\boldsymbol{\mu}$, namely μ_1, is nonzero. We consider a weight $r\boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is 1st standard basis and $r > 0$ which will be chosen later. We have

$$\begin{aligned}&\mathcal{L}_\infty^{\text{mask}}(r\boldsymbol{e}_1) - \mathcal{L}_\infty^{\text{mask}}(\mathbf{0}) \\ &= \sum_{k=1}^{p}\left[a_k\left(l\left(r\boldsymbol{e}_1^\top\boldsymbol{\mu}^{(k)}\right) - l(0)\right) + b_k\left(l\left(-r\boldsymbol{e}_1^\top\boldsymbol{\mu}^{(k)}\right) - l(0)\right) + c_k\left(l\left(r\boldsymbol{e}_1^\top\boldsymbol{\mu}\right) - l(0)\right)\right] \\ &= \sum_{k\in\mathcal{I}}\left[a_k\big(l\left(r\mu_1\right) - l(0)\big) + b_k\big(l\left(-r\mu_1\right) - l(0)\big) + c_k\big(l\left(r\mu_1\right) - l(0)\big)\right],\end{aligned}$$

where $\mathcal{I} \subset [p]$ is the index set satisfying 1st coordinate of $\boldsymbol{M}^{(k)}$ is 1 for each $k \in \mathcal{I}$. From our definition of a_k, b_k, c_k's, $a_k + b_k = c_k$, thus we have

$$\begin{aligned}&\sum_{k\in\mathcal{I}}\left[a_k\big(l\left(r\mu_1\right) - l(0)\big) + b_k\big(l\left(-r\mu_1\right) - l(0)\big) + c_k\big(l\left(r\mu_1\right) - l(0)\big)\right] \\ &= \sum_{k\in\mathcal{I}}\left[a_k\big(l\left(r\mu_1\right) - l\left(-r\mu_1\right)\big) + c_k\big(l\left(r\mu_1\right) - l(0)\big)\right] \\ &= \sum_{k\in\mathcal{I}}\left[-a_k r\mu_1 + c_k\big(l\left(r\mu_1\right) + l\left(-r\mu_1\right) - 2l(0)\big)\right] \\ &= \sum_{k\in\mathcal{I}} c_k r\mu_1\left(\frac{l\left(r\mu_1\right) + l\left(-r\mu_1\right) - 2l(0)}{r\mu_1} - \frac{a_k}{c_k}\right).\end{aligned}$$

Since $\lim_{z\to 0}\frac{l(z)+l(-z)-2l(0)}{z} = 0$, we can choose $r > 0$ small enough so that $\frac{l(r\mu_1)+l(-r\mu_1)-2l(0)}{r\mu_1} < \frac{a_k}{c_k}$ for all $k \in \mathcal{I}$. Then, we obtain $\mathcal{L}_\infty^{\text{mask}}(r\boldsymbol{e}_1) - \mathcal{L}_\infty^{\text{mask}}(\mathbf{0}) < 0$ and thus $\boldsymbol{w}_{\text{mask}}^*$ is nonzero.

Next, we would like to characterize the strong convexity of $\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w})$. For each $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\| \leq R$ and unit vector $\boldsymbol{v} \in \mathbb{R}^d$, Equation 22 implies

$$\begin{aligned}&\boldsymbol{v}^\top\nabla_{\boldsymbol{w}}^2\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w})\boldsymbol{v} \\ &= \sum_{k=1}^{p}\left[\left(a_k l''\left(\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)}\right) + b_k l''\left(-\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)}\right)\right)\left(\boldsymbol{v}^\top\boldsymbol{\mu}^{(k)}\right)^2 + c_k l''\left(\boldsymbol{w}^\top\boldsymbol{\mu}\right)\left(\boldsymbol{v}^\top\boldsymbol{\mu}\right)^2\right] \\ &\geq \frac{1}{2}\sum_{k=1}^{p}(a_k + b_k)l''\left(R\|\boldsymbol{\mu}^{(k)}\|\right)\left(\boldsymbol{v}^\top\boldsymbol{\mu}^{(k)}\right)^2.\end{aligned}$$

Recall that we have shown that $\boldsymbol{v} \mapsto \frac{1}{2}\sum_{k=1}^{p}(a_k + b_k)l''\left(R\|\boldsymbol{\mu}^{(k)}\|\right)\left(\boldsymbol{v}^\top\boldsymbol{\mu}^{(k)}\right)^2$ has minimum $\alpha > 0$ on the unit sphere $\{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\| = 1\}$ in Appendix C.1. Hence, $\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w})$ is α-strongly convex on $\left\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \leq R\right\}$. Since $\boldsymbol{w}_{\text{mask}}^*(n,\infty)$ and $\boldsymbol{w}_{\text{mask}}^*$ are contained in $\left\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \leq R\right\}$, we have

$$\frac{\alpha}{2}\|\boldsymbol{w}_{\text{mask}}^*(n,\infty) - \boldsymbol{w}_{\text{mask}}^*\|^2$$

$$
\begin{aligned}
&\leq \mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w}_{\text{mask}}^*(n,\infty)) - \mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w}_{\text{mask}}^*) \\
&= \left(\mathcal{L}_\infty^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*(n,\infty)\big) - \mathbb{E}_\infty\left[\mathcal{L}_S^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*(n,\infty)\big)\right]\right) \\
&\quad + \left(\mathbb{E}_\infty\left[\mathcal{L}_S^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*(n,\infty)\big)\right] - \mathbb{E}_\infty\left[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w}_{\text{mask}}^*)\right]\right) \\
&\quad + \left(\mathbb{E}_\infty\left[\mathcal{L}_S^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*\big)\right] - \mathcal{L}_\infty^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*\big)\right) \\
&\leq \left(\mathcal{L}_\infty^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*(n,\infty)\big) - \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*(n,\infty)\big)]\right) + \left(\mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*\big)] - \mathcal{L}_\infty^{\text{mask}}\big(\boldsymbol{w}_{\text{mask}}^*\big)\right).
\end{aligned}
$$

For each $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\| \leq R$,

$$
\begin{aligned}
&\left|\mathcal{L}_\infty^{\text{mask}}(\boldsymbol{w}) - \mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]\right| \\
&\leq \frac{1}{n}\left|l(\boldsymbol{w}^\top\boldsymbol{\mu})\right| + \frac{1}{n}\left|\sum_{k=1}^p \left(a_k l\left(\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)}\right) + b_k l\left(-\boldsymbol{w}^\top\boldsymbol{\mu}^{(k)}\right) + c_k l\left(\boldsymbol{w}^\top\boldsymbol{\mu}\right)\right)\right| \\
&\leq \frac{1}{n}\left(l(-R|\boldsymbol{\mu}\|) + \sum_{k=1}^p \left(a_k l\left(-R\left\|\boldsymbol{\mu}^{(k)}\right\|\right) + b_k l\left(-R\left\|\boldsymbol{\mu}^{(k)}\right\|\right) + c_k l\left(-R\|\boldsymbol{\mu}\|\right)\right)\right) \\
&= \frac{l(-R\|\boldsymbol{\mu}\|)}{n}\left(1 + \sum_{k=1}^p (a_k + b_k + c_k)\right) \\
&= \frac{2l(-R\|\boldsymbol{\mu}\|)}{n}.
\end{aligned}
$$

The last two inequalities are due to $\left\|\boldsymbol{\mu}^{(k)}\right\| = \|\boldsymbol{\mu}\|$ and definition of a_k, b_k, c_k's. Hence, we have

$$
\frac{\alpha}{2}\|\boldsymbol{w}_{\text{mask}}^*(n,\infty) - \boldsymbol{w}_{\text{mask}}^*\|^2 \leq \frac{4l(-R\|\boldsymbol{\mu}\|)}{n}. \tag{23}
$$

From triangular inequality, Equation 21, and Equation 23, we have

$$
\begin{aligned}
\|\boldsymbol{w}_{\text{mask}}^*\| &\leq \|\boldsymbol{w}_{\text{mask}}^*(n,\kappa)\| + \|\boldsymbol{w}_{\text{mask}}^*(n,\infty) - \boldsymbol{w}_{\text{mask}}^*(n,\kappa)\| + \|\boldsymbol{w}_{\text{mask}}^* - \boldsymbol{w}_{\text{mask}}^*(n,\infty)\| \\
&\leq \|\boldsymbol{w}_{\text{mask}}^*(n,\kappa)\| + \left(\frac{8l(-R\|\boldsymbol{\mu}\|)}{\alpha n}\right)^{1/2} + \left[\frac{4R}{\alpha\kappa^{1/2}}\left(\|\boldsymbol{\Sigma}\|^{1/2} + \sum_{k=1}^p (a_k + b_k + c_k)\left\|\boldsymbol{\Sigma}^{(k)}\right\|^{1/2}\right)\right]^{1/2}.
\end{aligned}
$$

Thus, if

$$
n \geq \frac{128 l(-R\|\boldsymbol{\mu}\|)}{\alpha\|\boldsymbol{w}_{\text{mask}}^*\|^2}, \kappa \geq \left(\frac{64R\left(\|\boldsymbol{\Sigma}\|^{1/2} + \sum_{k=1}^p (a_k + b_k + c_k)\left\|\boldsymbol{\Sigma}^{(k)}\right\|^{1/2}\right)}{\alpha\|\boldsymbol{w}_{\text{mask}}^*\|^2}\right)^2, \tag{24}
$$

then we have $\|\boldsymbol{w}_{\text{mask}}^*(n,\kappa)\| \geq \frac{\|\boldsymbol{w}_{\text{mask}}^*\|}{2}$. Notice that the lower bounds in Equation 24 are numerical constants independent of κ.

