Journal of Machine Learning Research 26 (2025) 1-72 Submitted 8/23; Revised 6/25; Published 11/25

General Loss Functions Lead to (Approximate) Interpolation in High Dimensions

Kuo-Wei Lai KLAI36@GATECH.EDU
School of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Vidya Muthukumar VMUTHUKUMAR8@GATECH.EDU
School of Electrical & Computer Engineering
H. Milton School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Editor: Genevera Allen

Abstract

We provide a unified framework that applies to a general family of convex losses across binary and multiclass settings in the overparameterized regime to approximately characterize the implicit bias of gradient descent in closed form. Specifically, we show that the implicit bias is approximated (but not exactly equal to) the minimum-norm interpolation in high dimensions, which arises from training on the squared loss. In contrast to prior work, which was tailored to exponentially-tailed losses and used the intermediate support-vector-machine formulation, our framework directly builds on the primal-dual analysis of Ji and Telgarsky (2021), allowing us to provide new approximate equivalences for general convex losses through a novel sensitivity analysis. Our framework also recovers existing exact equivalence results for exponentially-tailed losses across binary and multiclass settings. Finally, we provide evidence for the tightness of our techniques and use our results to demonstrate the effect of certain loss functions designed for *out-of-distribution* problems on the closed-form solution.

Keywords: implicit bias, convex loss, classification, high-dimensional regime, convex optimization

1. Introduction

The choice of loss function to optimize a model over training examples is an important cornerstone of the machine learning (ML) pipeline. This choice is particularly nuanced for the task of classification, which is evaluated by the 0-1 risk on test data. An elegant classical viewpoint is that training loss functions should be designed as continuous and optimizable surrogates (Bartlett et al., 2006; Zhang, 2004; Lugosi and Vayatis, 2004; Steinwart, 2005) to the 0-1 risk, as the training surrogate loss can often be related to the test surrogate risk, and the test surrogate risk can in turn be related to the test 0-1 risk. However, the first part of this reasoning breaks down in the modern high-dimensional regime, where infinitely many solutions can achieve zero training loss, but the test risk widely varies across these solutions (Zhang et al., 2021; Neyshabur et al., 2014).

The goal of this work is to provide a more transparent understanding of the impact of the training loss function on the eventual solution (and, thereby, its generalization) in this high-dimensional regime. Recent empirical and theoretical work provides a mixed and incomplete picture of the impact of loss. On one hand, large-scale empirical studies (Hui and Belkin, 2020; Kline and Berardi, 2005; Golik et al., 2013; Janocha and Czarnecki, 2017) have shown that the less popular squared loss generates surprisingly competitive performance to the popular cross-entropy loss (the multiclass extension of the binary logistic loss). On the other hand, the cross-entropy loss (and, more generally, the family of *exponentially-tailed losses*, Soudry et al., 2018; Ji and Telgarsky, 2019) is the only one that admits a direct relationship with maximization of the worst-case training data margin, which often correlates with good generalization (Bartlett et al., 1998; Bartlett and Mendelson, 2002). The empirically more challenging task of *out-of-distribution (OOD) generalization* (Mansour et al., 2008) yields further subtleties, with a diversity of loss functions that deviate significantly from this standard family of exponentially-tailed losses being recently designed and evaluated (Sagawa et al., 2019; Cao et al., 2019; Menon et al., 2020; Kini et al., 2021; Wang et al., 2021b). Even for high-dimensional linear models, a comprehensive theory for the impact of a general loss function on the ensuing solution (and, thereby, its generalization) is currently missing. While promising frameworks have been recently provided for the implicit bias of general losses through convex programming (Ji et al., 2020; Ji and Telgarsky, 2021), the properties of the implicit bias itself remain opaque. A separate recent line of work (Muthukumar et al., 2021; Hsu et al., 2021; Wang and Thrampoulidis, 2022; Wang et al., 2021a; Cao et al., 2021) shows that the squared loss and cross-entropy loss can yield identical solutions with high probability in high dimensions, complementing their aforementioned noticed similarities in empirical performance. In particular, both solutions are shown to exactly coincide with *minimum-norm interpolation* (MNI), which enjoys a closed-form expression and often generalizes well in high dimensions (Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2022; Kobak et al., 2020; Muthukumar et al., 2020, 2021). However, these proof techniques are highly tailored to exponentially-tailed losses and in particular the intermediate support-vector-machine (SVM) formulation (Soudry et al., 2018), leaving open whether such equivalences can be proved for more general losses.

1.1 Contributions

In this paper, we characterize the closed-form properties of the implicit bias of general convex losses arising from gradient descent in high dimensional linear models, by building on the primal-dual characterization of the implicit bias provided in Ji and Telgarsky (2021). **In Section 2.2** we show (Proposition 4 and Theorem 6) that general convex losses in conjunction with gradient descent yield solutions that are *approximately directionally close* to minimum-norm interpolation (MNI) on binary labels in a sufficiently high-dimensional regime with high probability. Our approximation error term is a decreasing function of an "effective dimension" which also appears in sufficient and necessary conditions for exact equivalence between the SVM and MNI (Hsu et al., 2021; Ardeshir et al., 2021). In contrast to all prior literature that works with the SVM, our analysis directly leverages the primal-dual framework of Ji and Telgarsky (2021), allowing us to recover the exact equivalence to MNI for exponentially-tailed losses (Hsu et al., 2021) through an alternative proof technique.

Our upper bounds on the approximation error utilize a novel sensitivity analysis of the dual implicit bias in high dimensions and are applicable to general convex losses.

In Section 3.2 we extend our framework and analysis in binary classification to the multiclass classification, where the primal-dual analysis in Ji and Telgarsky (2021) can be naturally extended. We also treat the cross-entropy loss separately and provide an alternative proof of exact equivalence to MNI that is conceptually simpler than the one provided in Wang et al. (2021a), in particular, not requiring any reparameterization of the dual.

Finally, **in Section 4** we provide partial evidence for the tightness of our arguments. First, in Proposition 13 we show that the conditions for exact equivalence in Theorem 4 are not only sufficient but necessary. We leverage this converse result to make an interpretable link between the popular techniques of *importance-weighting* on heavy-tailed losses (Wang et al., 2021b) and vector-scaling of exponentially-tailed losses (Kini et al., 2021) and a type of *cost-sensitive interpolation*, thereby providing a possible explanation for their success in addressing OOD generalization. Finally, under further assumptions on the data covariance, we provide a lower bound in Proposition 17 that in some sense "matches" the upper bound of Theorem 6.

1.2 Related Work

We organize our discussion of related work under three verticals.

1.2.1 Classical Perspectives on Loss Function Design

There are two classical perspectives on loss function design for classification. The first, supported by decades of research in the statistics community, advocates for choosing the loss function to match the negative logarithm of the maximum likelihood function and requires knowledge of the family of conditional distributions of the label. For binary (multiclass) labels, a popular family of conditional distributions is given by the *logistic* (multinomial) model, which yields the empirically popular choice of the *logistic (cross-entropy) loss*. The second and relatively more recent perspective, pioneered by the papers (Bartlett et al., 2006; Zhang, 2004; Lugosi and Vayatis, 2004; Steinwart, 2005), advocates for designing *continuous surrogates* to the discontinuous 0-1 test risk such that a bound on the 0-1 test risk can be easily obtained by inverting a bound on the surrogate test risk. In an indirect sense, this perspective suggests a type of equivalence in surrogate loss functions in terms of ensuing generalization bounds. However, principally because of the reliance on empirical-process-theory (to relate in turn the surrogate test risk to the surrogate training loss), this reasoning can frequently break down in high-dimensional settings, particularly when *perfectly fitting*, or *interpolating* models are considered. This is because infinitely many models interpolate the training data, but each of them suffers a different test risk that is fundamentally unrelated to the training loss. On the other hand, while the relations between test risks (e.g. Bartlett et al. 2006, Theorems 1 and 3) remain universally applicable, they also suffer from some shortcomings in high-dimensional settings—in particular, they are only powerful enough to provide faster *statistical rates* for classification tasks as compared to parameter recovery (Audibert and Tsybakov, 2007), rather than full separations in asymptotic consistency (many classic examples of such separations are considered in Devroye et al., 2013, but such separations were also shown more recently in the overparameterized regime in Muthukumar et al., 2021). The first statistical perspective is similarly not prescriptive in the high-dimensional regime where the maximum-likelihood estimator is no longer unique, and training loss, again, cannot be related to test risk.

1.2.2 Implicit Bias Characterization of Optimization Algorithms

In the modern high-dimensional regime, infinitely many solutions achieve zero training loss for most canonical choices of training loss functions. Therefore, it is not only the loss function but also the choice of optimization algorithm that determines the eventual solution, commonly called the *implicit bias.* An extensive body of work implicitly characterizes this implicit bias of optimization algorithms as solutions to various convex programs (Telgarsky, 2013; Soudry et al., 2018; Ji and Telgarsky, 2019; Ji et al., 2020; Ji and Telgarsky, 2021; Gunasekar et al., 2018a,b; Woodworth et al., 2020; Nacson et al., 2019). The convex program formulation typically does not admit a closed-form solution, except for gradient descent and the squared loss (which yields the MNI for linear models, Engl et al., 1996). Early work here was tailored to exponentially-tailed losses (Soudry et al., 2018; Ji and Telgarsky, 2019), and their established equivalence to the MNI and thereby the squared loss (Muthukumar et al., 2021; Hsu et al., 2021; Wang and Thrampoulidis, 2022; Wang et al., 2021a; Cao et al., 2021) in turn heavily rely on the intermediate SVM formulation. The more recent works (Nacson et al., 2019; Ji et al., 2020; Ji and Telgarsky, 2021) study some non-exponential losses, but leave the exact nature of the implicit bias somewhat mysterious, other than that the ensuing convex program no longer corresponds to the max-margin SVM. For example, Ji et al. (2020, Figure 1) provide a simulated example for which exponential and polynomial losses induce very different directions, and Ji et al. (2020, Proposition 12) provide an example under which the training data margin can be arbitrarily worse for polynomial losses. These are specialized examples of 2-dimensional data that is linearly separable; therefore, do not apply to the high-dimensional regime of interest. Whether such heavy-tailed losses are actually provably worse than exponentially-tailed losses is left open. Our results in this work imply intriguing similarities, but also differences, between heavy-tailed losses and exponential losses in the high-dimensional regime.

The recent papers (Ji et al., 2020; Ji and Telgarsky, 2021) provide promising avenues to understanding the nature of the implicit bias by formulating convex programs for general losses. Ji et al. (2020) make minimal assumptions on the loss function beyond convexity and differentiability, and characterize the implicit bias as the limit of a set of solutions to convex programs that minimize the training loss subject to an ℓ_2-norm constraint of increasing radius (i.e. a *regularization path*). Wang et al. (2021b, Appendix A) show for polynomially-tailed losses that this limit can itself be written as the solution to an explicit convex program, but their proof is tailored to polynomially-tailed losses and in particular their property of positive homogeneity—moreover, no closed-form characterization is provided. On the other hand, Ji and Telgarsky (2021) make slightly stronger assumptions on the loss function, but provide a clearer path to characterizing a closed-form solution for the implicit bias by understanding its *mirror-descent dual* as a solution to an explicit convex program (i.e. not a limit of solutions to convex programs on the regularization path). It is thus natural to attempt to obtain closed-form expressions for the "primal" implicit bias by understanding its "dual" for general losses[1]. A second advantage with analyzing the mirror-descent dual is that we show it automatically yields the non-trivial variable substitution of the multiclass SVM dual that was made in Wang et al. (2021a), resulting in a conceptually simpler proof of SVM equivalence to MNI for the cross-entropy loss. We also show that the primal-dual analysis is applicable to more general formulations of multiclass losses (Zhang, 2004; Tewari and Bartlett, 2007; Ji et al., 2021).

1.2.3 Generalization Analysis of Interpolating Predictors in High Dimensions

A comprehensive theory for overparameterized models arising from training with the squared loss (i.e. the MNI) was provided in work beginning with the papers that analyzed the test regression risk (Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2022; Kobak et al., 2020; Muthukumar et al., 2020). This theory critically utilizes the closed-form expression for the MNI. Sharply analyzing the classification risk poses distinct challenges, the most daunting of which is the lack of a closed-form expression for the solution arising from any other convex loss function used for classification. To tackle this challenge for the special case of exponential losses, Muthukumar et al. (2021) introduced a two-step recipe. First, they related the implicit bias of exponential losses (i.e. the SVM) to the MNI—in fact, by showing an *exact equivalence* result (which was since improved on by Hsu et al., 2021). Second, they sharply analyzed the classification test risk of the MNI and showed that it can achieve classification-consistency even when a corresponding regression task would not be consistent. It is worth noting that this type of consistency result cannot be easily recovered through any generalization bound that relies on empirical-process-theory, including margin-based data-dependent generalization bounds (as described in Muthukumar et al. 2021, Section 6). This recipe was since applied to binary and multiclass Gaussian and sub-Gaussian mixture models to identify new high-dimensional regimes in which classification-consistency is possible (Wang and Thrampoulidis, 2022; Cao et al., 2021; Wang et al., 2021a; Subramanian et al., 2022). To be able to apply this recipe to more general losses, corresponding equivalences would need to be established between general losses and the MNI, which is the focus of this paper.

Other than the approach described above, two other families of techniques are prevalent in the recent literature. The first applies to proportionally high-dimensional regimes (where $d \propto n$) and directly characterizes the limiting test risk as $(d, n) \to \infty$ as the solution to a system of nonlinear equations, beginning with the efforts tailored to exponential or exponentially tailed losses (Huang, 2017; Sur and Candès, 2019; Mai et al., 2019; Salehi et al., 2019; Deng et al., 2022; Montanari et al., 2019). More recently, Loureiro et al. (2021) provide precise asymptotic analysis for general losses and multiclass classification for Gaussian mixture models for *regularized empirical risk minimization* with general losses and regularizers[2]. However, they do not examine in detail the impact of loss functions on performance. In general, none of our results for general losses have direct implications for

1. This is especially true given that the mirror-descent dual for the case of exponentially-tailed losses turns out to exactly correspond to a scalar multiple of the SVM dual. Indeed, the proofs of SVM equivalence all construct a dual witness.
2. Note that this covers the implicit bias of gradient descent when the regularization proportion $\lambda \to 0$ due to the results of Ji et al. (2020).

this proportional regime. However, we believe the auxiliary convex program proposed in Lemma 3 might be of independent interest, particularly for the subset of approaches above that utilize Gordon's comparison theorems and the convex Gaussian min-max theorem.