Step 5: Sample Complexity for Directional Convergence

Assume n and κ is large enough so that satisfies Equation 24 and $\frac{\|\boldsymbol{w}_{\text{mask}}^*\|}{2}\epsilon \leq \|\boldsymbol{w}_{\text{mask}}^*(n,\kappa)\|\epsilon < \|\boldsymbol{w}_{\text{mask}}^*(n,\kappa)\| < \alpha_\kappa^{-1/2}$ We also assume the unique existence of $\hat{\boldsymbol{w}}_{\text{mask},S}^*$ which occurs almost surely. By Lemma 4.3, if

$$
n \geq \frac{16C_1' M_\kappa}{\alpha_\kappa^2\|\boldsymbol{w}_{\text{mask}}^*\|^4\epsilon^4}\log\left(\frac{3}{\delta}\max\left\{1, \left(\frac{8C_2' d^{1/2} R L_\kappa}{\alpha_\kappa\|\boldsymbol{w}_{\text{mask}}^*\|\epsilon^2}\right)^d\right\}\right) = \frac{\Theta(1)}{\epsilon^4}\left(1 + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right),
$$

then we have $\|\boldsymbol{w}_{\text{mask}}^*(n,\kappa) - \hat{\boldsymbol{w}}_{\text{mask},S}^*\| \leq \frac{\|\boldsymbol{w}_{\text{mask}}^*\|}{2}\epsilon \leq \|\boldsymbol{w}_{\text{mask}}^*(n,\kappa)\|\epsilon$ with probability at least $1-\delta$. Furthermore, if $\|\boldsymbol{w}_{\text{mask}}^*(n,\kappa) - \hat{\boldsymbol{w}}_{\text{mask},S}^*\| \leq \|\boldsymbol{w}_{\text{mask}}^*(n,\kappa)\|\epsilon$, then $\hat{\boldsymbol{w}}_{\text{mask},S}^*$ belongs to interior of $\mathcal{C}$ therefore, $\hat{\boldsymbol{w}}_{\text{mask},S}^*$ is a minimizer of $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})$ over the entire $\mathbb{R}^d$. Also, we have

$$
\text{sim}(\hat{\boldsymbol{w}}_{\text{mask},S}^*, \boldsymbol{w}_{\text{mask}}^*(n,\kappa)) = \left(1 - \sin^2\left(\angle\big(\hat{\boldsymbol{w}}_{\text{mask},S}^*, \boldsymbol{w}_{\text{mask}}^*(n,\kappa)\big)\right)\right)^{1/2}
$$

$$\geq 1 - \sin\left(\angle\big(\hat{\boldsymbol{w}}^*_{\text{mask},S}, \boldsymbol{w}^*_{\text{mask}}(n,\kappa)\big)\right)$$
$$\geq 1 - \epsilon.$$

Hence, we conclude that if $n = \frac{\Omega(1)}{\epsilon^4}\left(1 + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$, then with probability at least $1-\delta$, the masking based Mixup loss $\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})$ has a unique minimizer $\hat{\boldsymbol{w}}^*_{\text{mask},S}$ and $\text{sim}(\hat{\boldsymbol{w}}^*_{\text{mask},S}, \boldsymbol{w}^*_{\text{mask}}(n,\kappa)) \geq 1-\epsilon$. □

D. Technical Lemmas

In this section, we introduce several technical lemmas that previously appeared. Before moving on, we define some additional notation which will be used in this section.

Notation. For each (random) vector $\boldsymbol{u} \in \mathbb{R}^k$, we use $(\boldsymbol{u})_i$ to refer to the ith component of $\boldsymbol{u}$ and for each matrix $\boldsymbol{M} \in \mathbb{R}^{k\times k}$, we use $(\boldsymbol{M})_i$ to represent ith diagonal entry of $\boldsymbol{M}$ for $i \in [k]$. In addition, we use $\|\boldsymbol{u}\|_{\boldsymbol{M}}$ for $\left(\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u}\right)^{1/2}$.

D.1. The Bayes Optimal Classifier for $\mathcal{D}_\kappa$

To get the Bayes optimal classifier for $\mathcal{D}_\kappa$ for $\kappa \in (0,\infty)$, we have to solve the following problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_\kappa}\left[(2y-1)\boldsymbol{w}^\top\boldsymbol{x} > 0\right]. \tag{25}$$

From our definition of $\mathcal{D}_\kappa$ we have

$$\begin{aligned}\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_\kappa}\left[(2y-1)\boldsymbol{w}^\top\boldsymbol{x} > 0\right] &= \mathbb{P}_{\mathbf{x}\sim N(\boldsymbol{\mu},\kappa^{-1}\boldsymbol{\Sigma})}\left[\boldsymbol{w}^\top\mathbf{x} > 0\right] \\ &= \mathbb{P}_{Z\sim N(0,1)}\left[\kappa^{-1/2}\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}} Z + \boldsymbol{w}^\top\boldsymbol{\mu} > 0\right] \\ &= \mathbb{P}\left[Z > -\frac{\kappa^{1/2}\boldsymbol{w}^\top\boldsymbol{\mu}}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}}\right].\end{aligned}$$

Without loss of generality, consider the solution of Problem (25) with $\boldsymbol{w}^\top\boldsymbol{\mu} = 1$. We can change Problem (25) to the problem $\min_{\boldsymbol{w}^\top\boldsymbol{\mu}=1} \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}$ and by Lemma 3.3, we have our conclusion. □

D.2. Interchanging Differentiation and Expectation

We will introduce technical results related to interchanging differentiation and expectation. The following Lemma D.1 is a slight variant of Leibniz's rule.

Lemma D.1. *Let $f : U \times \mathbb{R}^k \to \mathbb{R}$ be a function where U is an open subset of $\mathbb{R}$. Suppose a probability distribution $\mathcal{P}$ on $\mathbb{R}^k$ satisfies the following conditions:*

1. *$\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\theta,\boldsymbol{\eta})] < \infty$ for all $\theta \in \mathbb{R}$.*
2. *For any $\theta \in U$, $\frac{\partial}{\partial\theta} f(\theta,\boldsymbol{\eta})$ exists for every $\boldsymbol{\eta} \in \mathbb{R}$.*
3. *There is $g : \mathbb{R}^k \to \mathbb{R}$ such that $|\frac{\partial}{\partial\theta} f(\theta,\boldsymbol{\eta})| \leq g(\boldsymbol{\eta})$ for each $\theta \in U$ and $\boldsymbol{\eta} \in \mathbb{R}^k$. In addition, $\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[g(\boldsymbol{\eta})] < \infty$.*

Then, $\frac{d}{d\theta}\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\theta,\boldsymbol{\eta})] = \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}\left[\frac{\partial}{\partial\theta} f(\theta,\boldsymbol{\eta})\right]$ for all $\theta \in U$.

Proof of Lemma D.1. Fix any $\theta \in U$ and let $\{h_m\}_{m\in\mathbb{N}}$ be any sequence of nonzero real numbers such that $h_m \to 0$ as $m \to \infty$ and $\theta + h_m \in U$. Define $f_{\theta,m} : \mathbb{R} \to \mathbb{R}$ as $f_{\theta,m}(\boldsymbol{\eta}) = \frac{1}{h_m}\left(f\left(\theta + h_m, \boldsymbol{\eta}\right) - f\left(\theta,\boldsymbol{\eta}\right)\right)$. Then, $f_{\theta,m}(\boldsymbol{\eta}) \to \frac{\partial}{\partial\theta} f(\theta,\boldsymbol{\eta})$ as $m \to \infty$. Therefore, for large enough m, $\left|\frac{\partial}{\partial\theta} f(\theta,\boldsymbol{\eta}) - f_{\theta,m}(\boldsymbol{\eta})\right| < 1$ holds, which implies $|f_{\theta,m}(\boldsymbol{\eta})| \leq g(\boldsymbol{\eta}) + 1$ and $\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[g(\boldsymbol{\eta}) + 1] = \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[g(\boldsymbol{\eta})] + 1 < \infty$. Then, by dominated convergence theorem,

$$\begin{aligned}\lim_{m\to\infty}\frac{\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\theta+h_m,\boldsymbol{\eta})] - \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\theta,\boldsymbol{\eta})]}{h_m} &= \lim_{m\to\infty}\frac{\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\theta+h_m,\boldsymbol{\eta}) - f(\theta,\boldsymbol{\eta})]}{h_m} \\ &= \lim_{m\to\infty}\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f_{\theta,m}(\boldsymbol{\eta})]\end{aligned}$$

$$= \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}\left[\frac{\partial}{\partial\theta}f(\theta,\boldsymbol{\eta})\right].$$

Since our choice of $\{h_m\}_{m\in\mathbb{N}}$ is arbitrary, $\frac{d}{d\theta}\mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}[f(\theta,\boldsymbol{\eta})] = \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{P}}\left[\frac{\partial}{\partial\theta}f(\theta,\boldsymbol{\eta})\right]$. □

By applying Lemma D.1, we can obtain the following lemma which makes us possible to investigate stationary points and strong convexity constants of expected losses in the proof of our main theorems.

Lemma D.2. *For any vector $\boldsymbol{u}\in\mathbb{R}^k$, positive definite matrix $\boldsymbol{M}\in\mathbb{R}^{k\times k}$, functions $f:[0,1]\to[-1,1], g:[0,1]\to[0,1]$ and probability distribution $\mathcal{P}$ with support contained in $[0,1]$, we have*

$$\nabla_{\boldsymbol{w}}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})]\right] = \mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l'(\boldsymbol{w}^\top\mathbf{x})\mathbf{x}]\right],$$

and

$$\nabla^2_{\boldsymbol{w}}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})]\right] = \mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l''(\boldsymbol{w}^\top\mathbf{x})\mathbf{x}\mathbf{x}^\top]\right].$$

Proof of Lemma D.2. For each $i\in[k]$ and $\boldsymbol{x}\in\mathbb{R}^k$, we have

$$\left|\frac{\partial}{\partial(\boldsymbol{w})_i}l(\boldsymbol{w}^\top\boldsymbol{x})\right| = |l'(\boldsymbol{w}^\top\boldsymbol{x})(\boldsymbol{x})_i| = \left|\frac{(\boldsymbol{x})_i}{1+e^{\boldsymbol{w}^\top\boldsymbol{x}}}\right| \le |(\boldsymbol{x})_i|.$$

Since $\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[|(\mathbf{x})_i|]<\infty$ for each $\eta\in[0,1]$, by Lemma D.1,

$$\frac{\partial}{\partial(\boldsymbol{w})_i}\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l'(\boldsymbol{w}^\top\mathbf{x})(\mathbf{x})_i].$$

Also,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[|(\mathbf{x})_i|] &\le \mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[(\mathbf{x})_i^2]^{1/2}\\ &= \left(g(\eta)(\boldsymbol{M})_i + f(\eta)^2(\boldsymbol{u})_i^2\right)^{1/2}\\ &\le \left((\boldsymbol{M})_i + (\boldsymbol{u})_i^2\right)^{1/2}.\end{aligned}$$

Hence, $\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[|(\mathbf{x})_i|]\right]<\infty$ because the inner expectation is uniformly bounded for all $\eta\in[0,1]$. Applying Lemma D.1 again, we have

$$\frac{\partial}{\partial(\boldsymbol{w})_i}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})]\right] = \mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{\mu},g(\eta)\kappa^{-1}\boldsymbol{\Sigma})}[l'(\boldsymbol{w}^\top\mathbf{x})(\mathbf{x})_i]\right],$$

and we conclude

$$\nabla_{\boldsymbol{w}}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\kappa^{-1}\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})]\right] = \mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\kappa^{-1}\boldsymbol{M})}[l'(\boldsymbol{w}^\top\mathbf{x})\mathbf{x}]\right].$$

For each $i,j\in[k]$ and $\boldsymbol{x}\in\mathbb{R}^k$,

$$\left|\frac{\partial^2}{\partial(\boldsymbol{w})_j\partial(\boldsymbol{w})_i}l(\boldsymbol{w}^\top\boldsymbol{x})\right| = \left|l''(\boldsymbol{w}^\top\boldsymbol{x})(\boldsymbol{x})_i(\boldsymbol{x})_j\right| \le \frac{|(\boldsymbol{x})_i(\boldsymbol{x})_j|}{4}.$$

Since $\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[|(\mathbf{x})_i(\mathbf{x})_j|]<\infty$ for each $\eta\in[0,1]$, by Lemma D.1, we have

$$\begin{aligned}\frac{\partial^2}{\partial(\boldsymbol{w})_j\partial(\boldsymbol{w})_i}\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})] &= \frac{\partial}{\partial(\boldsymbol{w})_j}\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}\left[\frac{\partial}{\partial(\boldsymbol{w})_i}l(\boldsymbol{w}^\top\mathbf{x})\right]\\ &= \mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}\left[\frac{\partial^2}{\partial(\boldsymbol{w})_j\partial(\boldsymbol{w})_i}l(\boldsymbol{w}^\top\mathbf{x})\right]\\ &= \mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l''(\boldsymbol{w}^\top\mathbf{x})(\mathbf{x})_i(\mathbf{x})_j].\end{aligned}$$

Also, $\left|\frac{\partial^2}{\partial(\boldsymbol{w})_j\partial(\boldsymbol{w})_i}\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})]\right| \leq \frac{1}{4}\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[|(\mathbf{x})_i(\mathbf{x})_j|]$ and

$$\begin{aligned}\frac{1}{4}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[|(\mathbf{x})_i(\mathbf{x})_j|]\right] &\leq \frac{1}{8}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[(\mathbf{x})_i^2+(\mathbf{x})_j^2]\right]\\ &= \frac{1}{8}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[g(\eta)((\boldsymbol{M})_i+(\boldsymbol{M})_j)+f(\eta)^2((\boldsymbol{u})_i^2+(\boldsymbol{u})_j^2)]\right]\\ &\leq \frac{1}{8}(\mathrm{Tr}(\boldsymbol{M})+\|\boldsymbol{u}\|^2)<\infty.\end{aligned}$$

By applying Lemma D.1 again, for each $i,j\in[k]$,

$$\frac{\partial^2}{\partial(\boldsymbol{w})_j\partial(\boldsymbol{w})_i}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})]\right]=\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}\left[\frac{\partial^2}{\partial(\boldsymbol{w})_j\partial(\boldsymbol{w})_i}l(\boldsymbol{w}^\top\mathbf{x})\right]\right],$$

and we conclude

$$\nabla^2_{\boldsymbol{w}}\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l(\boldsymbol{w}^\top\mathbf{x})]\right]=\mathbb{E}_{\eta\sim\mathcal{P}}\left[\mathbb{E}_{\mathbf{x}\sim N(f(\eta)\boldsymbol{u},g(\eta)\boldsymbol{M})}[l''(\boldsymbol{w}^\top\mathbf{x})\mathbf{x}\mathbf{x}^\top]\right].$$

□

D.3. Inequalities

Let us introduce some inequalities which are used in the proof of main theorems.

The following lemma provides us computable lower bound on the strong convexity constant.