The second technique was proposed by Chatterji and Long (2021) for directly analyzing the generalization error of the implicit bias of exponential losses on a sub-Gaussian mixture model. The key technical innovation is to prove a "loss ratio" bound: under sufficiently overparameterized settings, Chatterji and Long (2021) show that the training losses of any two examples are within a constant factor of each other throughout the optimization path of gradient descent. This proof technique is quite generally applicable and was since used for polynomially-tailed losses (Wang et al., 2021b), deep linear networks (Chatterji et al., 2022) and certain 2-layer neural networks on high-dimensional data (Frei et al., 2022). However, the loss-ratio bound often requires a much larger data dimension $d \gg n^2$ to hold as compared to the MNI-equivalence approach to analyzing the SVM, as shown explicitly in Wang and Thrampoulidis (2022). It is not clear whether this dimension requirement is tight even in the worst case. A natural question of interest is whether a loss-ratio bound implies exact or approximate equivalence of solutions, or vice versa. Behnia et al. (2022) showed recently that a loss-ratio bound can imply exact equivalence to the MNI in the case of exponential losses, but this is a research direction that is otherwise largely unexplored.

1.2.4 Comparison to Related Work

In Table 1, we succinctly situate our work in the literature on the implicit bias of classification-oriented loss functions. In sum, we go *beyond worst-case characterizations* (by investigating an approximate equivalence to the MNI under sufficiently high-dimensional random data) of the implicit bias of gradient descent on *general convex loss functions* (going beyond previous work that only established an approximate equivalence for the class of exponentially-tailed losses). While the beyond-worst-case aspect had been previously explored on exponentially-tailed losses (Muthukumar et al., 2021; Hsu et al., 2021; Wang et al., 2021a), and a worst-case characterization of general losses was provided (Ji et al., 2020; Ji and Telgarsky, 2021), prior to our work these had not been studied together. Our starting point for analyzing the implicit bias of general losses is the insightful *dual* convex program characterization provided by Ji and Telgarsky (2021). We introduce several novel ideas over and above their work; prominent among them a new, and simpler to analyze, auxiliary convex program for the dual (Lemma 3), as well as a new sensitivity analysis of this auxiliary program that is "fixed-design" in nature (Theorem 6). Our main sensitivity theorem can easily be applied in conjunction with standard results on high-dimensional probability, e.g. random matrix concentration, to establish approximate equivalence to the MNI for general losses and a variety of random data models (Corollary 7).

1.3 Notations

We use lower-case boldface (e.g. $\boldsymbol{x}$) to denote vector notation and upper-case boldface (e.g. $\boldsymbol{X}$) to denote matrix notation. We use $\|\cdot\|_p$ to denote the ℓ_p-norm of a vector for $p \in [1, \infty)$ and $\|\cdot\|_2$ to additionally denote the operator norm of a matrix. $\mathrm{diag}(\boldsymbol{x})$ denotes the diagonal matrix whose entries are given by the vector $\boldsymbol{x}$. For a 1-dimensional function $h(\cdot) : \mathbb{R} \to \mathbb{R}$, we frequently overload notation and denote its element-wise operation on a

	Implicit bias for worst-case data	Implicit bias vs. MNI under high-dimensional random data
Exponentially-tailed losses (e.g. exponential loss, logistic loss)	Soudry et al. (2018) Ji and Telgarsky (2019) Ravi et al. (2024)	Muthukumar et al. (2021) Hsu et al. (2021) Wang et al. (2021a)
General convex losses (e.g. polynomially-tailed loss)	Ji et al. (2020) Ji and Telgarsky (2021)	This work

Table 1: Our result, contextualized in related implicit bias literature.

vector by $h(\boldsymbol{x}) \coloneqq \big(h(x_1), \cdots, h(x_n)\big)^\top$. All other appearances of the notation $h(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^k$ instead denote a function that takes a vector-valued argument. We denote first and second derivatives by $'$ and $''$ respectively, and use ∂ to denote a partial derivative. We use the shorthand notation $[n]$ to denote the set of natural numbers $\{1, \ldots, n\}$.

2. Approximate Equivalences for Binary Classification

Since our results build on the primal-dual analysis presented in Ji and Telgarsky (2021), we reproduce their assumptions on the data and loss function below.

2.1 Problem Setup

We consider a labeled data set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ satisfies the normalization $\|\boldsymbol{x}_i\|_2 \leq 1$ (which can be done without loss of generality) and the labels $y_i \in \{-1, 1\}$ are binary. We denote $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} \coloneqq (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$. We focus on an unbounded, unregularized empirical risk minimization (ERM) problem with a margin-based loss function and a linear classifier

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \mathcal{R}(\boldsymbol{w}) \coloneqq \frac{1}{n} \sum_{i=1}^n \ell\left(-y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right) = \frac{1}{n} \sum_{i=1}^n \ell\left(y_i \langle \boldsymbol{w}, \boldsymbol{z}_i \rangle\right), \tag{1}$$

where we denote $\boldsymbol{z}_i \coloneqq -\boldsymbol{x}_i$, $\boldsymbol{Z} \coloneqq -\boldsymbol{X}$, and $\boldsymbol{w} \in \mathbb{R}^d$ is the set of parameters of the linear classifier.

Assumption 1 (From Ji and Telgarsky, 2021, Assumption 1) *The loss function $\ell(\cdot)$ is twice differentiable, and satisfies:*

1. ℓ, ℓ', $\ell'' > 0$, *and* $\lim_{z \to -\infty} \ell(z) = 0$.
2. $z\ell'(z)/\ell(z)$ *is increasing on* $(-\infty, 0)$, *and* $\lim_{z \to -\infty} z\ell'(z) = 0$.
3. *For all $b \geq 1$, there exists $c > 0$ (which may depend on b), such that for all $a > 0$, we have* $\ell'\big(\ell^{-1}(a)\big)/\ell'\big(\ell^{-1}(ab)\big) \geq c$.
4. *Given $\boldsymbol{\xi} \in \mathbb{R}^n$, we define*

$$\mathcal{L}(\boldsymbol{\xi}) \coloneqq \sum_{i=1}^n \ell(\xi_i), \text{ and } \psi(\boldsymbol{\xi}) \coloneqq \ell^{-1}\left(\mathcal{L}(\boldsymbol{\xi})\right),$$

and the "generalized sum" ψ is convex and β-smooth with respect to ℓ_∞ norm.

Next, we show that for any loss function $\ell(\cdot)$ that satisfies Assumption 1, there exists an explicit analytical function $g(\cdot)$, derived as the limit of a certain ratio of derivatives of inverses of the loss function $\ell(\cdot)$, that will be instrumental in our analysis of the implicit bias. This lemma is a direct implication of Assumption 1, without any additional assumptions.

Lemma 1 *Under Assumption 1, the limit* $\lim_{a\to 0}\frac{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)}$ *exists for every* $0 < z \leq 1$*. Moreover, there exists a function* $g(\cdot)$ *such that* $g(z) := \lim_{a\to 0}\frac{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)}$ *for* $0 < z \leq 1$*, where* $g : (0,1] \to (0,1]$ *is a non-negative, strictly increasing, convex function satisfying* $g(1) = 1$*.*

The proof of Lemma 1 can be found in Appendix A.1. Lemma 1 is central to all of our results, since different loss functions $\ell(\cdot)$ may result in different functions $g(\cdot)$. In particular, we critically use the convexity of the function $g(\cdot)$ to obtain a simplified auxiliary convex program, that is equivalent in optimal solution, underlying the dual of the implicit bias. Figure 1 displays various examples of the form the function $g(\cdot)$ takes for specific, commonly used loss functions.

2.1.1 The Implicit Bias Formulation

We use the gradient descent algorithm to solve this unregularized empirical risk minimization problem with initial weights $\boldsymbol{w}_0$ and the update rule: $\boldsymbol{w}_{t+1} := \boldsymbol{w}_t - \eta_t \nabla \mathcal{R}(\boldsymbol{w}_t)$ for $t \geq 0$. We also denote, in the context of mirror-descent analysis, the "primal" $\boldsymbol{p}_t := \text{diag}(\boldsymbol{y})\,\boldsymbol{Z}\boldsymbol{w}_t \in \mathbb{R}^n$ and its corresponding "dual" $\boldsymbol{q}_t := \nabla\psi(\boldsymbol{p}_t) \in \mathbb{R}^n$, where

$$q_{t,i} = \frac{\ell'\left(p_{t,i}\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\ell\left(p_{t,i}\right)\right)\right)} = \frac{\ell'\left(p_{t,i}\right)}{\ell'\left(\psi\left(\boldsymbol{p}_t\right)\right)}. \tag{2}$$

These mirror-descent primal and dual terms were defined in Ji and Telgarsky (2021).

Next, we assume that the data can be interpolated or perfectly fitted, which corresponds to a full-rank assumption on the Gram matrix $\boldsymbol{X}\boldsymbol{X}^\top$. Note that this in turn implies that the data set is linearly separable. This full-rank assumption is satisfied with high probability in the overparameterized regime $d \gg n$ for most canonical data distributions; see, e.g. Hsu et al. (2021).

Assumption 2 *We assume that* $d \geq n$ *and the data Gram matrix satisfies* $\boldsymbol{X}\boldsymbol{X}^\top \succ \boldsymbol{0}$*. This in turn implies that there exists a linear separator* $\boldsymbol{u} \in \mathbb{R}^d$ *that* $y_i \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle > 0$ *for all* $i \in [n]$*.*

We restate the primal-dual implicit bias formulation of Ji and Telgarsky (2021, Theorem 5) below.

Lemma 2 (From Ji and Telgarsky, 2021, Theorem 5) *Under Assumptions 1 and 2, and provided that* $\hat{\eta}_t := \eta_t \ell'\left(\psi\left(\boldsymbol{p}_t\right)\right)/n \leq 1/\beta$ *is nonincreasing and* $\sum_{t=0}^{\infty}\hat{\eta}_t = \infty$*, the primal and dual implicit bias* $(\bar{\boldsymbol{w}}, \bar{\boldsymbol{q}})$ *are given by*

$$\bar{\boldsymbol{w}} := \lim_{t\to\infty}\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2} = \frac{-\boldsymbol{Z}^\top \mathit{diag}(\boldsymbol{y})\,\bar{\boldsymbol{q}}}{\left\|-\boldsymbol{Z}^\top \mathit{diag}(\boldsymbol{y})\,\bar{\boldsymbol{q}}\right\|_2} = \frac{\boldsymbol{X}^\top \mathit{diag}(\boldsymbol{y})\,\bar{\boldsymbol{q}}}{\left\|\boldsymbol{X}^\top \mathit{diag}(\boldsymbol{y})\,\bar{\boldsymbol{q}}\right\|_2}, \tag{3}$$

and

$$\bar{\boldsymbol{q}} \in \underset{\psi^*(\boldsymbol{q})\leq 0}{\arg\min}\, f\left(\boldsymbol{q}\right), \tag{4}$$

where ψ^* *denotes the convex conjugate of* ψ*, and we define* $f\left(\boldsymbol{q}\right) \coloneqq \frac{1}{2}\left\|\boldsymbol{X}^\top \mathit{diag}\left(\boldsymbol{y}\right)\boldsymbol{q}\right\|_2^2$.

Consequently, a characterization of any solution $\bar{\boldsymbol{q}}$ to the convex program (4) defined in Lemma 2 would directly characterize the desired primal implicit bias $\bar{\boldsymbol{w}}$. Accordingly, our techniques and results largely focus on characterizing a suitable solution $\bar{\boldsymbol{q}}$ to (4).

2.1.2 Minimum-norm Interpolation

We are especially interested in relating the primal implicit bias $\bar{\boldsymbol{w}}$ to the minimum-norm interpolation (MNI) $\boldsymbol{w}_{\mathsf{MNI}} \coloneqq \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}$. The MNI arises as the implicit bias of gradient descent applied to the square loss under a sufficiently small step size and initialization $\boldsymbol{w}_0 = \boldsymbol{0}$ (Engl et al., 1996). For example, it is easy to see that the candidate dual solution $\boldsymbol{q} \coloneqq \mathrm{diag}\left(\boldsymbol{y}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}$ would correspond to a primal solution proportional to $\boldsymbol{w}_{\mathsf{MNI}}$; we will utilize this candidate solution in our equivalence results.

2.2 Main Results

The convex program defined in (4) is challenging to directly work with and analyze. This is primarily because the convex conjugate constraint $\psi^*\left(\boldsymbol{q}\right)$ is in general an implicitly defined function on $\boldsymbol{q}$ (except for the exact exponential loss as shown in Ji and Telgarsky, 2021), and therefore its non-positivity can be difficult to verify. To make progress, we present a simple but critical auxiliary convex program that recovers the same dual implicit bias solution in Lemma 3 that critically utilizes the convex function $g\left(\cdot\right)$ that we defined in Lemma 1.

Lemma 3 *Under Assumptions 1 and 2, any solution to the auxiliary convex program*

$$\bar{\boldsymbol{q}} \in \underset{\boldsymbol{q}\in\mathbb{R}^n}{\arg\min}\, \underbrace{\frac{1}{2}\boldsymbol{q}^\top \mathit{diag}\left(\boldsymbol{y}\right)\boldsymbol{X}\boldsymbol{X}^\top \mathit{diag}\left(\boldsymbol{y}\right)\boldsymbol{q}}_{f(\boldsymbol{q})} \tag{5}$$

$$\textit{subject to} \quad -q_i < 0 \quad \textit{for all } i \in [n], \textit{ and} \quad 1 - \sum_{i=1}^{n} g^{-1}\left(q_i\right) \leq 0,$$

is also an optimal solution to the original convex program (4).