Lemma D.3. *For any $z\in\mathbb{R}$,*

$$l''(z)=\frac{e^z}{(e^z+1)^2}\geq\frac{1}{4}e^{-z^2/2}.$$

Proof of Lemma D.3. Define a function $f:\mathbb{R}\to\mathbb{R}$ as $f(z)=\frac{1}{2}z^2-2\log(1+e^z)$ for each $z\in\mathbb{R}$. Then, we have

$$f'(z)=z-\frac{2e^z}{1+e^z},\ f''(z)=1-\frac{2e^z}{(e^z+1)^2}>0,$$

for each $z\in\mathbb{R}$. Hence, f is a convex function and $z\mapsto -z-2\log 2$ is a tangent line of the graph of $f(z)$ at $z=0$. Therefore, for each $z\in\mathbb{R}$, we have

$$f(z)=\frac{1}{2}z^2-2\log(1+e^z)\geq -z-2\log 2,$$

which implies

$$\frac{1}{2}z^2\geq -z+2\log(1+e^z)-2\log 2=-\log\left(\frac{4e^z}{(e^z+1)^2}\right).$$

Thus, we conclude $l''(z)=\frac{e^z}{(e^z+1)^2}\geq\frac{1}{4}e^{-z^2/2}$ for each $z\in\mathbb{R}$. □

When we get lower bounds on $\|\boldsymbol{w}^*(\kappa)\|$ and $\|\boldsymbol{w}^*_{\mathrm{mask}}(n,\kappa)\|$, the following lemma is useful.

Lemma D.4. *For any $z\in\mathbb{R}$,*

$$\frac{z}{1+e^z}\geq\frac{1}{2}z-\frac{1}{4}z^2.$$

Proof of Lemma D.4. Define a function $f:\mathbb{R}\to\mathbb{R}$ as $f(z)=\frac{1}{1+e^z}$ for each $z\in\mathbb{R}$. We have

$$f'(z)=-\frac{e^z}{(1+e^z)^2},\ f''(z)=\frac{e^z(e^z-1)}{(1+e^z)^3}.$$

Therefore, f is convex on $[0, \infty)$ and concave on $(-\infty, 0]$. Since $z \mapsto -\frac{1}{4}z + \frac{1}{2}$ is tangent line of $f(z)$ at $z = 0$, we have

$$f(z) = \frac{1}{1+e^z} \geq -\frac{1}{4}z + \frac{1}{2},$$

for any $z \geq 0$ and

$$f(z) = \frac{1}{1+e^z} \leq -\frac{1}{4}z + \frac{1}{2},$$

for any $z \leq 0$. By multiplying z to the inequalities above, we have our conclusion. □

Together with Lemma D.3, the following lemma provide us lower bounds on strong convexity constant of expected loss $\mathbb{E}_\kappa[\mathcal{L}_S(\boldsymbol{w})]$, $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mix}}(\boldsymbol{w})]$ and $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$.

Lemma D.5. *Let $\boldsymbol{u} \in \mathbb{R}^k$ be a vector and $\boldsymbol{M} \in \mathbb{R}^{k\times k}$ be a positive definite matrix. For any vector $\boldsymbol{w} \in \mathbb{R}^k$ and unit vector $\boldsymbol{v} \in \mathbb{R}^k$, we have*

$$\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{u},\boldsymbol{M})}\left[(\boldsymbol{v}^\top\mathbf{x})^2 e^{-(\boldsymbol{w}^\top\mathbf{x})^2/2}\right] \geq \max\left\{(\|\boldsymbol{M}^{-1}\| + \|\boldsymbol{w}\|^2)^{-1}, \frac{1}{2}(\boldsymbol{v}^\top\boldsymbol{u})^2 - \|\boldsymbol{M}\|^2\|\boldsymbol{w}\|^4\|\boldsymbol{u}\|^2\right\}$$
$$\cdot \left(\|\boldsymbol{M}\|(\|\boldsymbol{M}^{-1}\| + \|\boldsymbol{w}\|^2)\right)^{-k/2} \exp\left(-\|\boldsymbol{w}\|^2\|\boldsymbol{M}^{-1}\|\|\boldsymbol{M}\|\|\boldsymbol{u}\|^2\right).$$

Proof of Lemma D.5. By changing expectation to integral form, we have

$$\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{u},\boldsymbol{M})}\left[(\boldsymbol{v}^\top\mathbf{x})^2 e^{-(\boldsymbol{w}^\top\mathbf{x})^2/2}\right] = \int_{\mathbb{R}^k}(2\pi)^{-k/2}\underbrace{\det(\boldsymbol{M})^{-1/2}}_{\text{(a)}}(\boldsymbol{v}^\top\boldsymbol{x})^2\underbrace{\exp\left(-\frac{1}{2}\left(\|\boldsymbol{x}-\boldsymbol{u}\|_{\boldsymbol{M}^{-1}}^2 + (\boldsymbol{w}^\top\boldsymbol{x})^2\right)\right)}_{\text{(b)}}d\boldsymbol{x}.$$

We can rewrite the term (a) as

$$\begin{aligned}\det(\boldsymbol{M})^{-1/2} &= \det(\boldsymbol{M}(\boldsymbol{M}^{-1} + \boldsymbol{w}\boldsymbol{w}^\top)(\boldsymbol{M}^{-1} + \boldsymbol{w}\boldsymbol{w}^\top)^{-1})^{-1/2}\\ &= \det((\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)(\boldsymbol{M}^{-1} + \boldsymbol{w}\boldsymbol{w}^\top)^{-1})^{-1/2}\\ &= \det(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1/2}\det(\boldsymbol{M}^{-1} + \boldsymbol{w}\boldsymbol{w}^\top)^{1/2},\end{aligned}$$

and the term (b) as

$$\begin{aligned}&\exp\left(-\frac{1}{2}\left(\|\boldsymbol{x}-\boldsymbol{u}\|_{\boldsymbol{M}^{-1}}^2 + (\boldsymbol{w}^\top\boldsymbol{x})^2\right)\right)\\ &= \exp\left(-\frac{1}{2}\left(\|\boldsymbol{x}\|_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top} - 2\boldsymbol{u}^\top\boldsymbol{M}^{-1}\boldsymbol{x} + \|\boldsymbol{u}\|_{\boldsymbol{M}^{-1}}\right)\right)\\ &= \exp\left(-\frac{1}{2}\left(\|\boldsymbol{x}\|_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top}^2 - 2\boldsymbol{u}^\top\boldsymbol{M}^{-1}(\boldsymbol{M}^{-1} + \boldsymbol{w}\boldsymbol{w}^\top)^{-1}(\boldsymbol{M}^{-1} + \boldsymbol{w}\boldsymbol{w}^\top)\boldsymbol{x} + \|\boldsymbol{u}\|_{\boldsymbol{M}^{-1}}^2\right)\right)\\ &= \exp\left(-\frac{1}{2}\left(\|\boldsymbol{x}\|_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top}^2 - 2\boldsymbol{u}^\top(\boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})^{-1}(\boldsymbol{M}^{-1} + \boldsymbol{w}\boldsymbol{w}^\top)\boldsymbol{x} + \|\boldsymbol{u}\|_{\boldsymbol{M}^{-1}}^2\right)\right)\\ &= \exp\left(-\frac{1}{2}\left(\left\|\boldsymbol{x} - \left(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top\right)^{-1}\boldsymbol{u}\right\|_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top}^2 + \|\boldsymbol{u}\|_{\boldsymbol{M}^{-1}}^2 - \left\|\left(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top\right)^{-1}\boldsymbol{u}\right\|_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top}^2\right)\right)\\ &= \exp\left(-\frac{1}{2}\left\|\boldsymbol{x} - \left(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top\right)^{-1}u\right\|_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top}^2\right)\cdot\exp\left(-\frac{1}{2}\underbrace{\left(\|\boldsymbol{u}\|_{\boldsymbol{M}^{-1}}^2 - \left\|(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u}\right\|_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top}^2\right)}_{\text{(c)}}\right).\end{aligned}$$

Also, the term (c) can be simplified as

$$\begin{aligned}&\|\boldsymbol{u}\|_{\boldsymbol{M}^{-1}}^2 - \|(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u}\|_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top}^2\\ &= \boldsymbol{u}^\top\boldsymbol{M}^{-1}\boldsymbol{u} - ((\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u})^\top(\boldsymbol{M}^{-1} + \boldsymbol{w}\boldsymbol{w}^\top)((\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u})\\ &= \boldsymbol{u}^\top\boldsymbol{M}^{-1}\boldsymbol{u} - (\boldsymbol{u}^\top(\boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})^{-1})(\boldsymbol{M}^{-1}(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top))((\boldsymbol{I} + \boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u})\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{u}^\top \boldsymbol{M}^{-1}\boldsymbol{u} - \boldsymbol{u}^\top(\boldsymbol{I}+\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})^{-1}\boldsymbol{M}^{-1}\boldsymbol{u} \\
&= \boldsymbol{u}^\top(\boldsymbol{I}+\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})(\boldsymbol{I}+\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})^{-1}\boldsymbol{M}^{-1}\boldsymbol{u} - \boldsymbol{u}^\top(\boldsymbol{I}+\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})^{-1}\boldsymbol{M}^{-1}\boldsymbol{u} \\
&= \boldsymbol{u}^\top(\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})(\boldsymbol{I}+\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})^{-1}\boldsymbol{M}^{-1}\boldsymbol{u} \\
&= \boldsymbol{u}^\top(\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})(\boldsymbol{M}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})^{-1}\boldsymbol{u} \\
&= \boldsymbol{u}^\top(\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{M})((\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)\boldsymbol{M})^{-1}\boldsymbol{u} \\
&= \boldsymbol{u}^\top\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim N(\boldsymbol{u},\boldsymbol{M})}\left[(\boldsymbol{v}^\top\mathbf{x})^2 e^{-(\boldsymbol{w}^\top\mathbf{x})^2/2}\right] \\
&= \underbrace{\int_{\mathbb{R}^d}(2\pi)^{-d/2}\det(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{1/2}(\boldsymbol{v}^\top\boldsymbol{x})^2\exp\left(-\frac{1}{2}\left(\|\boldsymbol{x}-(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u}\right)\|^2_{\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top}\right)d\boldsymbol{x}}_{\text{(d)}} \\
&\quad\cdot\det(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u}\right).
\end{aligned}
$$

By changing integral into expectation, we have

$$
\begin{aligned}
\text{(d)} &= \mathbb{E}_{\mathbf{x}\sim N((\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u},(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{-1})}\left[\left(\boldsymbol{v}^\top\mathbf{x}\right)^2\right] \\
&= \mathbb{E}_{\mathrm{z}\sim N(\boldsymbol{v}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u},\boldsymbol{v}^\top(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{v})}\left[\mathrm{z}^2\right].
\end{aligned}
$$

From $\mathbb{E}_{\mathrm{z}\sim N(m,\sigma^2)}[\mathrm{z}^2] = m^2+\sigma^2$ for each $m\in\mathbb{R}, \sigma>0$, note that

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{z}\sim N(\boldsymbol{v}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u},\boldsymbol{v}^\top(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{v})}[\mathrm{z}^2] \\
&\geq \boldsymbol{v}^\top(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{v} \geq \|\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top\|^{-1} \\
&\geq \left(\|\boldsymbol{M}^{-1}\|+\|\boldsymbol{w}\|^2\right)^{-1},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{z}\sim N(\boldsymbol{v}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u},\boldsymbol{v}^\top(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{v})}[\mathrm{z}^2] \\
&\geq (\boldsymbol{v}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u})^2 = (\boldsymbol{v}^\top\boldsymbol{u}-\boldsymbol{v}^\top\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u})^2 \\
&\geq \frac{1}{2}(\boldsymbol{v}^\top\boldsymbol{u})^2 - (\boldsymbol{v}^\top\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u})^2 \\
&\geq \frac{1}{2}(\boldsymbol{v}^\top\boldsymbol{u})^2 - \|\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top\|^2\|(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\|\|\boldsymbol{u}\|^2 \\
&\geq \frac{1}{2}(\boldsymbol{v}^\top\boldsymbol{u})^2 - \|\boldsymbol{M}\|^2\|\boldsymbol{w}\|^4\|\boldsymbol{u}\|^2.
\end{aligned}
$$

Also,

$$
\det(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top) = \det(\boldsymbol{M})\det(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top) \leq (\|\boldsymbol{M}\|\|\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top\|)^k \leq (\|\boldsymbol{M}\|(\|\boldsymbol{M}^{-1}\|+\|\boldsymbol{w}\|^2))^k,
$$

and

$$
\begin{aligned}
\boldsymbol{u}^\top\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{I}+\boldsymbol{M}\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{u} &= \boldsymbol{u}^\top\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{M}^{-1}\boldsymbol{u} \\
&\leq \|\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\boldsymbol{M}^{-1}\|\|\boldsymbol{u}\|^2 \\
&\leq \|\boldsymbol{w}\boldsymbol{w}^\top\|\|(\boldsymbol{M}^{-1}+\boldsymbol{w}\boldsymbol{w}^\top)^{-1}\|\|\boldsymbol{M}^{-1}\|\|\boldsymbol{u}\|^2 \\
&\leq \|\boldsymbol{w}\|^2\|\boldsymbol{M}\|\|\boldsymbol{M}^{-1}\|\|\boldsymbol{u}\|^2.
\end{aligned}
$$

Hence, we have our conclusion. □

The following lemma makes us obtain M value in Lemma 3.5 and Lemma 4.3 when we prove the sample complexity results.

Lemma D.6. *Let $\boldsymbol{M} \in \mathbb{R}^{k\times k}$ be a positive definite matrix. Then, we have*

$$\mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\boldsymbol{M})}\left[e^{\|\mathbf{z}\|}\right] \leq e^{4\|\boldsymbol{M}\|} + 2^{k/2}.$$

Proof of Lemma D.6. We have

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\boldsymbol{M})}\left[e^{\|\mathbf{z}\|}\right] &= \mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\boldsymbol{M})}\left[e^{\|\mathbf{z}\|}\mathbf{1}_{\|\mathbf{z}\|\leq 4\|\boldsymbol{M}\|}\right] + \mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\boldsymbol{M})}\left[e^{\|\mathbf{z}\|}\mathbf{1}_{\|\mathbf{z}\|> 4\|\boldsymbol{M}\|}\right] \\ &\leq e^{4\|\boldsymbol{M}\|} + \mathbb{E}_{Z\sim N(\mathbf{0},\boldsymbol{M})}\left[e^{\frac{1}{4}\|\boldsymbol{M}\|^{-1}\|\mathbf{z}\|^2}\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim N(\mathbf{0},\boldsymbol{M})}\left[e^{\frac{1}{4}\|\boldsymbol{M}\|^{-1}\|\mathbf{z}\|^2}\right] &= \int_{\mathbb{R}^k}(2\pi)^{-k/2}\det(\boldsymbol{M})^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\left(\boldsymbol{M}^{-1}-\frac{1}{2}\|\boldsymbol{M}\|^{-1}\boldsymbol{I}\right)\boldsymbol{x}\right)d\boldsymbol{x} \\ &= \det\left(\boldsymbol{M}^{-1}-\frac{1}{2}\|\boldsymbol{M}\|^{-1}\boldsymbol{I}\right)^{-1/2}\det(\boldsymbol{M})^{-1/2} = \det\left(\boldsymbol{I}-\frac{1}{2}\|M\|^{-1}\right)^{-1/2} \\ &\leq \left\|\left(\boldsymbol{I}-\frac{1}{2}\|\boldsymbol{M}\|^{-1}\boldsymbol{M}\right)^{-1}\right\|^{k/2} \leq 2^{k/2}.\end{aligned}$$

Hence, we have our conclusion. □

Lastly, we introduce the lemma used in showing uniform convergence of $\mathbb{E}_\kappa[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ to $\mathbb{E}_\infty[\mathcal{L}_S^{\text{mask}}(\boldsymbol{w})]$ as $\kappa\to\infty$.

Lemma D.7. *For each $m\in\mathbb{R}$ and $\sigma>0$,*

$$0 \leq \mathbb{E}_{X\sim N(m,\sigma^2)}[l(X)] - l(m) \leq \sigma.$$

Proof of Lemma D.7. Since $l(\cdot)$ is convex, by Jensen's inequality, we have

$$\mathbb{E}_{X\sim N(m,\sigma^2)}[l(X)] \geq l(\mathbb{E}_{X\sim N(m,\sigma^2)}[X]) = l(m).$$

Also, we have

$$l(m) - \mathbb{E}_{X\sim N(m,\sigma^2)}[l(X)] = \mathbb{E}_{X\sim N(m,\sigma^2)}[l(m)-l(X)] \geq \mathbb{E}_{X\sim N(m,\sigma^2)}[l'(X)(m-X)] \geq -\mathbb{E}[|X-m|],$$

where the last inequality used $|l'(z)|\leq 1$ for all $z\in\mathbb{R}$. By Cauchy-Schwartz inequality, $\mathbb{E}_{X\sim N(m,\sigma^2)}[|X-m|] \leq \mathbb{E}_{X\sim N(m,\sigma^2)}[(X-m)^2]^{1/2} = \sigma$. Thus, we have our conclusion. □

D.4. Concentration Bounds

We introduce concentration bounds for i.i.d. random variables which we use in the proof of Lemma 3.5.