The full proof for Lemma 3 is contained in Appendix A.2. The proof of Lemma 3 follows via a two-part argument. We first show that the convex conjugate constraint in the convex program (4) must be active at optimality, which implies that $\psi^*\left(\bar{\boldsymbol{q}}\right) = 0$. We then demonstrate that the condition $\sum_{i=1}^n g^{-1}\left(q_i\right) = 1$, derived from the Karush-Kuhn-Tucker (KKT) (Karush, 1939) conditions for the auxiliary convex program (5), is sufficient to ensure that $\psi^*\left(\boldsymbol{q}\right) = 0$. Next, we show that any solution to the original convex program (4) also satisfies $\sum_{i=1}^n g^{-1}\left(q_i\right) = 1$. Therefore, every solution to the auxiliary convex program (5) is also a solution to the original program (4). The idea is illustrated in Figure 2.

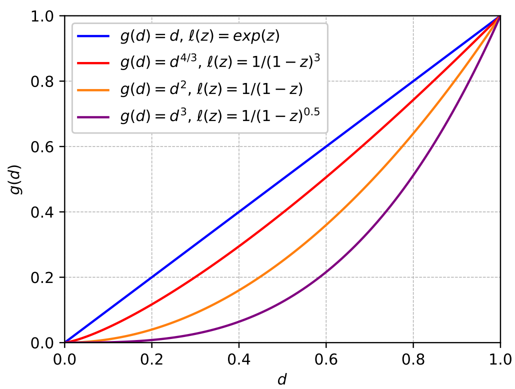


Figure 1: Plots of $g(\cdot)$ for different losses.

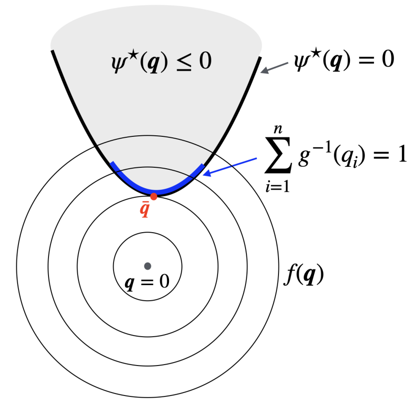


Figure 2: Illustration of the original convex program (4) and how it relates to the auxiliary convex program (5).

2.2.1 Warm-up: Conditions for Exact Equivalence to MNI

Although the auxiliary convex program in (5) is simpler to analyze, it still does not admit a closed-form solution in general. We begin by providing a warm-up result characterizing settings under which (5) does admit a closed-form solution, which turns out to yield the MNI primal $\boldsymbol{w}_{\mathsf{MNI}}$.

Proposition 4 *Under Assumptions 1 and 2, the following statements hold:*

1. *If $\boldsymbol{y}$ is an exact eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$, the implicit bias $\bar{\boldsymbol{w}}$ is parallel to the MNI $\boldsymbol{w}_{\mathsf{MNI}}$, i.e. $\frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} = \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2}$.*
2. *For any loss function that admits the identity function $g(d) = d$, the implicit bias $\bar{\boldsymbol{w}}$ is parallel to the MNI $\boldsymbol{w}_{\mathsf{MNI}}$, i.e. $\frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} = \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2}$ iff*

$$\boldsymbol{X}\boldsymbol{X}^\top \succ \boldsymbol{0} \text{ and } \boldsymbol{\beta} := \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y} \text{ satisfies } y_i\beta_i > 0 \text{ for all } i \in [n]. \tag{6}$$

The full proof of Proposition 4 is provided in Appendix A.3 and works directly with the KKT conditions of the auxiliary convex program (5). We make a few remarks here about this proposition. First, note that Part 2 of Proposition 4 recovers the sufficient and necessary condition for the equivalence between the SVM and the MNI, i.e. *support-vector-proliferation* (SVP) originally studied in Muthukumar et al. (2021); Hsu et al. (2021). This makes sense, as the class of loss functions that admits the identity function $g(d) = d$ corresponds to the class of *exponentially-tailed losses*, which are well-known to generate implicit bias that is parallel to the SVM (Soudry et al., 2018). Next, note that the condition for general losses in Part 1 (that $\boldsymbol{y}$ is an exact eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$) is significantly stronger than the condition in Part 2—while $\boldsymbol{y}$ being an exact eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$ implies Equation (6), the reverse implication does not hold. We show in Proposition 13 in Section 4 that the exact-eigenvector condition is in fact necessary for any loss function that does not admit the identity function $g(d) = d$. Finally, we informally remark on some sufficient conditions under which the exact-eigenvector condition would hold. One easily verifiable case is when the Gram matrix is an exact multiple of the identity, as stated below.

Corollary 5 *If $\boldsymbol{X}\boldsymbol{X}^\top = \alpha \boldsymbol{I}$ for some $\alpha > 0$, then we have $\bar{\boldsymbol{q}} \propto \boldsymbol{1}$ and $\bar{\boldsymbol{w}} \propto \boldsymbol{w}_{\mathsf{MNI}}$ for any loss satisfying Assumption 1.*

Corollary 5 describes a scenario that will not arise in practice, as in general the Gram matrix $\boldsymbol{X}$ will be random. Muthukumar et al. (2020) showed that the scenario $\boldsymbol{X}\boldsymbol{X}^\top = \alpha \boldsymbol{I}$ can, however, arise with data that is *uniformly spaced* in conjunction with certain feature families. Uniformly-spaced data models also appear in some pedagogical analyses of nonparametric statistics, as they often provide a simpler analysis as compared to random data (Nemirovski, 2000; Tsybakov, 2009).

2.2.2 Main Result: Approximate Equivalence to MNI in High Dimensions

We now turn to more realistic scenarios to handle random data. In general, we only expect the Gram matrix to be *close* to a multiple of the identity (in the sense that the operator norm of the difference $\left\|\boldsymbol{X}\boldsymbol{X}^\top - \alpha \boldsymbol{I}\right\|_2$ is typically controlled in high dimensions). This leads to whether the solution $\bar{\boldsymbol{w}}$ is now close in its direction to $\boldsymbol{w}_{\mathsf{MNI}}$. Theorem 6 below addresses this question.

Theorem 6 *Under Assumptions 1 and 2, consider any value of $\alpha > 0$ satisfying $\frac{\|\boldsymbol{X}\boldsymbol{X}^\top - \alpha \boldsymbol{I}\|_2}{\alpha} \leq \frac{1}{3}$. Then, the implicit bias $\bar{\boldsymbol{w}}$ converges in direction to $\boldsymbol{w}_{\mathsf{MNI}}$ at the rate*

$$\left\| \frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} - \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2} \right\|_2 \leq \frac{C \left\| \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y} - \alpha \boldsymbol{y} \right\|_2}{\alpha \|\boldsymbol{y}\|_2}, \tag{7}$$

where C is a universal constant that does not depend on $\alpha, \boldsymbol{X}$ or $\boldsymbol{y}$.

Theorem 6 shows that every loss function satisfying Assumption 1 yields an approximately equivalent implicit bias in high dimensions. It also recovers Corollary 5 as a special case (as in this case the RHS of Equation 7 becomes equal to 0).

Before discussing how to prove Theorem 6, we describe a canonical high-dimensional statistical ensemble under which it implies directional convergence of the implicit bias $\bar{\boldsymbol{w}}$ to the MNI $\boldsymbol{w}_{\mathsf{MNI}}$.

Corollary 7 *Assume independent and identically distributed data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ such that each covariate satisfies one of the following: a) $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, and we denote the spectrum of $\boldsymbol{\Sigma}$ by $\boldsymbol{\lambda}$;* or *b) $\boldsymbol{x}_i = diag(\boldsymbol{\lambda})^{1/2} \boldsymbol{z}_i$, where $\boldsymbol{z}_i$ has independent entries such that each z_{ij} is mean-zero, unit-variance, and sub-Gaussian with parameter $v > 0$ (i.e. $\mathbb{E}[z_{ij}] = 0, \mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[e^{tz_{ij}}] \leq e^{vt^2/2}$ for all $t \in \mathbb{R}$). In both cases, define the effective dimensions $d_2 := \frac{\|\boldsymbol{\lambda}\|_1^2}{\|\boldsymbol{\lambda}\|_2^2}$ and*

$d_\infty := \frac{\|\boldsymbol{\lambda}\|_1}{\|\boldsymbol{\lambda}\|_\infty}$ *and assume that* $d_2 \gg v^2 n$ *and* $d_\infty \gg vn$*. Then, Theorem 6 implies that*

$$\left\| \frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} - \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2} \right\|_2 \leq C \cdot v \cdot \max\left\{ \sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right\},$$

with probability at least $1 - 4e^{-cn}$*, where* $C, c > 0$ *are appropriately chosen universal constants. This implies that* $\left\| \frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} - \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2} \right\|_2$ *is vanishingly small for any high-dimensional ensemble* $\{(n, d, \boldsymbol{\lambda})\}_{n \geq 1}$ *satisfying* $d_2 \gg v^2 n$ *and* $d_\infty \gg vn$*.*

The proof of Corollary 7 is in Appendix A.5 and applies the operator norm concentration inequality of Hsu et al. (2021, Lemma 8) (which in turn uses a volume argument from Pisier, 1999). The corollary demonstrates the role of a sufficiently high-dimensional ensemble in ensuring that the implicit bias from a general convex loss eventually converges, in a directional sense, to the MNI. As a special case, consider the *isotropic* high-dimensional ensemble for which $\boldsymbol{\lambda} = \mathbf{1}$ and $v = 1$. Here, we have $d_2 = d_\infty = d$, and the required effective dimension conditions reduce to $d \gg n$. Hsu et al. (2021) show that when $d_2 \gg v^2 n$ and $d_\infty \gg vn \log n$, the stronger phenomenon of SVP would occur[3], working from the condition in Proposition 4 Part 2. The anisotropic Gaussian or independent sub-Gaussian model for covariates considered in Corollary 7 does not directly cover certain high-dimensional ensembles for which conditions for SVP have been characterized; in particular, mixture models (Wang and Thrampoulidis, 2022; Wang et al., 2021a; Cao et al., 2021). We believe that results similar to Corollary 7 can also be established for these cases.

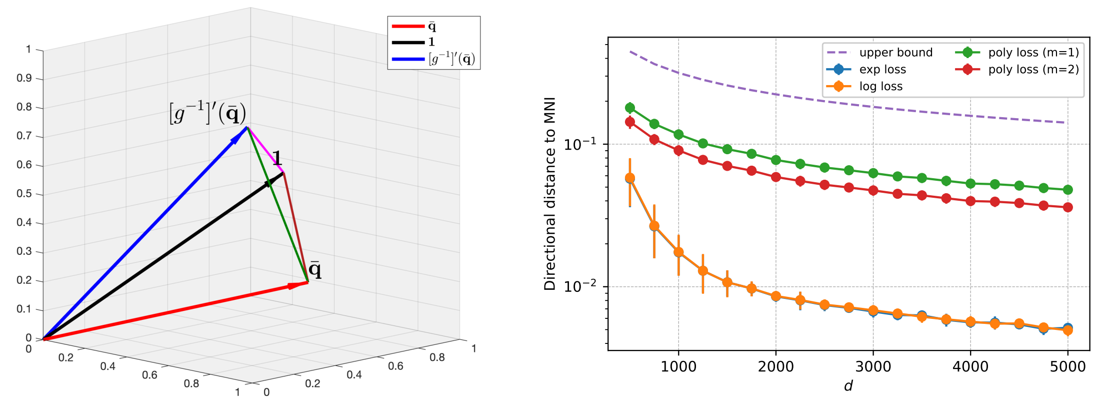
2.2.3 Proof Sketch for Theorem 6

The full proof of Theorem 6 is in Appendix A.4. We divide the proof in four steps. **In Step 1,** we begin with the auxiliary convex program (5), and determine necessary characteristic equations for the solution $\bar{\boldsymbol{q}}$; in particular, we show that it is *necessary* for $\bar{\boldsymbol{q}}$ to solve the system of nonlinear equations $\boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}} = \mu \mathrm{diag}(\boldsymbol{y}) \left[g^{-1}\right]'(\bar{\boldsymbol{q}})$ for some $\mu > 0$. **In Step 2,** we use the relative closeness (in an operator-norm sense) of $\boldsymbol{X}\boldsymbol{X}^\top$ to a multiple of $\boldsymbol{I}$ to show that the nonlinear equation above implies that the vectors $\bar{\boldsymbol{q}}$ and $\left[g^{-1}\right]'(\bar{\boldsymbol{q}})$ are close in a directional sense in Equation (26).

Next, **Step 3** proves a simple but non-trivial observation which, as pictured in Figure 3a, states that the vector $\mathbf{1}$ is *in between* the vectors $\bar{\boldsymbol{q}}$ and $\left[g^{-1}\right]'(\bar{\boldsymbol{q}})$ in Equation (28) (implying that its angle with either of the vectors is smaller than the angle between $\bar{\boldsymbol{q}}$ and $\left[g^{-1}\right]'(\bar{\boldsymbol{q}})$). The proof of this observation critically uses the convexity of $g(\cdot)$ which turns out to lead to an application of Chebyshev's sum inequality (Hardy et al., 1952) to complete the desired argument. Steps 1, 2 and 3 together give a rate on the directional convergence of the *dual optimal solution* $\bar{\boldsymbol{q}}$ to $\mathbf{1}$ in Equation (29).

The final **Step 4** uses the primal-dual relationship in Equation (2) to show that the primal convergence rate is identical to the dual convergence rate up to universal constant

3. The careful reader might notice that the SVP result has an extra $\log n$ factor in the required condition on the effective dimension d_∞, that in fact turns out to be necessary (Ardeshir et al., 2021). There is no contradiction with our results, because SVP describes a stronger phenomenon of *exact* equivalence that holds even when n and d are finite, as opposed to our directional convergence result, which only gives exact asymptotic equivalence as $(n, d) \to \infty$.