Lemma D.8. *Let $X_1,\dots,X_N \overset{\text{i.i.d.}}{\sim} \mathcal{P}$ where $\mathcal{P}$ is a probability distribution on $\mathbb{R}$. Suppose $\mathbb{E}_{X\sim\mathcal{P}}\left[e^{|X-\mathbb{E}_{X\sim\mathcal{P}}[X]|}\right]\leq M$ for some constant $M>0$. Then, for any $0<\epsilon<1$,*

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^N X_i - \mathbb{E}_{X\sim\mathcal{P}}[X] > \epsilon\right] \leq \exp\left(-\frac{C\epsilon^2 N}{M}\right),$$

and

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^N X_i - \mathbb{E}_{X\sim\mathcal{P}}[X] < -\epsilon\right] \leq \exp\left(-\frac{C\epsilon^2 N}{M}\right),$$

where $C>0$ is a universal constant.

Proof of Lemma D.8. From our definition of M, we have $M \geq 1$. Choose $t = \frac{\epsilon}{16M}$, then we have $0 < t < \frac{1}{2}$ since $0 < \epsilon < 1$ and $M \geq 1$. From Chernoff bound, we have

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}_{X\sim\mathcal{P}}[X] > \epsilon\right] \leq \mathbb{E}\left[e^{t\left(\sum_{i=1}^{N} X_i - \mathbb{E}_{X\sim\mathcal{P}}[X]\right)}\right] e^{-tN\epsilon},$$

and

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}_{X\sim\mathcal{P}}[X] < -\epsilon\right] \leq \mathbb{E}\left[e^{t\left(\sum_{i=1}^{N} \mathbb{E}_{X\sim\mathcal{P}}[X] - X_i\right)}\right] e^{-tN\epsilon}.$$

Since $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} \mathcal{P}$, we have

$$\mathbb{E}\left[e^{t\left(\sum_{i=1}^{N} X_i - \mathbb{E}_{X\sim\mathcal{P}}[X]\right)}\right] e^{-tN\epsilon} = \mathbb{E}\left[e^{t\sum_{i=1}^{N}(X_i - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon)}\right] = \mathbb{E}_{X\sim\mathcal{P}}\left[e^{t(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon)}\right]^N,$$

and

$$\mathbb{E}\left[e^{t\left(\sum_{i=1}^{N} \mathbb{E}_{X\sim\mathcal{P}}[X] - X_i\right)}\right] e^{-tN\epsilon} = \mathbb{E}\left[e^{t\sum_{i=1}^{N}(\mathbb{E}_{X\sim\mathcal{P}}[X] - X_i - \epsilon)}\right] = \mathbb{E}_{X\sim\mathcal{P}}\left[e^{t(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon)}\right]^N.$$

For each $x \in \mathbb{R}$, $e^x \leq 1 + x + \frac{1}{2}x^2 e^{|x|}$. Thus,

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{P}}\left[e^{t(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon)}\right] \\
&\leq \mathbb{E}_{X\sim\mathcal{P}}\left[1 + t\left(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon\right) + \frac{t^2}{2}\left(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon\right)^2 e^{|t(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon)|}\right] \\
&= 1 - \epsilon t + \frac{t^2}{2}\mathbb{E}_{X\sim\mathcal{P}}\left[\left(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon\right)^2 e^{|t(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon)|}\right],
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{P}}\left[e^{t(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon)}\right] \\
&\leq \mathbb{E}_{X\sim\mathcal{P}}\left[1 + t\left(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon\right) + \frac{t^2}{2}\left(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon\right)^2 e^{|t(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon)|}\right] \\
&= 1 - \epsilon t + \frac{t^2}{2}\mathbb{E}_{X\sim\mathcal{P}}\left[\left(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon\right)^2 e^{|t(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon)|}\right].
\end{aligned}$$

Also, since $x^2 \leq 4e^{|x|/2}$ for each $x \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{P}}\left[\left(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon\right)^2 e^{|t(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon)|}\right] \\
&\leq 4\mathbb{E}_{X\sim\mathcal{P}}\left[e^{\left(t + \frac{1}{2}\right)|X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon|}\right] \leq 4\mathbb{E}_{X\sim\mathcal{P}}\left[e^{|X - \mathbb{E}_{X\sim\mathcal{P}}[X]| + \epsilon}\right] \\
&\leq 16M,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{P}}\left[\left(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon\right)^2 e^{|t(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon)|}\right] \\
&\leq 4\mathbb{E}_{X\sim\mathcal{P}}\left[e^{\left(t + \frac{1}{2}\right)|\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon|}\right] \leq 4\mathbb{E}_{X\sim\mathcal{P}}\left[e^{|X - \mathbb{E}_{X\sim\mathcal{P}}[X]| + \epsilon}\right] \\
&\leq 16M.
\end{aligned}$$

Therefore, by substituting $t = \frac{\epsilon}{16M}$ we get

$$\mathbb{E}_{X\sim\mathcal{P}}\left[e^{t(X - \mathbb{E}_{X\sim\mathcal{P}}[X] - \epsilon)}\right] \leq 1 - \epsilon t + 8Mt^2 = 1 - \frac{\epsilon^2}{32M} \leq \exp\left(-\frac{\epsilon^2}{32M}\right),$$

and

$$\mathbb{E}_{X\sim\mathcal{P}}\left[e^{t(\mathbb{E}_{X\sim\mathcal{P}}[X] - X - \epsilon)}\right] \leq 1 - \epsilon t + 8Mt^2 = 1 - \frac{\epsilon^2}{32M} \leq \exp\left(-\frac{\epsilon^2}{32M}\right).$$

We conclude

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}_{X\sim\mathcal{P}}[X] > \epsilon\right] \leq \exp\left(-\frac{C\epsilon^2 N}{M}\right),$$

and

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}_{X\sim\mathcal{P}}[X] < -\epsilon\right] \leq \exp\left(-\frac{C\epsilon^2 N}{M}\right).$$

where $C = \frac{1}{32}$. □

We extend Lemma D.8 to non i.i.d. setting random variables with special types of dependency using the following two technical lemmas.

Lemma D.9. *There are disjoint sets $P_1, \ldots, P_m \subset [N] \times [N]$ such that $[N] \times [N] = \bigcup_{i=1}^{m} P_i \cup \{(1,1), \ldots, (N,N)\}$ and*

$$m = \begin{cases} 2N & \text{if } N \text{ is odd,} \\ 2(N-1) & \text{if } N \text{ is even,} \end{cases} \quad |P_k| = \begin{cases} \frac{N-1}{2} & \text{if } N \text{ is odd,} \\ \frac{N}{2} & \text{if } N \text{ is even,} \end{cases} \text{ for all } k \in [m].$$

That is, $P_1, \ldots, P_m$ and $\{(1,1), \ldots, (N,N)\}$ together form a partition of $[N] \times [N]$. Furthermore, for each $k \in [m]$ and for any distinct $(i_1, j_1), (i_2, j_2) \in P_k$, $\{i_1\} \cup \{j_2\}$ and $\{i_2\} \cup \{j_2\}$ are disjoint.

Proof of Lemma D.9.
Case 1: N is odd
For each $k \in [N]$, define

$$P_{2k-1} = \{(i,j) \mid i + j \equiv k \pmod{N}, i < j\},$$
$$P_{2k} = \{(i,j) \mid i + j \equiv k \pmod{N}, i > j\}.$$

It can be easily checked that these $P_1, \ldots, P_{2N}$ are what we desired.
Case 2: N is even
For each $k \in [N-1]$, there is unique $i_k \in [N-1]$ such that $2i_k \equiv k \pmod{(N-1)}$. For each $k \in [N-1]$, define

$$P_{2k-1} = \{(i,j) \mid i + j \equiv k \pmod{(N-1)}, i < j\} \cup \{(i_k, N)\},$$
$$P_{2k} = \{(i,j) \mid i + j \equiv k \pmod{(N-1)}, i > j\} \cup \{(N, i_k)\},$$

and it can be easily checked that these $P_1, \ldots, P_{2(N-1)}$ are what we desired. □

This is a generalized version of Cauchy-Schwartz inequality.