(a) Proof idea of Theorem 6 (b) Simulation for binary classification

Figure 3: Panel (a) illustrates the relationship between the vectors $\bar{\boldsymbol{q}}$, $\left[g^{-1}\right]'(\bar{\boldsymbol{q}})$ and $\mathbf{1}$ for the loss function $\ell(z) = 1/(1-z)$. Panel (b) is a simulation that compares the implicit bias of gradient descent to the MNI. The covariate-response pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ are independently and identically distributed (IID) with a fixed sample size $n = 100$ and varying data dimension d, where $\boldsymbol{x}_i$ is isotropic Gaussian and y_i is uniformly distributed in $\{\pm 1\}$. Gradient descent is run for the minimum of 10^3 iterations or when the empirical risk falls below 10^{-12}. The results demonstrate that the directional distance to the MNI is upper bounded by the theoretical guarantee in Theorem 6. For exponentially-tailed loss functions, exact convergence to the MNI is not observed, as it only occurs when the number of iterations of gradient descent is infinite. Each experiment is repeated over 100 independent trials.

factors and is proved through a series of algebraic manipulations which repeatedly utilize the operator-norm concentration of $\boldsymbol{X}\boldsymbol{X}^\top$ around $\alpha\boldsymbol{I}$.

2.2.4 Loss Functions Satisfying Assumption 1

We conclude this section with a brief discussion of popular loss functions that satisfy Assumption 1, and to which Proposition 4 and Theorem 6 are therefore applicable. These loss functions are also discussed in Ji and Telgarsky (2021, Sec. 5).

Proposition 8 *Assumption 1 is satisfied by the following losses with the corresponding values of the function $g(\cdot)$ provided:*

$$\text{Exponential loss: } \ell_{exp} \coloneqq \exp(z)\,,\ g_{exp}(d) = d$$
$$\text{Logistic loss: } \ell_{log} \coloneqq \ln\big(1 + \exp(z)\big)\,,\ g_{log}(d) = d$$
$$\text{Polynomial loss (degree } m > 0\text{): } \ell_{poly}(z) \coloneqq \begin{cases} \frac{1}{(1-z)^m} & z \le 0 \\ \frac{1}{(1+z)^m} + 2mz & z > 0 \end{cases}, \ g_{poly}(d) = d^{\frac{m+1}{m}}.$$

The proof of Proposition 8 is provided in Appendix A.6. A plot of the function $g(\cdot)$ that underlies each loss is given in Figure 1. Note that $g(\cdot)$ that deviate more from $g(d) = d$

are heavier-tailed, that the furthest pictured such function corresponds to the purple line $g(d) = d^3$ for the polynomial loss with degree $m = 0.5$.

3. Approximate Equivalences for Multiclass Classification

We now extend our analysis from the binary to the multiclass setting.

3.1 Multiclass Problem Setup

We consider a labeled data set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in [K]$. We assume there is at least one example in each class. For each class k, we assign a weight vector $\boldsymbol{w}_k \in \mathbb{R}^d$. We denote as shorthand $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times d}$, and an n-dimensional encoding of the multiclass labels $\boldsymbol{c}_k(\alpha,\beta) = \big(c_{k,1}(\alpha,\beta), \cdots, c_{k,n}(\alpha,\beta)\big)^\top \in \mathbb{R}^n$, where $c_{k,i}(\alpha,\beta) = \begin{cases} \alpha & k = y_i \\ -\beta & k \neq y_i \end{cases}$, for all $\alpha, \beta > 0$ for all $i \in [n]$ and $k \in [K]$. (We frequently omit the arguments (α,β) and simply write $\boldsymbol{c}_k$ when the values of α and β are clear from context.)

We assume w.l.o.g. that $\|\boldsymbol{x}_i\|_2 \leq \max\limits_{k\in[K]} |c_{k,i}|$. We concatenate the weight vector $\boldsymbol{W} \in \mathbb{R}^{Kd}$, data matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{Kn\times Kd}$ and label matrix across classes $\boldsymbol{C} \in \mathbb{R}^{Kn\times Kn}$ as below

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_k \end{bmatrix}, \ \tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{X} \end{bmatrix}, \ \boldsymbol{C} = \begin{bmatrix} \operatorname{diag}\big(\boldsymbol{c}_1^{-1}\big) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \operatorname{diag}\big(\boldsymbol{c}_K^{-1}\big) \end{bmatrix}.$$

We focus on an unbounded, unregularized ERM problem with a linear classifier

$$\min_{\boldsymbol{W}\in\mathbb{R}^{Kd}} \mathcal{R}(\boldsymbol{W}) := \frac{1}{n}\sum_{i=1}^n \mathfrak{L}\left(-\left\{c_{k,i}^{-1}\langle \boldsymbol{w}_k, \boldsymbol{x}_i\rangle\right\}_{k=1}^K\right) = \frac{1}{n}\sum_{i=1}^n \mathfrak{L}\left(\left\{c_{k,i}^{-1}\langle \boldsymbol{w}_k, \boldsymbol{z}_i\rangle\right\}_{k=1}^K\right), \quad (8)$$

where we denote $\boldsymbol{z}_i := -\boldsymbol{x}_i$, and therefore $\boldsymbol{Z} := -\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\tilde{\boldsymbol{Z}} := -\tilde{\boldsymbol{X}} \in \mathbb{R}^{Kn\times Kd}$. Next, we introduce different variants of the multiclass loss function, which we denote by $\mathfrak{L}$.

Assumption 3 (One-vs-all multiclass loss) *The multiclass loss function satisfies*

$$\mathfrak{L}\left(\left\{c_{k,i}^{-1}\langle \boldsymbol{w}_k, \boldsymbol{z}_i\rangle\right\}_{k=1}^K\right) = \sum_{k=1}^K \ell\left(c_{k,i}^{-1}\langle \boldsymbol{w}_k, \boldsymbol{z}_i\rangle\right)$$

where ℓ follows Assumption 1 Parts 1, 2 and 3. Additionally, given $\boldsymbol{\xi}_k \in \mathbb{R}^n$ and $\boldsymbol{\Xi} = \left(\boldsymbol{\xi}_1^\top, \cdots, \boldsymbol{\xi}_K^\top\right)^\top \in \mathbb{R}^{Kn}$, we define $\mathcal{L}(\boldsymbol{\Xi}) := \sum_{i=1}^n \mathfrak{L}\left(\{\xi_{k,i}\}_{k=1}^K\right)$ and $\psi(\boldsymbol{\Xi}) := \ell^{-1}\left(\mathcal{L}(\boldsymbol{\Xi})\right)$, where ψ is jointly convex and β-smooth with respect to the ℓ_∞ norm.

Our framework is able to handle general losses satisfying Assumption 1 under the popular one-vs-all framework. Finally, we treat the popular cross-entropy loss, which is a generalization of the binary logistic loss, separately.

Assumption 4 (Cross-entropy loss) *The loss function* $\mathfrak{L}$ *satisfies*

$$\mathfrak{L}\left(\left\{c_{k,i}^{-1}\left\langle \boldsymbol{w}_k, \boldsymbol{z}_i\right\rangle\right\}_{k=1}^{K}\right) = -\ln\left(\frac{\exp\left(\left\langle \boldsymbol{w}_{y_i}, \boldsymbol{x}_i\right\rangle\right)}{\sum_{k=1}^{K}\exp\left(\left\langle \boldsymbol{w}_k, \boldsymbol{x}_i\right\rangle\right)}\right)$$
$$= \ln\left(1 + \sum_{k\neq y_i}^{K}\exp\left(c_{y_i,i}\left(c_{y_i,i}^{-1}\left\langle \boldsymbol{w}_{y_i}, \boldsymbol{z}_i\right\rangle\right) - c_{k,i}\left(c_{k,i}^{-1}\left\langle \boldsymbol{w}_k, \boldsymbol{z}_i\right\rangle\right)\right)\right).$$

Given $\boldsymbol{\xi}_k \in \mathbb{R}^n$, $\boldsymbol{\Xi} = \left(\boldsymbol{\xi}_1^\top, \cdots, \boldsymbol{\xi}_K^\top\right)^\top \in \mathbb{R}^{Kn}$, *and* $\ell(z) = \ln\left(1 + \exp(z)\right)$, *we define*

$$\mathcal{L}(\boldsymbol{\Xi}) \coloneqq \sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right) = \sum_{i=1}^{n}\ln\left(1 + \sum_{k\neq y_i}^{K}\exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right)\right)$$
$$\text{and } \psi(\boldsymbol{\Xi}) \coloneqq \ell^{-1}\left(\mathcal{L}(\boldsymbol{\Xi})\right),$$

where ψ *is individually convex with respect to each* $\boldsymbol{\xi}_k$, *and* β*-smooth with respect to* ℓ_∞ *norm.*

For the loss functions that satisfy Assumption 3, we use the "equal assignment" encoding of the labels, $\alpha = \beta = 1$; for cross-entropy loss under Assumption 4, we use the "simplex representation" encoding of the labels (Lee et al., 2004; Wang et al., 2021a) with $\alpha = \frac{K-1}{K}$ and $\beta = \frac{1}{K}$. In Appendix C.3 we show that the properties of convexity and β-smoothness of ψ carry over to the multiclass case; interestingly, we can only prove *individual convexity* for cross-entropy loss under Assumption 4.

3.1.1 Multiclass Minimum-norm Interpolation

Analogous to the case of binary labels, we define the minimum-norm interpolator (MNI) of multiclass labels as $\boldsymbol{W}_{\mathsf{MNI}} \coloneqq \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k$ where $\boldsymbol{c}_k$ is a specific encoding of the multiclass labels as defined at the beginning of this section. Specifically, gradient descent run with the square loss on labels encoded with the "equal assignment" choice $\alpha = \beta = 1$ would result in what we call the *one-vs-all MNI*, given by $\boldsymbol{W}_{\mathsf{OvA}} \coloneqq \left(\boldsymbol{w}_{\mathsf{OvA},1}^\top, \cdots, \boldsymbol{w}_{\mathsf{OvA},K}^\top\right)^\top \in \mathbb{R}^{Kd}$ where $\boldsymbol{w}_{\mathsf{OvA},k} \coloneqq \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k(1,1)$. Similarly, gradient descent run with the square loss on labels encoded with the "simplex representation" $\alpha = \frac{K-1}{K}$ and $\beta = \frac{1}{K}$ would result in what we call the *simplex MNI*, given by $\boldsymbol{W}_{\mathsf{simplex}} \coloneqq \left(\boldsymbol{w}_{\mathsf{simplex},1}^\top, \cdots, \boldsymbol{w}_{\mathsf{simplex},K}^\top\right)^\top \in \mathbb{R}^{Kd}$ where $\boldsymbol{w}_{\mathsf{simplex},k} \coloneqq \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{c}_k\left(\frac{K-1}{K}, \frac{1}{K}\right)$.

3.2 Main Results

First, we extend the primal-dual framework from Ji and Telgarsky (2021) to the multiclass case. We again use gradient descent to solve this unregularized ERM problem with initialization $\boldsymbol{W}_0$ and update rule: $\boldsymbol{W}_{t+1} \coloneqq \boldsymbol{W}_t - \eta_t\nabla\mathcal{R}(\boldsymbol{W}_t)$ for $t \geq 0$. We denote, in the context of mirror-descent analysis, the "primal" $\boldsymbol{P}_t \coloneqq \boldsymbol{C}\tilde{\boldsymbol{Z}}\boldsymbol{W}_t = \left(\boldsymbol{p}_{t,1}^\top, \cdots, \boldsymbol{p}_{t,K}^\top\right)^\top \in \mathbb{R}^{Kn}$, and its corresponding "dual" $\boldsymbol{Q}_t \coloneqq \nabla\psi(\boldsymbol{P}_t) = \left(\boldsymbol{q}_{t,1}^\top, \cdots, \boldsymbol{q}_{t,K}^\top\right)^\top \in \mathbb{R}^{Kn}$, where $\boldsymbol{p}_{t,k} = \operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right)\boldsymbol{Z}\boldsymbol{w}_{t,k} \in \mathbb{R}^n$ and $\boldsymbol{q}_{t,k} = \nabla_{\boldsymbol{p}_{t,k}}\psi(\boldsymbol{P}_t) \in \mathbb{R}^n$ for all $k \in [K]$ and $t \geq 0$. This concatenated representation together with Assumption 3, (or 4) and Assumption 2 ensure that the setup is identical to that of Ji and Telgarsky (2021). Therefore, we can directly apply their primal-dual result, which we restate below in our notation specific to the multiclass setting.

Lemma 9 *Under Assumption 3, (or 4) and 2, when all t with $\psi\big(\boldsymbol{C}\tilde{\boldsymbol{Z}}\boldsymbol{W}_t\big) \leq 0$, and iteration of gradient descent goes to infinity with $\hat{\eta}_t = \eta_t \ell'\big(\psi\big(\boldsymbol{C}\tilde{\boldsymbol{Z}}\boldsymbol{W}_t\big)\big)/n \leq 1/\beta$ is nonincreasing and $\sum_{t=0}^{\infty} \hat{\eta}_t = \infty$, we have the implicit bias $\bar{\boldsymbol{W}} := \lim_{t\to\infty} \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|_2} = \frac{\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}}{\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}\right\|_2}$, where*

$$\bar{\boldsymbol{Q}} \in \underset{\psi^*(\boldsymbol{Q})\leq 0}{\arg\min}\, F\left(\boldsymbol{Q}\right), \text{ and } F\left(\boldsymbol{Q}\right) := \frac{1}{2}\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}\right\|_2^2 \tag{9}$$

We provide the details of this proof, which is mostly an extension of Ji and Telgarsky (2021), in Appendix B. One subtlety is that we were only able to establish individual convexity in ψ for cross-entropy loss in Assumption 4. Lemma 29 shows that this is sufficient to recover Lemma 9, and joint convexity is only required to prove the tightness of the convergence rates in Ji and Telgarsky (2021).