Lemma D.10. *Suppose $X_1, \ldots, X_k$ are nonnegative random variables. Then,*

$$\mathbb{E}\left[\prod_{i=1}^{k} X_i\right] \leq \left(\prod_{i=1}^{k} \mathbb{E}\left[X_i^k\right]\right)^{\frac{1}{k}}.$$

Proof of Lemma D.10. We prove this by using induction on k. Note that the case $k = 1$ is trivial and $k = 2$ is Cauchy-Schwartz inequality. Suppose Lemma D.10 holds for $k = m$. Let $X_1, \ldots, X_{m+1}$ be nonnegative random variables. By Hölder inequality,

$$\mathbb{E}[X_1 X_2 \cdots X_{m+1}] \leq \mathbb{E}\left[(X_1 \cdots X_m)^{\frac{m+1}{m}}\right]^{\frac{m}{m+1}} \mathbb{E}\left[X_{m+1}^{m+1}\right]^{\frac{1}{m+1}}.$$

From the induction hypothesis, we have

$$\mathbb{E}\left[(X_1 \cdots X_m)^{\frac{m+1}{m}}\right] \leq \left(\prod_{i=1}^{m} \mathbb{E}\left[\left(X_i^{\frac{m+1}{m}}\right)^m\right]\right)^{\frac{1}{m}} = \left(\prod_{i=1}^{m} \mathbb{E}\left[X_i^{m+1}\right]\right)^{\frac{1}{m}}.$$

Therefore, we have

$$\mathbb{E}[X_1 X_2 \cdots X_{m+1}] \leq \left(\prod_{i=1}^{m+1} \mathbb{E}\left[X_i^{m+1}\right] \right)^{\frac{1}{m+1}}.$$

By the principle of mathematical induction, our conclusion holds for all $k \in \mathbb{N}$. □

Using the two lemmas above, we prove the following lemma which is used in the proof of Lemma 4.3.

Lemma D.11. *Let $\{X_{i,j}\}_{i,j\in[N]}$ be real-valued random variables satisfy followings.*

- $\mathbb{E}\left[e^{|X_{i,j}-\mathbb{E}[X_{i,j}]|}\right] \leq M$ *for some* $M > 0$.
- *If $\{i_1\} \cup \{j_1\}$ and $\{i_2\} \cup \{j_1\}$ are disjoint for $i_1, i_2, j_1, j_2 \in [n]$, then X_{i_1,j_1} and X_{i_2,j_2} are independent.*

Then, for any $0 < \epsilon < 1$,

$$\mathbb{P}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}(X_{i,j} - \mathbb{E}[X_{i,j}]) > \epsilon\right] \leq \exp\left(-\frac{C'\epsilon^2 N}{M}\right),$$

and

$$\mathbb{P}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}(X_{i,j} - \mathbb{E}[X_{i,j}]) < -\epsilon\right] \leq \exp\left(-\frac{C'\epsilon^2 N}{M}\right),$$

where $C' > 0$ is a universal constant.

Proof of Lemma D.11. From Chernoff bound, we have

$$\mathbb{P}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}(X_{i,j} - \mathbb{E}[X_{i,j}]) > \epsilon\right] \leq \mathbb{E}\left[e^{t\sum_{i,j=1}^{N}(X_{i,j}-\mathbb{E}[X_{i,j}]-\epsilon)}\right],$$

and

$$\mathbb{P}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}(X_{i,j} - \mathbb{E}[X_{i,j}]) < -\epsilon\right] \leq \mathbb{E}\left[e^{t\sum_{i,j=1}^{N}(\mathbb{E}[X_{i,j}]-X_{i,j}-\epsilon)}\right],$$

for any $t > 0$. We would like to get an upper bound on $\mathbb{E}\left[e^{t\sum_{i,j=1}^{N}(X_{i,j}-\mathbb{E}[X_{i,j}]-\epsilon)}\right]$ and $\mathbb{E}\left[e^{t\sum_{i,j=1}^{N}(\mathbb{E}[X_{i,j}]-X_{i,j}-\epsilon)}\right]$, but the problem is that $\{X_{i,j}\}_{i,j\in[N]}$ are not independent of one another. We overcome this obstacle by applying the two lemmas above.

For our N, consider m and $P_1, \ldots, P_m \subset [N] \times [N]$ that we can obtain from Lemma D.9. From our definition of M, we have $M \geq 1$ and then let $t = \frac{\epsilon}{16(m+1)M}$. Since we have $0 < \epsilon < 1$ and $M \geq 1$, we can check that $0 < (m+1)t < \frac{1}{2}$.

By Lemma D.10, we have

$$\begin{aligned}
&\mathbb{E}\left[e^{t\sum_{i,j=1}^{N}(X_{i,j}-\mathbb{E}[X_{i,j}]-\epsilon)}\right] \\
&= \mathbb{E}\left[e^{t\sum_{i=1}^{N}(X_{i,i}-\mathbb{E}[X_{i,i}]-\epsilon)} \prod_{k=1}^{m} e^{t\sum_{(i,j)\in P_k}(X_{i,j}-\mathbb{E}[X_{i,j}]-\epsilon)}\right] \\
&\leq \left(\mathbb{E}\left[e^{(m+1)t\sum_{i=1}^{N}(X_{i,i}-\mathbb{E}[X_{i,i}]-\epsilon)}\right] \prod_{k=1}^{m} \mathbb{E}\left[e^{(m+1)t\sum_{(i,j)\in P_k}(X_{i,j}-\mathbb{E}[X_{i,j}]-\epsilon))}\right]\right)^{\frac{1}{m+1}} \\
&= \left(\prod_{i=1}^{N} \mathbb{E}\left[e^{(m+1)t(X_{i,i}-\mathbb{E}[X_{i,i}]-\epsilon)}\right] \prod_{k=1}^{m}\left(\prod_{(i,j)\in P_k} \mathbb{E}\left[e^{(m+1)t(X_{i,j}-\mathbb{E}[X_{i,j}]-\epsilon)}\right]\right)\right)^{\frac{1}{m+1}}
\end{aligned}$$

$$= \left(\prod_{i,j=1}^{N} \mathbb{E}\left[e^{(m+1)t(X_{i,j} - \mathbb{E}[X_{i,j}] - \epsilon)} \right] \right)^{\frac{1}{m+1}},$$

and

$$\begin{aligned}
&\mathbb{E}\left[e^{t\sum_{i,j=1}^{N}(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon)} \right] \\
&= \mathbb{E}\left[e^{t\sum_{i=1}^{N}(\mathbb{E}[X_{i,i}] - X_{i,i} - \epsilon)} \prod_{k=1}^{m} e^{t\sum_{(i,j)\in P_k}(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon)} \right] \\
&\leq \left(\mathbb{E}\left[e^{(m+1)t\sum_{i=1}^{N}(\mathbb{E}[X_{i,i}] - X_{i,i} - \epsilon)} \right] \prod_{k=1}^{m} \mathbb{E}\left[e^{(m+1)t\sum_{(i,j)\in P_k}(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon))} \right] \right)^{\frac{1}{m+1}} \\
&= \left(\prod_{i=1}^{N} \mathbb{E}\left[e^{(m+1)t(\mathbb{E}[X_{i,i}] - X_{i,i} - \epsilon)} \right] \prod_{k=1}^{m} \left(\prod_{(i,j)\in P_k} \mathbb{E}\left[e^{(m+1)t(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon)} \right] \right) \right)^{\frac{1}{m+1}} \\
&= \left(\prod_{i,j=1}^{N} \mathbb{E}\left[e^{(m+1)t(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon)} \right] \right)^{\frac{1}{m+1}}.
\end{aligned}$$

For each $x \in \mathbb{R}, e^x \leq 1 + x + \frac{1}{2}x^2 e^{|x|}$and $x^2 \leq 4e^{|x|/2}$. Thus, for each $(i,j) \in [N] \times [N]$, we have

$$\begin{aligned}
&\mathbb{E}\left[e^{(m+1)t(X_{i,j} - \mathbb{E}[X_{i,j}] - \epsilon)} \right] \\
&\leq 1 - \epsilon(m+1)t + \frac{1}{2}(m+1)^2 t^2 \mathbb{E}\left[(X_{i,j} - \mathbb{E}[X_{i,j}] - \epsilon)^2 e^{|(m+1)t(X_{i,j} - \mathbb{E}[X_{i,j}] - \epsilon)|} \right] \\
&\leq 1 - \epsilon(m+1)t + 2(m+1)^2 t^2 \mathbb{E}\left[e^{((m+1)t + \frac{1}{2})|X_{i,j} - \mathbb{E}[X_{i,j}] - \epsilon|} \right] \\
&\leq 1 - \epsilon(m+1)t + 8M(m+1)^2 t^2 = 1 - \frac{\epsilon^2}{32M} \\
&\leq \exp\left(-\frac{\epsilon^2}{32M} \right).
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[e^{(m+1)t(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon)} \right] \\
&\leq 1 - \epsilon(m+1)t + \frac{1}{2}(m+1)^2 t^2 \mathbb{E}\left[(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon)^2 e^{|(m+1)t(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon)|} \right] \\
&\leq 1 - \epsilon(m+1)t + 2(m+1)^2 t^2 \mathbb{E}\left[e^{((m+1)t + \frac{1}{2})|\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon|} \right] \\
&\leq 1 - \epsilon(m+1)t + 8M(m+1)^2 t^2 = 1 - \frac{\epsilon^2}{32M} \\
&\leq \exp\left(-\frac{\epsilon^2}{32M} \right).
\end{aligned}$$