We now present the main results of this section. We first show that for any multiclass loss satisfying Assumption 3, the implicit bias solution $\bar{\boldsymbol{W}}$ is approximately close to the one-vs-all MNI $\boldsymbol{W}_{\mathsf{OvA}}$. This result is analogous to Theorem 6 which we proved for the binary case.

Theorem 10 *Under Assumptions 3 and 2, consider any value of $\alpha > 0$ satisfying $\frac{\|\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I}\|_2}{\alpha} \leq \frac{1}{3}$. Then, for every class $k \in [K]$, the implicit bias $\bar{\boldsymbol{w}}_k$ converges in direction to $\boldsymbol{w}_{\mathsf{OvA},k}$ at the rate*

$$\left\|\frac{\bar{\boldsymbol{w}}_k}{\|\bar{\boldsymbol{w}}_k\|_2} - \frac{\boldsymbol{w}_{\mathsf{OvA},k}}{\left\|\boldsymbol{w}_{\mathsf{OvA},k}\right\|_2}\right\|_2 \leq \frac{C\left\|\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{c}_k - \alpha\boldsymbol{c}_k\right\|_2}{\alpha\left\|\boldsymbol{c}_k\right\|_2}, \tag{10}$$

where C is a universal constant that does not depend on α, $\boldsymbol{X}$ or $\boldsymbol{c}_k$.

The proof of Theorem 10 is provided in Appendix C.1 and is a simple extension of the proof of Theorem 6. We now state a corollary (analogous to Corollary 7) showing that the canonical high-dimensional ensembles that admit directional convergence in probability of the implicit bias to the MNI on binary labels also do so for the one-vs-all MNI on one-hot-encoded labels.

Corollary 11 *Assume independent and identically distributed data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ such that each covariate satisfies one of the following: a) $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, and we denote the spectrum of $\boldsymbol{\Sigma}$ by $\boldsymbol{\lambda}$;* or *b) $\boldsymbol{x}_i = diag(\boldsymbol{\lambda})^{1/2}\boldsymbol{z}_i$, where $\boldsymbol{z}_i$ has independent entries such that each z_{ij} is mean-zero, unit-variance, and sub-Gaussian with parameter $v > 0$ (i.e. $\mathbb{E}[z_{ij}] = 0, \mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[e^{tz_{ij}}] \leq e^{vt^2/2}$ for all $t \in \mathbb{R}$). In both cases, define the effective dimensions $d_2 := \frac{\|\boldsymbol{\lambda}\|_1^2}{\|\boldsymbol{\lambda}\|_2^2}$ and $d_\infty := \frac{\|\boldsymbol{\lambda}\|_1}{\|\boldsymbol{\lambda}\|_\infty}$ and assume that $d_2 \gg v^2 n$ and $d_\infty \gg vn$. Then, Theorem 10 implies that for each class $k \in [K]$, we have*

$$\left\|\frac{\bar{\boldsymbol{w}}_k}{\|\bar{\boldsymbol{w}}_k\|_2} - \frac{\boldsymbol{w}_{\mathsf{OvA},k}}{\left\|\boldsymbol{w}_{\mathsf{OvA},k}\right\|_2}\right\|_2 \leq C \cdot v \cdot \max\left\{\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right\},$$

with probability at least $1 - 4e^{-cn}$*, where* $C, c > 0$ *are appropriately chosen universal constants. This implies that* $\left\| \frac{\bar{\boldsymbol{w}}_k}{\|\bar{\boldsymbol{w}}_k\|_2} - \frac{\boldsymbol{w}_{\mathsf{OvA},k}}{\|\boldsymbol{w}_{\mathsf{OvA},k}\|_2} \right\|_2$ *is vanishingly small for any high-dimensional ensemble* $\{(n, d, \boldsymbol{\lambda})\}_{n\geq 1}$ *satisfying* $d_2 \gg v^2 n$ *and* $d_\infty \gg vn$*.*

The proof of Corollary 11 is identical to the proof of Corollary 7, only with $\boldsymbol{y}$ replaced by $\boldsymbol{c}_k$; therefore, we omit the details.

The next theorem shows an *exact* equivalence to the simplex MNI for cross-entropy loss under Assumption 4. This result is the multiclass analog of Proposition 4 Part 2.

Theorem 12 *Under Assumption 4, the implicit bias is parallel to the simplex MNI* $\boldsymbol{W}_{\mathsf{simplex}}$ *iff* $\boldsymbol{X}\boldsymbol{X}^\top \succ \boldsymbol{0}$ *and* $\boldsymbol{\beta}_k := \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k$ *satisfies* $c_{k,i}\beta_{k,i} > 0$ *for all* $i \in [n]$ *and* $k \in [K]$*.*

The proof of Theorem 12 is in provided in Appendix C.2. Note that Theorem 12 recovers the exact equivalence condition of Wang et al. (2021a) without using the intermediate multiclass SVM formulation of the implicit bias primal. Interestingly, the convex programs on $\bar{\boldsymbol{q}}_k$ for all $k \in [K]$ that are formulated in the proof of Theorem 12 already contains the novel equality constraints that Wang et al. (2021a) were only able to obtain after applying a non-trivial transformation to the multiclass SVM dual variables. This suggests that the mirror-descent dual is the more natural dual to analyze in the multiclass case.

4. A Converse Result

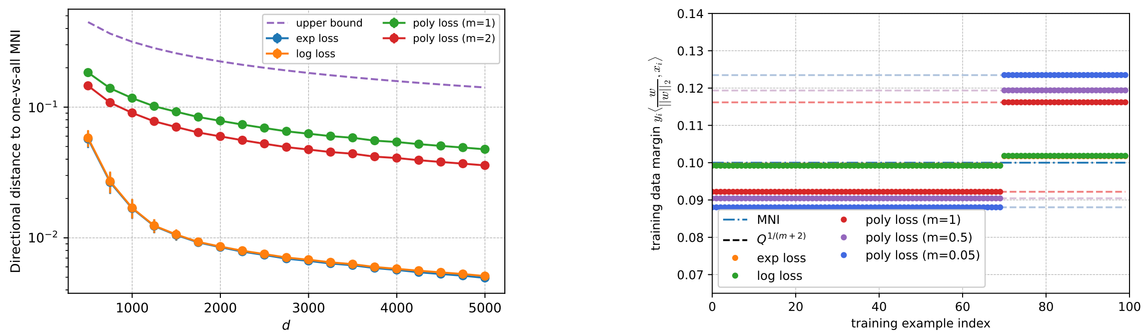
We now show that the condition for exact equivalence in Proposition 4 is necessary. For conciseness, we consider binary labels, but these proofs can easily be extended to the multiclass case.

Proposition 13 *Consider any loss function that satisfies Assumption 1 with a strictly convex function* $g(d) \neq d$*. Define* $h(d) := \left[g^{-1}\right]'(d)$ *and* $f(d) := \frac{h(d)}{d}$*. Then, the following statements are true about the optimal solution* $\bar{\boldsymbol{q}}$ *to the dual convex program* (4)*:*

1. *If* $\boldsymbol{y}$ *is not an exact eigenvector of* $\boldsymbol{X}\boldsymbol{X}^\top$*, then at least two of the entries in* $\bar{\boldsymbol{q}}$ *need to be distinct, i.e.* $\bar{\boldsymbol{q}}$ *cannot be parallel to* $\boldsymbol{1}$*; therefore,* $\bar{\boldsymbol{w}}$ *is not parallel to* $\boldsymbol{w}_{\mathsf{MNI}}$*.*

2. *If* $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{D} = diag(\boldsymbol{d})$*, the primal solution* $\bar{\boldsymbol{w}}$ *interpolates the adjusted labels* $\tilde{y}_i = y_i d_i \cdot f^{-1}\left(\frac{d_i}{\mu}\right)$*, where* $\mu > 0$ *is any solution to the equation* $\sum_{i=1}^n g^{-1}\left(f^{-1}\left(\frac{d_i}{\mu}\right)\right) = 1$*.*

Proposition 13 is proved in Appendix D.1 and also utilizes the relaxed convex program of Lemma 3. The proposition shows that the condition for exact equivalence in Equation (6) only applies to the implicit bias of exponentially-tailed losses, which satisfy Assumption 1 with the identity mapping $g(d) = d$. Moreover, Part 1 of Proposition 4 is a sufficient *and* necessary condition for exact equivalence for any non-exponential loss with a non-identity mapping $g(d) \neq d$. Part 2 of Proposition 13 provides explicit counterexamples in the form of Gram matrices $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{D}$ that can easily be verified to satisfy the SVM equivalence condition $y_i(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} \succ 0$, yet, induce a very different solution from the MNI that interpolates labels adjusted differently *per training example.*



(a) Simulation for multiclass classification using one-vs-all losses (Assumption 3).

(b) Simulation for importance weighting on different loss functions (Corollary 14).

Figure 4: Panel (a) compares the implicit bias of gradient descent to the one-vs-all MNI. The results demonstrate that the directional distance to the MNI is upper bounded by the theoretical guarantee in Theorem 10, analogous to the binary case. The simulation setup is the same as Figure 3b with $K = 5$ classes, and labels drawn uniformly at random in $[K]$. Panel (b) visualizes the *normalized* training data margins induced by importance weighting on different loss functions in Corollary 14. We consider the idealized assumption $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{I}$ with $n = 100$ and $d = 5000$. The first 70 examples are *majority examples* and labeled as $y_i = +1$, and the rest of the 30 examples are *minority examples* labeled as $y_i = -1$. Note that we apply the importance weighting factor $Q = 2.0$ only to the minority examples. We run gradient descent on different loss functions for a minimum of 10^4 iterations, or when the empirical risk falls below 10^{-12}. As predicted by Corollary 14, the margins of exponentially-tailed losses are not impacted by importance weighting and are almost intact to those of the MNI, but polynomially-tailed losses interpolate adjusted labels to different extents depending on the value of m. In Appendix E, we provide corresponding simulations on random data.

To drive home this point, we use Corollary 14 to characterize the impact of the **importance weighting procedure** with polynomial losses. This procedure, parameterized by a subset of underrepresented examples $S \subset [n]$ and weight $Q > 1$ and applied with a loss function $\ell(\cdot)$, minimizes the weighted risk $\mathcal{R}\left(\boldsymbol{w}; (Q, S)\right) \coloneqq \frac{1}{n}\sum_{i=1}^n Q^{\mathbb{I}[i \in S]} \cdot \ell\left(-y_i \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle\right)$. Recently, Wang et al. (2021b) proposed applying this procedure with polynomial losses to address OOD generalization.

Corollary 14 *Consider the idealized data matrix* $\boldsymbol{X}\boldsymbol{X}^\top = \alpha \boldsymbol{I}$ *for some* $\alpha > 0$*, as in Corollary 5. Then, importance weighting with a polynomial loss of degree* m *leads to implicit bias* $\bar{\boldsymbol{w}}$ *that interpolates per-example-adjusted labels* $\tilde{y}_i \propto Q^{\frac{1}{m+2}\cdot \mathbb{I}[i\in S]} y_i$*. We call the implicit bias* $\bar{\boldsymbol{w}}$ *the cost-sensitive MNI.*

Corollary 14 is proved in Appendix D.2 and implies that importance weighting with polynomial losses will interpolate labels that are larger in magnitude on minority points. As shown in Kini et al. (2021); Behnia et al. (2022), this type of *cost-sensitive interpolation* is provably beneficial for OOD generalization. Since $Q > 1$, heavier-tailed polynomial losses (corresponding to smaller values of m) lead to a stronger importance-weighting effect. In Figure 4b, we illustrate how different loss functions influence the training data margins (and also the interpolated adjusted labels) with an identical choice of importance weighting Q. This visualization clearly demonstrates that heavier-tailed losses (e.g. smaller values of m in the polynomially-tailed loss) increase the margin on minority examples. Interestingly, we also observe a corresponding slight *decrease* in the margin on majority examples. This is because we normalized the training data margins (i.e. use the normalized weights of the linear model $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$) in order to provide a fair comparison of the directional differences between solutions. Appendix E shows that similar patterns manifest on randomly generated data, for which Corollary 14 does not apply. One can compare this interpolation to that induced by the **vector-scaling** (VS-loss) (Ye et al., 2020; Kini et al., 2021), defined as a per-example loss function $\ell_{\text{vs}}\left(z_i; (Q, S)\right) := \ln\Big(1 + \exp\left(Q^{\mathbb{I}[i \in S]} z_i\right)\Big)$. Behnia et al. (2022) show[4] that in our high-dimensional regime, this will lead to cost-sensitive interpolation of the adjusted labels $\tilde{y}_i \propto Q^{\mathbb{I}[i \in S]} y_i$, which is in fact a stronger interpolation effect.

Finally, we present converse results on multiclass data that are analogous to Proposition 13 and Corollary 14 respectively.

Proposition 15 *Consider any loss function that satisfies Assumption 3 with a strictly convex function $g(d) \neq d$. Define $h(d) := \left[g^{-1}\right]'(d)$ and $f(d) := \frac{h(d)}{d}$. Then, the following statements are true about the optimal solution $\bar{\boldsymbol{q}}_k$ for $k \in [K]$ to the dual convex program (9):*

1. *If $\boldsymbol{c}_k$ is not an exact eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$, then at least two of the entries in $\bar{\boldsymbol{q}}_k$ need to be distinct, i.e. $\bar{\boldsymbol{q}}_k$ cannot be parallel to $\mathbf{1}$; therefore, $\bar{\boldsymbol{w}}_k$ is not parallel to $\boldsymbol{w}_{\mathsf{OvA},k}$.*
2. *If $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{D} = diag(\boldsymbol{d})$, the primal solution $\bar{\boldsymbol{w}}_k$ interpolates the adjusted labels $\tilde{\boldsymbol{c}}_{k,i} = c_{k,i} d_i \cdot f^{-1}\left(\frac{d_i}{\mu}\right)$ for each $k \in [K]$, where $\mu > 0$ is any solution to the equation $\sum_{i=1}^n \sum_{k=1}^K g^{-1}\Big(f^{-1}\left(\frac{d_i}{\mu}\right)\Big) = 1$.*

Corollary 16 *Consider the idealized data matrix $\boldsymbol{X}\boldsymbol{X}^\top = \alpha \boldsymbol{I}$ for some $\alpha > 0$. Then, importance weighting with a polynomial loss of degree m leads to implicit bias $\bar{\boldsymbol{w}}_k$ that interpolates per-example-adjusted labels $\tilde{\boldsymbol{c}}_{k,i} \propto Q^{\frac{1}{m+2} \cdot \mathbb{I}[\{k,i\} \in S]} c_{k,i}$ for each $k \in [K]$.*

The proof of Proposition 15 is identical to the proof of Proposition 13, and the proof of Corollary 16 is identical to the proof of Corollary 14, since they analyze the same characteristic equation—Equation (67a) in the binary case and Equation (44a) in multiclass case with $\boldsymbol{y}$ replaced by $\boldsymbol{c}_k$ for each class $k \in [K]$. Therefore, we omit the details.