Thus, we obtain

$$\mathbb{E}\left[e^{t\sum_{i,j=1}^{N}(X_{i,j} - \mathbb{E}[X_{i,j}] - \epsilon)} \right] \leq \exp\left(-\frac{\epsilon^2 N^2}{32(m+1)M} \right) \leq \exp\left(-\frac{C'\epsilon^2 N}{M} \right),$$

and

$$\mathbb{E}\left[e^{t\sum_{i,j=1}^{N}(\mathbb{E}[X_{i,j}] - X_{i,j} - \epsilon)} \right] \leq \exp\left(-\frac{\epsilon^2 N^2}{32(m+1)M} \right) \leq \exp\left(-\frac{C'\epsilon^2 N}{M} \right),$$

where $C' > 0$ is a universal constant, and we have our conclusion. □

E. Detailed Experimental Settings and Additional Results of Section 6.1

E.1. Detailed Settings

In Section 6.1, we intentionally selected values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ such that $\boldsymbol{\mu}$ is not an eigenvector of $\boldsymbol{\Sigma}$. This was done to ensure that $\boldsymbol{\mu}$ is distinct in a direction from $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Our selected value for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is as follows, but we note that any other general choices would work.

$$\boldsymbol{\mu} = \begin{pmatrix} -0.1067 \\ 0.2572 \\ -0.2392 \\ 0.4135 \\ -0.2179 \\ -0.3995 \\ -0.1437 \\ 0.5950 \\ 0.1786 \\ -0.2839 \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.4481 & 0.0904 & -0.0128 & -0.0245 & 0.1082 & -0.2444 & 0.1817 & 0.0881 & 0.0308 & 0.0450 \\ 0.0904 & 0.4727 & 0.0578 & -0.1620 & 0.0481 & -0.0629 & 0.0509 & -0.1300 & -0.1013 & -0.1706 \\ -0.0128 & 0.0578 & 0.2477 & -0.0728 & -0.0490 & 0.1214 & 0.0189 & 0.0159 & 0.0064 & 0.1649 \\ -0.0245 & -0.1620 & -0.0728 & 0.4457 & 0.0462 & -0.1026 & 0.1188 & -0.0066 & -0.0757 & 0.1065 \\ 0.1082 & 0.0481 & -0.0490 & 0.0462 & 0.2892 & 0.0268 & 0.1117 & -0.1799 & 0.0617 & 0.1787 \\ -0.2444 & -0.0629 & 0.1214 & -0.1026 & 0.0268 & 0.4248 & -0.0868 & 0.0565 & 0.0482 & 0.2182 \\ 0.1817 & 0.0509 & 0.0189 & 0.1188 & 0.1117 & -0.0868 & 0.3638 & -0.0980 & -0.0279 & 0.1658 \\ 0.0881 & -0.1300 & 0.0159 & -0.0066 & -0.1799 & 0.0565 & -0.0980 & 0.4999 & 0.0010 & -0.0318 \\ 0.0308 & -0.1013 & 0.0064 & -0.0757 & 0.0617 & 0.0482 & -0.0279 & 0.0010 & 0.1550 & 0.1723 \\ 0.0450 & -0.1706 & 0.1649 & 0.1065 & 0.1787 & 0.2182 & 0.1658 & -0.0318 & 0.1723 & 0.6230 \end{pmatrix}$$

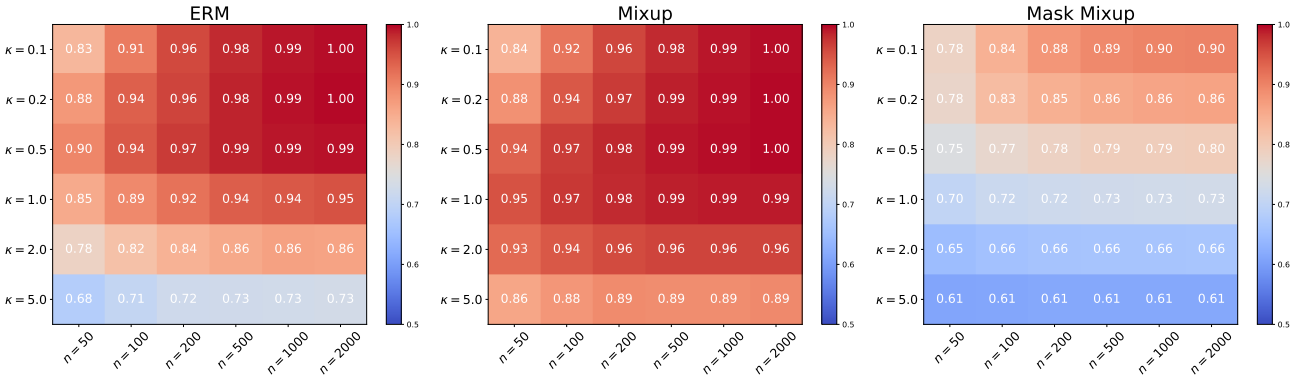
E.2. Addtional Results of Section 6.1

We provide additional experimental results of Section 6.1. We follow the same setting described in Section 6.1 and Appendix E.1 without fixing initial weights for various choices on the number of samples $n = 50, 100, 200, 500, 1000, 2000$ and the separability constant $\kappa = 0.1, 0.2, 0.5, 1.0, 2.0, 5.0$. We plot the average cosine similarity between the Bayes optimal direction and learned weights in Figure 4(a) and one may check that the experiments align with our theoretical findings.

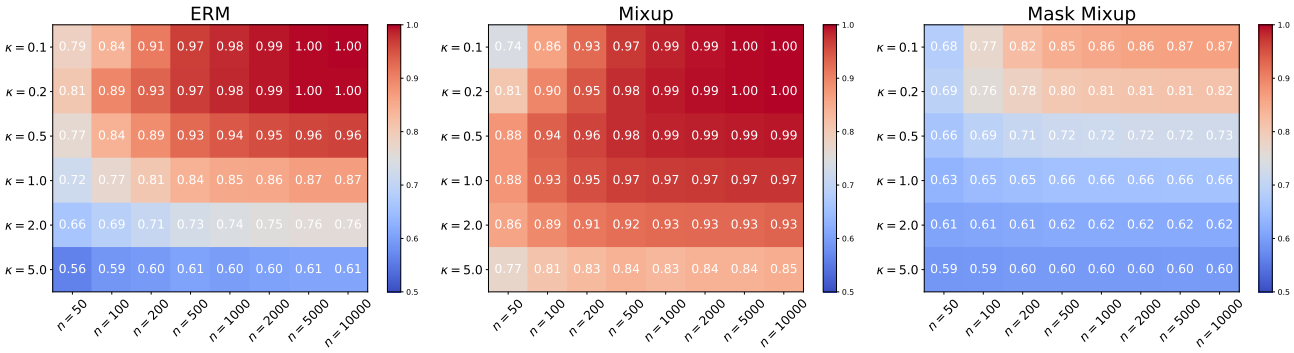
In addition, we provide results on $d = 20$ in Figure 4(b) in order to demonstrate a dependency of a sample complexity on a data dimension d. For the results with $d = 20$, we use additional values on the number of samples $n = 5000, 10000$ and we choose $\boldsymbol{\mu} \in \mathbb{R}^{20}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{20\times 20}$ as

$$\boldsymbol{\mu} = \left(\boldsymbol{\mu}_0^\top, \boldsymbol{\mu}_0^\top\right)^\top, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_0 \end{pmatrix},$$

where $\boldsymbol{\mu}_0 \in \mathbb{R}^{10}$ and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{10\times 10}$ is choice of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ described in Appendix E.1. This choice makes it easy to compare two cases $d = 10$ and $d = 20$. Comparison between Figure 4(a) and Figure 4(b) shows that sample complexities for finding optimal decision boundary significantly increase in dimension d.



(a) $d = 10$



(b) $d = 20$

Figure 4. The average cosine similarity between the Bayes optimal direction and learned weights