4.1 Lower Bound on Directional Convergence Between $\bar{q}$ and 1

The preceding Proposition 13 addressed the question of tightness of our *exact* equivalence theorem (Theorem 4). This section addresses whether we can obtain a lower bound on the approximation error that matches Theorem 6. We show that we can obtain a lower bound

4. This result is also recoverable in our framework, although we omit the details for brevity.

on the approximation error for loss functions with homogeneous function $h(z)$ such that, in some sense, "matches" our upper bound.

Proposition 17 *Consider any loss function $\ell(z)$ satisfying Assumption 1, and additionally assume that its corresponding function $h(q) := \left[g^{-1}\right]'(q)$ is a homogeneous function, i.e. $h(ab) = a^{\gamma} h(b)$ for $a, b \geq 0$ and $\gamma \in \mathbb{R}$. Further, assume that $\left\| \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{\mathbf{1}}{\sqrt{n}} \right\|_2 \leq \frac{\delta}{\sqrt{n}}$ for some $\delta \in (0,1)$. Then, the dual implicit bias is lower bounded (in its directional distance from the dual-MNI) as*

$$\left\| \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{\mathbf{1}}{\sqrt{n}} \right\|_2 \geq \frac{1}{2\sqrt{n}} \min_{\alpha>0} \min \left\{ \frac{\left\| \boldsymbol{X}\boldsymbol{X}^{\top}\boldsymbol{y} - \alpha \boldsymbol{y} \right\|_2}{k\alpha}, \frac{\left\| \boldsymbol{X}\boldsymbol{X}^{\top}\boldsymbol{y} - \alpha \boldsymbol{y} \right\|_2}{\left\| \boldsymbol{X}\boldsymbol{X}^{\top} \right\|_2} \right\}, \tag{11}$$

where $k = \max\left(\frac{\bar{h}(1-\delta)-1}{\delta}, \frac{1-\bar{h}(1+\delta)}{\delta} \right)$, and $\bar{h}(z) = ah(z)$ for some $a > 0$ such that $\bar{h}(1) = 1$.

Proposition 17 is proved in Appendix D.3. We first remark on the sense in which Equation (11) is tight with respect to the upper bound in Theorem 6. If the best value of α is one for which $\left\| \boldsymbol{X}\boldsymbol{X}^{\top} \right\|_2 \leq \frac{4\alpha}{3}$ (which is the assumption made in Theorem 6), then the lower bound becomes $\min \left\{ \frac{\left\| \boldsymbol{X}\boldsymbol{X}^{\top}\boldsymbol{y} - \alpha \boldsymbol{y} \right\|_2}{2k\alpha\|\boldsymbol{y}\|_2}, \frac{3\left\| \boldsymbol{X}\boldsymbol{X}^{\top}\boldsymbol{y} - \alpha \boldsymbol{y} \right\|_2}{8\alpha\|\boldsymbol{y}\|_2} \right\}$, which matches the upper bound (Equation 7) up to the constant factor k. Next, we briefly comment on the extra assumptions appearing in the proposition, starting with the assumption of homogeneity on $h(q)$. In particular, the special case of polynomial loss has $h(q) = \frac{m}{m+1} q^{\frac{-1}{m+1}}$ which is a homogeneous function; therefore, Proposition 17 applies. We also comment on the requirement that $\left\| \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{\mathbf{1}}{\sqrt{n}} \right\|_2 \leq \frac{\delta}{\sqrt{n}}$ for some $\delta \in (0,1)$. Note that Corollary 7 directly implies that this condition would be satisfied w.h.p. if $d_2 \gg v^2 n^2$ and $d_\infty \gg v n^{\frac{3}{2}}$; i.e. under a very high-dimensional regime. We believe that the extra $\frac{1}{\sqrt{n}}$ factor in the upper bound above is not required, and could be removed if one were able to show that all entries of the directional error vector $\frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{\mathbf{1}}{\sqrt{n}}$ were within constant factors of one another. Showing this (and, relatedly, providing tight upper and lower bounds on the ℓ_∞-directional error) is an important direction for future work.

Finally, we present a corollary (analogous to Proposition 17) that lower bounds the approximation error for multiclass losses under Assumption 3.

Corollary 18 *Consider any multiclass loss function satisfying Assumption 3, and additionally assume that its corresponding function $h(q) := \left[g^{-1}\right]'(q)$ is a homogeneous function, i.e. $h(ab) = a^{\gamma} h(b)$ for $a, b \geq 0$ and $\gamma \in \mathbb{R}$. Further, assume that $\left\| \frac{\bar{\boldsymbol{q}}_k}{\|\bar{\boldsymbol{q}}_k\|_2} - \frac{\mathbf{1}}{\sqrt{n}} \right\|_2 \leq \frac{\delta}{\sqrt{n}}$ for some $\delta \in (0,1)$ for all $k \in [K]$. Then, the dual implicit bias for each class k is lower bounded (in its directional distance from the dual-MNI) as*

$$\left\| \frac{\bar{\boldsymbol{q}}_k}{\|\bar{\boldsymbol{q}}_k\|_2} - \frac{\mathbf{1}}{\sqrt{n}} \right\|_2 \geq \frac{1}{2\sqrt{n}} \min_{\alpha>0} \min \left\{ \frac{\left\| \boldsymbol{X}\boldsymbol{X}^{\top}\boldsymbol{c}_k - \alpha \boldsymbol{c}_k \right\|_2}{t\alpha}, \frac{\left\| \boldsymbol{X}\boldsymbol{X}^{\top}\boldsymbol{c}_k - \alpha \boldsymbol{c}_k \right\|_2}{\left\| \boldsymbol{X}\boldsymbol{X}^{\top} \right\|_2} \right\}, \tag{12}$$

where $t = \max\left(\frac{\bar{h}(1-\delta)-1}{\delta}, \frac{1-\bar{h}(1+\delta)}{\delta}\right)$, *and* $\bar{h}(z) = ah(z)$ *for some* $a > 0$ *such that* $\bar{h}(1) = 1$.

The proof of Corollary 18 is identical to the proof of Proposition 17, as it analyzes the same characteristic equation—Equation (67a) in the binary case and Equation (44a) in multiclass case with $\boldsymbol{y}$ replaced by $\boldsymbol{c}_k$ for each class $k \in [K]$. Therefore, we omit the details.

5. Discussion

Our results show that once we move away from the exponentially-tailed family of losses, general losses exhibit a variety of influence on the eventual solution, with similarities for "in-distribution"-oriented loss functions but differences for "out-of-distribution"-oriented loss functions. We believe that these results show the potential of the primal-dual framework to study closed-form properties of the implicit bias. It would be interesting to provide similar closed-form characterizations for the implicit bias of other optimization algorithms and/or for nonlinear models. Specific to linear models and gradient descent, there are still many open questions. Based on converse results in Hsu et al. (2021); Ardeshir et al. (2021) for exponential losses, the effective overparameterization conditions in Corollary 7 appear necessary for asymptotic directional convergence of the implicit bias to MNI. However, whether Theorem 6 provides the optimal rate of convergence (beyond the partial converse result in Proposition 17) is unclear. Also, of interest is whether it is possible to obtain results similar to Propositions 4 and Theorem 6 under even fewer assumptions on losses, such as in Ji et al. (2020); Bartlett et al. (2006). Finally, we are interested in using these closed-form characterizations to obtain tight non-asymptotic bounds on the test risk.

Acknowledgments

We gratefully acknowledge the support of the ARC-ACO Fellowship provided by Georgia Tech, the NSF (through awards CCF-2239151 and award IIS-2212182), an Adobe Data Science Research Award, and an Amazon Research Award.

Appendix

Table of Contents

Appendix A. Proofs of All Binary Results

In this section, we include all the detailed proofs for our analysis for the binary case.

A.1 Proof of Lemma 1 (Existence of g Function)

In this section, we prove Lemma 1, which establishes the existence of a function $g(\cdot)$ that is strictly increasing, convex, and a function of the original loss function $\ell(\cdot)$. To do so, we first introduce a different co-convergent sequence $\{\bar{g}_a(\cdot)\}_{a>0}$. We start with a lemma proving the existence of its limit, which implies the existence of the limit of the original sequence of functions of interest, $g_a(z) := \frac{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)}$.

Lemma 19 *Under Assumption 1, define two sequences* $g_a(z) := \frac{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)}$ *and* $\bar{g}_a(z) := \frac{\ell^{-1}(a)\cdot z}{\ell^{-1}(a\cdot z)}$, *for* $0 < a < \ell(0)$ *and* $z \in (0,1]$. *Then, these two sequences converge to the same limit, i.e.,* $\lim_{a\to 0} g_a(z) = \lim_{a\to 0} \bar{g}_a(z)$ *for all* $z \in (0,1]$.

Proof (of Lemma 19) We first show that both sequences are equivalent in the limit through the following chain of equalities

$$\lim_{a\to 0} \bar{g}_a(z) = \lim_{a\to 0} \frac{\ell^{-1}(a)\cdot z}{\ell^{-1}(a\cdot z)} = \lim_{a\to 0} \frac{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)} = \lim_{a\to 0} g_a(z).$$

Above, the second equality follows from l'Hospital's rule (because $\lim_{a\to 0} \frac{\ell^{-1}(a)\cdot z}{\ell^{-1}(a\cdot z)}$ is in indeterminate form). Therefore, it suffices to show the existence of the limit of $\bar{g}_a(z)$ as $a \to 0$. To do this, we will show that $\bar{g}_a(z)$ is decreasing in a as well as bounded above for all $a > 0$.

Following Ji and Telgarsky (2021, Lemma 6), we define the function $\sigma(s) := \ell'\left(\ell^{-1}(s)\right)\cdot\ell^{-1}(s)$, where Parts 1 and 2 of Assumption 1 together imply that $\lim_{s\to 0} \sigma(s) = 0$ and the function $\sigma(s)/s$ is increasing in $s \in \left(0, \ell(0)\right)$. Additionally, since $\ell^{-1}(s) < 0$ for $s \in \left(0, \ell(0)\right)$, $\sigma(s)/s$ is non-positive in $s \in \left(0, \ell(0)\right)$. Using these properties, we will show that $\bar{g}_a(z)$ is decreasing in $a \in \left(0, \ell(0)\right)$ by showing that $\frac{\partial \bar{g}_a(z)}{\partial a} \leq 0$. In particular, we have

$$\begin{aligned}
\frac{\partial \bar{g}_a(z)}{\partial a} &= \frac{\frac{\ell^{-1}(a\cdot z)\cdot z}{\ell'\left(\ell^{-1}\left(a\right)\right)} - \frac{\ell^{-1}(a)\cdot z^2}{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}}{\left[\ell^{-1}(a\cdot z)\right]^2} \\
&= \frac{\frac{az^2}{\ell'\left(\ell^{-1}\left(a\right)\right)\cdot\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}}{\left[\ell^{-1}(a\cdot z)\right]^2} \cdot \left(\frac{\ell^{-1}(a\cdot z)\cdot\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}{a\cdot z} - \frac{\ell^{-1}(a)\cdot\ell'\left(\ell^{-1}\left(a\right)\right)}{a}\right) \\
&= \frac{\frac{az^2}{\ell'\left(\ell^{-1}\left(a\right)\right)\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}}{\left[\ell^{-1}(a\cdot z)\right]^2} \cdot \left(\frac{\sigma(a\cdot z)}{a\cdot z} - \frac{\sigma(a)}{a}\right) \leq 0,
\end{aligned}$$

where the last step follows for $z \in (0,1]$ as the function $\sigma(s)/s$ is increasing in s. Finally, since $\ell'\big(\ell^{-1}(\cdot)\big)$ is an increasing function, we have $\frac{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)} \leq 1$ for all $a \in \big(0, \ell\left(0\right)\big)$ and $z \in (0,1]$, meaning that $\bar{g}_a\left(z\right) \leq 1$.

Thus, we have shown that $\bar{g}_a\left(z\right)$ is decreasing in a (therefore, increasing as $a \downarrow 0$) and is bounded above by 1. By the monotone convergence theorem, its limit exists as $a \to 0$. This completes the proof of the lemma. ■

Armed with Lemma 19, we now provide the proof of Lemma 1.

Proof (of Lemma 1) We showed in Lemma 19 that the limit of the sequence of functions $\{g_a(z)\}_{a>0}$ exists as $a \to 0$. Accordingly, we define $g\left(z\right) \coloneqq \lim\limits_{a\to 0}\frac{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)}$ for $z \in (0,1]$. It remains for us to show that $g\left(z\right)$ is strictly increasing and convex. We first show that the derivative of $g\left(z\right)$ exists. For this, we reuse the co-convergent sequence defined in Lemma 19, i.e. $\bar{g}_a\left(z\right) \coloneqq \frac{\ell^{-1}(a)\cdot z}{\ell^{-1}(a\cdot z)}$. Specifically, we want to show that

$$g'\left(z\right) = \frac{\partial}{\partial z}\lim_{a\to 0}\frac{\ell^{-1}\left(a\right)\cdot z}{\ell^{-1}\left(a\cdot z\right)} = \lim_{a\to 0}\frac{\partial}{\partial z}\frac{\ell^{-1}\left(a\right)\cdot z}{\ell^{-1}\left(a\cdot z\right)} = \lim_{a\to 0}\bar{g}_a'\left(z\right), \tag{13}$$

where in the above, the second equality will hold if $\lim\limits_{a\to 0}\bar{g}_a'\left(z\right)$ converges uniformly. To show this, we provide a direct calculation as below

$$\begin{aligned}
\bar{g}_a'\left(z\right) &= \frac{\ell^{-1}\left(a\right)\cdot\ell^{-1}\left(a\cdot z\right) - \ell^{-1}\left(a\right)\cdot z\cdot\frac{a}{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)}}{\left[\ell^{-1}\left(a\cdot z\right)\right]^2} \\
&= \frac{\bar{g}_a\left(z\right)}{z}\left(1 - \frac{a\cdot z}{\ell'\left(\ell^{-1}\left(a\cdot z\right)\right)\cdot\ell^{-1}\left(a\cdot z\right)}\right) \\
&= \frac{\bar{g}_a\left(z\right)}{z}\left(1 - \frac{a\cdot z}{\sigma\left(a\cdot z\right)}\right),
\end{aligned}$$

where we defined the function $\sigma\left(s\right)$ in the proof of Lemma 19. Since the function $\sigma\left(s\right)/s$ is increasing on $s \in \big(0, \ell\left(0\right)\big)$ and non-positive, the function $s/\sigma\left(s\right)$ is decreasing in s (therefore, increasing as $s \downarrow 0$), non-positive and bounded above by 0. By the monotone convergence theorem, we then have $\lim\limits_{a\to 0}\frac{a\cdot z}{\sigma(a\cdot z)} = \lim\limits_{s\to 0}\frac{s}{\sigma(s)} = c \leq 0$. As a result, we have

$$\begin{aligned}
\lim_{a\to 0}\bar{g}_a'\left(z\right) &= \lim_{a\to 0}\frac{\bar{g}_a\left(z\right)}{z}\left(1 - \frac{a\cdot z}{\sigma\left(a\cdot z\right)}\right) \\
&= \lim_{a\to 0}\frac{\bar{g}_a\left(z\right)}{z}\cdot\lim_{a\to 0}\left(1 - \frac{a\cdot z}{\sigma\left(a\cdot z\right)}\right) \\
&= \frac{g\left(z\right)}{z}\cdot\left(1-c\right).
\end{aligned} \tag{14}$$

Therefore, Equation (13) holds with $g'\left(z\right) = \frac{(1-c)g(z)}{z}$. Because $c \leq 0$, $g(z) > 0$ and $z > 0$, we can conclude that $g'\left(z\right) > 0$ and so $g(\cdot)$ is a strictly increasing function. It remains to show convexity. Using a similar procedure to the above, we can show that

$$
\begin{aligned}
g''(z) &= \frac{\partial}{\partial z} \lim_{a\to 0} \bar{g}'_a(z) \\
&= \frac{\partial}{\partial z} \frac{(1-c)\, g(z)}{z} = \frac{(1-c)\big(g'(z)\cdot z - g(z)\big)}{z^2} = g(z)\left(\frac{(1-c)(-c)}{z^2}\right).
\end{aligned}
$$

Finally, since $c \leq 0$ and $g(z)$ is non-negative, we can conclude that $g''(z) \geq 0$ and therefore the function $g(\cdot)$ is convex. This completes the proof of the lemma. ■

A.2 Proof of Lemma 3 (Auxiliary Convex Program)

In this section, we prove Lemma 3. The proof follows a two-step procedure. First, we show that the new equality constraint in (5) is sufficient to imply the original equality constraint in (4). Second, we show that any solution to the original program also satisfies the new equality constraint; therefore, the solution sets of both programs coincide. The following lemma demonstrates the first part of this reasoning.

Lemma 20 *Under Assumption 1, for any* $\boldsymbol{q} \in \mathbb{R}^n$ *such that* $q_i = g(z_i) = \lim_{a\to 0} \frac{\ell'\left(\ell^{-1}\left(a\cdot z_i\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)} > 0$*, for all* $i \in [n]$ *with* $z_i \in (0,1]$ *and* $\sum_{i=1}^n z_i = 1$*, it implies* $\psi^*(\boldsymbol{q}) = 0$*.*

Proof (of Lemma 20) In this proof, we substantially apply the convex analysis in the astral space introduced in Dudík et al. (2022). Informally, astral space $\overline{\mathbb{R}^n}$ consists of the union of the set $\mathbb{R}^n$ and all "astral points", i.e. n-dimensional points at infinity. Accordingly, we define the astral extension function to take into account of astral points naturally (Dudík et al., 2022, Chapter 7); e.g. $\overline{\ell} : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$.

Using this framework, we show that $\boldsymbol{q}$ is a subdifferential of the astral extension of ψ for some $\overline{\boldsymbol{p}} \in \overline{\mathbb{R}^n}$, and then the astral convex conjugate (which is equivalent to the original convex conjugate) is equal to zero. We start with writing $\boldsymbol{q}$ in the astral format. Since q_i is finite and continuous in a, we can take the limit inside the function, obtaining

$$
q_i = \lim_{a\to 0} \frac{\ell'\big(\ell^{-1}(a\cdot z_i)\big)}{\ell'\big(\ell^{-1}(a)\big)} = \frac{\ell'\Big(\ell^{-1}\Big(\lim_{a\to 0} a\cdot z_i\Big)\Big)}{\ell'\Big(\ell^{-1}\Big(\lim_{a\to 0} a\Big)\Big)} = \frac{\overline{\ell}'\Big(\overline{\ell}^{-1}\Big(\lim_{a\to 0} a\cdot z_i\Big)\Big)}{\overline{\ell}'\Big(\overline{\ell}^{-1}\Big(\lim_{a\to 0} a\Big)\Big)}, \tag{15}
$$

where in the last equality, we replace the original functions with their astral extensions. Next, we define an astral point $\overline{\boldsymbol{p}} \in \overline{\mathbb{R}^n}$ such that $\overline{p}_i = \lim_{a\to 0} \overline{\ell}^{-1}(a\cdot z_i) = \overline{\ell}^{-1}\Big(\lim_{a\to 0} a\cdot z_i\Big)$ for all $i \in [n]$. This also implies $\sum_{i=1}^n \overline{\ell}(\overline{p}_i) = \lim_{a\to 0} a$. Substituting these values in Equation (15), we can write q_i as

$$
q_i = \frac{\overline{\ell}'(\overline{p}_i)}{\overline{\ell}'\Big(\overline{\ell}^{-1}\Big(\sum_{i=1}^n \overline{\ell}(\overline{p}_i)\Big)\Big)}
$$

for all $i \in [n]$. On the other hand, according to Lemma 1, we can also define q_i in the limit of a different co-convergent sequence as

$$q_i = \lim_{a\to 0} \frac{\ell^{-1}(a)\cdot z_i}{\ell^{-1}(a\cdot z_i)} = \frac{\ell^{-1}\left(\lim\limits_{a\to 0} a\right)\cdot z_i}{\ell^{-1}\left(\lim\limits_{a\to 0} a\cdot z_i\right)} = \frac{\overline{\ell}^{-1}\left(\lim\limits_{a\to 0} a\right)\cdot z_i}{\overline{\ell}^{-1}\left(\lim\limits_{a\to 0} a\cdot z_i\right)} = \frac{\overline{\ell}^{-1}\left(\sum_{i=1}^n \overline{\ell}(\overline{p}_i)\right)\cdot z_i}{\overline{p}_i} \tag{16}$$

for all $i \in [n]$. Next, we show that $\boldsymbol{q}$ is a subdifferential of $\overline{\psi}(\overline{\boldsymbol{p}})$. By the definition of $\overline{\psi}: \overline{\mathbb{R}^n} \to \overline{\mathbb{R}}$ and for any $\boldsymbol{p} \in \overline{\mathbb{R}^n}$, we have

$$\overline{\psi}(\boldsymbol{p}) = \overline{\ell}^{-1}\left(\sum_{i=1}^n \overline{\ell}(p_i)\right), \text{ and } \frac{\partial}{\partial p_i}\overline{\psi}(\boldsymbol{p}) = \frac{\overline{\ell}'(p_i)}{\overline{\ell}'\left(\overline{\ell}^{-1}\left(\sum_{i=1}^n \overline{\ell}(p_i)\right)\right)},$$

for all $i \in [n]$. Therefore, according to Equation (15), it implies that $\boldsymbol{q}$ is in the subdifferential of $\overline{\psi}(\overline{\boldsymbol{p}})$ such that $\boldsymbol{q} \in \partial\overline{\psi}(\overline{\boldsymbol{p}})$. Hence, we can further apply the property of Fenchel–Young inequality in the convex conjugate, obtaining

$$\overline{\psi}^*(\boldsymbol{q}) = \sup_{\boldsymbol{p}\in\overline{\mathbb{R}^n}} \langle \boldsymbol{p}, \boldsymbol{q}\rangle - \overline{\psi}(\boldsymbol{p}) = \langle \overline{\boldsymbol{p}}, \boldsymbol{q}\rangle - \overline{\psi}(\overline{\boldsymbol{p}}).$$

As a result, by Equation (16) and the definition of $\overline{\psi}(\cdot)$, we can write

$$\overline{\psi}^*(\boldsymbol{q}) = \langle \overline{\boldsymbol{p}}, \boldsymbol{q}\rangle - \overline{\psi}(\overline{\boldsymbol{p}}) = \sum_{i=1}^n \overline{p}_i \cdot \frac{\overline{\ell}^{-1}\left(\sum_{i=1}^n \overline{\ell}(\overline{p}_i)\right)\cdot z_i}{\overline{p}_i} - \overline{\ell}^{-1}\left(\sum_{i=1}^n \overline{\ell}(\overline{p}_i)\right) = 0.$$

Finally, by Dudík et al. (2022, Proposition 8.5), we have $\psi^*(\boldsymbol{q}) = \overline{\psi}^*(\boldsymbol{q}) = 0$. This completes the proof of the lemma.

■

Next, we show in the following lemma that $\bar{\boldsymbol{q}} = \lim_{t\to\infty} \boldsymbol{q}_t$ under Assumption 2.

Lemma 21 *Under Assumption 1 and 2, the gradient descent dual variable $\boldsymbol{q}_t$ converges to $\bar{\boldsymbol{q}}$ when $t \to \infty$. It gives $\bar{q}_i = \lim_{t\to\infty} q_{t,i} = g(z_i) > 0$ for all $i \in [n]$ with some $z_i \in (0,1]$ and $\sum_{i=1}^n z_i = 1$.*

Proof (of Lemma 21) According to Ji and Telgarsky (2021, Theorem 5), we have $\lim\limits_{t\to\infty} \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y})\, \boldsymbol{q}_t = \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}}$, and $\boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}}$ is the same for all $\bar{\boldsymbol{q}} \in \arg\min\limits_{\psi^*(\boldsymbol{q})\le 0} f(\boldsymbol{q})$. Based on the definition of $q_{t,i}$ in Equation (2) and considering $\ell'(\cdot)$ is an increasing function with $p_{t,i} \le \psi(\boldsymbol{p}_t)$ for all $i \in [n]$ and $t \ge 0$, it follows that $0 < q_{t,i} \le 1$; hence, $\lim\limits_{t\to\infty} \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y})\, \boldsymbol{q}_t = \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y}) \lim\limits_{t\to\infty} \boldsymbol{q}_t = \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}}$. Next, by Assumption 2, $\boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y})$ has full column rank, and $\mathrm{diag}(\boldsymbol{y})\, \boldsymbol{Z}\boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y}) \succ \boldsymbol{0}$. Therefore, we can multiply the pseudo-inverse of $\boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y})$ on both sides of $\boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y}) \lim\limits_{t\to\infty} \boldsymbol{q}_t = \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}}$, which implies $\lim\limits_{t\to\infty} \boldsymbol{q}_t = \bar{\boldsymbol{q}}$. Next, by the definition of $q_{t,i}$ in Equation (2) and the primal convergence in Ji and Telgarsky (2021, Theorem 1) such that $\lim_{t\to\infty} \sum_{i=1}^n \ell\left(p_{t,i}\right) = 0$, we have

$$\lim_{t\to\infty} q_{t,i} = \lim_{t\to\infty} \frac{\ell'\big(\ell^{-1}\big(\ell\left(p_{t,i}\right)\big)\big)}{\ell'\big(\ell^{-1}\big(\sum_{i=1}^n \ell\left(p_{t,i}\right)\big)\big)} = \lim_{a\to 0} \frac{\ell'\big(\ell^{-1}\big(a\cdot z_i\big)\big)}{\ell'\big(\ell^{-1}\big(a\big)\big)} = g(z_i),$$

where we let $a = \sum_{i=1}^n \ell\left(p_{t,i}\right)$, $\ell\left(p_{t,i}\right) = a \cdot z_i$, and $z_i = \lim_{t\to\infty} \frac{\ell(p_{t,i})}{\sum_{i=1}^n \ell(p_{t,i})}$. Finally, Assumption 1 guarantees that $\bar{q}_i = g(z_i) > 0$. This completes the proof of the lemma. ■

Armed with Lemmas 20 and 21, we can prove Lemma 3.

Proof (of Lemma 3) We start with the original convex program in (4)

$$\bar{\boldsymbol{q}} \in \arg\min_{\psi^*(\boldsymbol{q})\le 0} f\left(\boldsymbol{q}\right).$$

By complementary slackness in the KKT conditions, if the constraint is inactive such that $\psi^*\left(\boldsymbol{q}\right) < 0$, we get an invalid solution $\bar{\boldsymbol{q}} = \mathbf{0}$. Therefore, the constraint is active and $\bar{\boldsymbol{q}}$ satisfies $\psi^*\left(\boldsymbol{q}\right) = 0$. In other words, we can write

$$\bar{\boldsymbol{q}} \in \arg\min_{\psi^*(\boldsymbol{q}) = 0} f\left(\boldsymbol{q}\right). \tag{17}$$

Now, Lemma 21 directly implies that $\bar{\boldsymbol{q}}$ must satisfy $\sum_{i=1}^n g^{-1}(\bar{q}_i) = 1$ and $\bar{q}_i > 0$ for all $i \in [n]$. This means that we can further tighten (17) to obtain

$$\begin{aligned} &\bar{\boldsymbol{q}} \in \arg\min_{\boldsymbol{q}\in\mathbb{R}^n} f\left(\boldsymbol{q}\right) && (18)\\ \text{subject to} \quad & \psi^*\left(\boldsymbol{q}\right) = 0, \\ & -q_i < 0 \qquad \text{for all } i \in [n], \\ \text{and} \quad & 1 - \sum_{i=1}^n g^{-1}\left(q_i\right) = 0. \end{aligned}$$

Next, Lemma 20 tells us that $1 - \sum_{i=1}^n g^{-1}\left(q_i\right) = 0 \implies \psi^*\left(\boldsymbol{q}\right) = 0$, meaning that the constraint $\psi^*\left(\boldsymbol{q}\right) = 0$ is redundant and can simply be omitted, leading to the simplified program

$$\begin{aligned} &\bar{\boldsymbol{q}} \in \arg\min_{\boldsymbol{q}\in\mathbb{R}^n} f\left(\boldsymbol{q}\right) && (19)\\ \text{subject to} \quad & -q_i < 0 \qquad \text{for all } i \in [n], \\ \text{and} \quad & 1 - \sum_{i=1}^n g^{-1}\left(q_i\right) = 0. \end{aligned}$$

The final step is to derive an auxiliary convex program

$$\begin{aligned} \boldsymbol{q}^\star \in \underset{\boldsymbol{q}\in\mathbb{R}^n}{\arg\min}\ & \underbrace{\frac{1}{2}\boldsymbol{q}^\top \mathrm{diag}\left(\boldsymbol{y}\right) \boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}\left(\boldsymbol{y}\right) \boldsymbol{q}}_{f(\boldsymbol{q})} \\ \text{subject to} \quad & -q_i < 0 \qquad \text{for all } i \in [n], \\ \text{and} \quad & 1 - \sum_{i=1}^n g^{-1}\left(q_i\right) \leq 0. \end{aligned} \tag{20}$$

Note that in the above, we have relaxed the equality constraint $1 - \sum_{i=1}^n g^{-1}\left(q_i\right) = 0$ to an inequality constraint, $1 - \sum_{i=1}^n g^{-1}\left(q_i\right) \leq 0$.

To complete the proof, it remains to show that any optimal solution to (20) satisfies $\sum_{i=1}^n g^{-1}\left(q_i\right) = 1$. From (19), this directly implies that the set of optima of (4) and (20) are identical. We now show this final step. It is necessary and sufficient for any optimal solution $\boldsymbol{q}^\star$ to the auxiliary convex program (20) to satisfy its KKT conditions, listed below

$$-q_i < 0 \qquad \text{for all } i \in [n], \tag{21a}$$

$$1 - \sum_{i=1}^n g^{-1}\left(q_i\right) \leq 0, \tag{21b}$$

$$\lambda_i \geq 0 \qquad \text{for all } i \in [n], \tag{21c}$$

$$\mu \geq 0, \tag{21d}$$

$$-\lambda_i q_i = 0 \qquad \text{for all } i \in [n], \tag{21e}$$

$$\mu\left(1 - \sum_{i=1}^n g^{-1}\left(q_i\right)\right) = 0, \tag{21f}$$

$$\mathrm{diag}\left(\boldsymbol{y}\right) \boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}\left(\boldsymbol{y}\right) \boldsymbol{q} - \boldsymbol{\lambda} - \mu\left[g^{-1}\right]'\left(\boldsymbol{q}\right) = \boldsymbol{0}, \tag{21g}$$

where $\left[g^{-1}\right]'\left(\boldsymbol{q}\right) := \left(\left[g^{-1}\right]'\left(q_1\right), \cdots, \left[g^{-1}\right]'\left(q_n\right)\right)^\top$. First, we claim that any optimal solution $\boldsymbol{q}^\star$ needs to satisfy $1 - \sum_{i=1}^n g^{-1}\left(q_i^\star\right) = 0$. This follows because we need to set $\mu > 0$ for a valid solution; together with Equation (21f) this implies that we need $1 - \sum_{i=1}^n g^{-1}\left(q_i^\star\right) = 0$. To see why we need to set $\mu > 0$, consider the alternative choice $\mu = 0$. Note that Equations (21a) and (21e) together also require $\boldsymbol{\lambda} = \boldsymbol{0}$. Equation (21g) would then become

$$\boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}\left(\boldsymbol{y}\right) \boldsymbol{q}^\star = \boldsymbol{0} \iff \mathrm{diag}\left(\boldsymbol{y}\right) \boldsymbol{q}^\star = \boldsymbol{0} \iff \boldsymbol{q}^\star = \boldsymbol{0},$$

where the first *iff* statement follows because we have assumed that $\boldsymbol{X}\boldsymbol{X}^\top \succ \boldsymbol{0}$. However, this $\boldsymbol{q}^\star$ is not a valid solution as it violates Equation (21a). Hence, we can conclude both $\bar{\boldsymbol{q}}$ and $\boldsymbol{q}^\star$ satisfy $1 - \sum_{i=1}^n g^{-1}\left(q_i\right) = 0$, and $\boldsymbol{q}^\star = \bar{\boldsymbol{q}}$. This completes the proof of the lemma. ■

A.3 Proof of Proposition 4 (Exact Equivalence to MNI)

Proof (of Proposition 4) The proof of Proposition 4 is divided into two parts.

Proof of Part 1. By Lemma 3 it suffices to characterize an optimal solution to the relaxed convex program (5), reproduced below.

$$\bar{\boldsymbol{q}} = \arg\min_{\boldsymbol{q}\in\mathbb{R}^n} \underbrace{\frac{1}{2}\boldsymbol{q}^\top \operatorname{diag}(\boldsymbol{y})\, \boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\, \boldsymbol{q}}_{f(\boldsymbol{q})}$$

$$\text{subject to} \qquad -q_i < 0 \qquad \text{for all } i \in [n],$$

$$\text{and} \quad 1 - \sum_{i=1}^{n} g^{-1}(q_i) \leq 0.$$

Any optimal solution must satisfy the KKT conditions for this convex program, listed in Equation (21). Let $k > 0$ be the positive eigenvalue corresponding to the exact eigenvector $\boldsymbol{y}$, i.e. we consider $\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y} = k\boldsymbol{y}$. We choose the candidate solution $\bar{\boldsymbol{q}} = g\left(\frac{1}{n}\right)\mathbf{1} \propto \mathbf{1}$, and verify that it satisfies all the KKT conditions below.

- The primal feasibility equations, Equations (21a) and (21b), are satisfied because $g\left(\frac{1}{n}\right) > 0$ (as $g(\cdot)$ is non-negative), and $1 - \sum_{i=1}^{n} g^{-1}(q_i) = 1 - \sum_{i=1}^{n} \frac{1}{n} = 0$.

- The dual feasibility equations, Equations (21c) and (21d), are satisfied by setting $\boldsymbol{\lambda} = \mathbf{0}$ and $\mu = \frac{kg\left(\frac{1}{n}\right)}{\left[g^{-1}\right]'\left(g\left(\frac{1}{n}\right)\right)}$.

- The complementary slackness equations, Equations (21e) and (21f), are satisfied because $\boldsymbol{\lambda} = \mathbf{0}$ and $1 - \sum_{i=1}^{n} \left[g^{-1}\right]'(q_i) = 0$.

- The stationary condition, Equation (21g), is satisfied because

$$\begin{aligned}
&\operatorname{diag}(\boldsymbol{y})\, \boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\, \boldsymbol{q} - \boldsymbol{\lambda} - \mu \left[g^{-1}\right]'(\boldsymbol{q}) \\
&= kg\left(\frac{1}{n}\right) \cdot \operatorname{diag}(\boldsymbol{y})\, \boldsymbol{y} - \mu \left[g^{-1}\right]'\left(g\left(\frac{1}{n}\right)\right) \cdot \mathbf{1} \\
&= kg\left(\frac{1}{n}\right) \mathbf{1} - kg\left(\frac{1}{n}\right) \mathbf{1} = \mathbf{0}.
\end{aligned}$$

This shows that the candidate solution $\bar{\boldsymbol{q}} = g\left(\frac{1}{n}\right)\mathbf{1}$ is indeed optimal. By Lemma 2, we have $\bar{\boldsymbol{w}} = \lim_{t\to\infty} \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2} = \frac{\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}}{\left\|\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}\right\|_2} = \frac{g(1/n)\boldsymbol{X}^\top \boldsymbol{y}}{\left\|g(1/n)\boldsymbol{X}^\top \boldsymbol{y}\right\|_2}$. On the other hand, since $\boldsymbol{y}$ is an exact eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$, we have $\boldsymbol{w}_{\mathsf{MNI}} = \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y} = \frac{1}{k}\boldsymbol{X}^\top\boldsymbol{y}$ for some positive eigenvalue $k > 0$. Therefore, $\boldsymbol{w}_{\mathsf{MNI}} \propto \bar{\boldsymbol{w}}$. This completes the proof of Part 1 of the proposition.

Proof of Part 2. As with the proof of Part 1 of the proposition, we start by analyzing the convex program (5). In the special case $g(d) = d$, the KKT conditions reduce to

$$
\begin{aligned}
-q_i &< 0 \qquad \text{for all } i \in [n], && \text{(22a)}\\
1 - \sum_{i=1}^{n} q_i &\le 0, && \text{(22b)}\\
\lambda_i &\ge 0 \qquad \text{for all } i \in [n], && \text{(22c)}\\
\mu &\ge 0, && \text{(22d)}\\
-\lambda_i q_i &= 0 \qquad \text{for all } i \in [n], && \text{(22e)}\\
\mu\left(1 - \sum_{i=1}^{n} q_i\right) &= 0, && \text{(22f)}\\
\operatorname{diag}(\boldsymbol{y})\,\boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\,\boldsymbol{q} - \boldsymbol{\lambda} - \mu\mathbf{1} &= \mathbf{0}. && \text{(22g)}
\end{aligned}
$$

In this case, we pick the candidate solution $\bar{\boldsymbol{q}} = \frac{\operatorname{diag}(\boldsymbol{y})\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}}{\left\|\operatorname{diag}(\boldsymbol{y})\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_1}$ and verify that it satisfies all the KKT conditions below.

- The primal feasibility equations, Equations (22a) and (22b), are satisfied because of our assumed condition $\operatorname{diag}(\boldsymbol{y})\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y} \succ \mathbf{0}$ and because, by definition, $\sum_{i=1}^{n} \bar{q}_i = 1$.

- The dual feasibility equations, Equations (22c) and (22d), are satisfied by setting $\boldsymbol{\lambda} = \mathbf{0}$ and $\mu = \frac{1}{\left\|\operatorname{diag}(\boldsymbol{y})\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_1}$.

- The complementary slackness equations, Equations (22e) and (22f), are satisfied because we have set $\boldsymbol{\lambda} = \mathbf{0}$ and $1 - \sum_{i=1}^{n} q_i = 0$.

- The stationary condition, Equation (22g), is satisfied because

$$
\begin{aligned}
&\operatorname{diag}(\boldsymbol{y})\,\boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\,\boldsymbol{q} - \boldsymbol{\lambda} - \mu\mathbf{1}\\
&= \frac{1}{\left\|\operatorname{diag}(\boldsymbol{y})\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_1}\mathbf{1} - \frac{1}{\left\|\operatorname{diag}(\boldsymbol{y})\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_1}\mathbf{1} = \mathbf{0},
\end{aligned}
$$

  where we have used $\operatorname{diag}(\boldsymbol{y})\operatorname{diag}(\boldsymbol{y}) = \boldsymbol{I}$ due to the labels being binary, i.e. $y_i = \pm 1$.

Therefore, the candidate solution $\bar{\boldsymbol{q}}$ is optimal. By Lemma 2, we have $\bar{\boldsymbol{w}} = \lim_{t\to\infty} \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2} = \frac{\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}}{\left\|\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}\right\|_2} = \frac{\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}}{\left\|\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2}$. Therefore, $\boldsymbol{w}_{\mathsf{MNI}} \propto \bar{\boldsymbol{w}}$. This completes the proof of Part 2 of the proposition. ■

A.4 Proof of Theorem 6 (Approximate Equivalence to MNI Upper Bound)

In this section, we present the proof of Theorem 6. We divide the proof in four steps.

Step 1. Our proof starts with the relaxed convex program (5) and identifies a necessary set of characteristic equations that the optimal solution $\bar{\boldsymbol{q}}$ needs to satisfy. The KKT conditions for this convex program are given in Equation (21). Lemma 3 postulates that any optimal solution must satisfy $\boldsymbol{\lambda} = \boldsymbol{0}$ and $\mu > 0$; therefore, it is necessary for $\bar{\boldsymbol{q}}$ to satisfy the following characteristic equations

$$\operatorname{diag}(\boldsymbol{y})\, \boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}} = \mu h(\bar{\boldsymbol{q}}), \tag{23a}$$

$$\mu > 0, \tag{23b}$$

$$\sum_{i=1}^{n} g^{-1}(\bar{q}_i) = 1, \tag{23c}$$

where we have denoted $h(\boldsymbol{q}) := \left[g^{-1}\right]'(\boldsymbol{q})$ as shorthand.

Step 2. Next, we use the nonlinear characteristic equations in Equation (23) to determine a relationship between the directions of the vectors $\bar{\boldsymbol{q}}$ and $h(\bar{\boldsymbol{q}})$. We denote $\bar{\boldsymbol{q}}_{\boldsymbol{y}} := \operatorname{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}}$ and $\epsilon_\alpha(\boldsymbol{q}) := \frac{\left\|\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{q} - \alpha \boldsymbol{q}\right\|_2}{\|\boldsymbol{q}\|_2}$ as shorthand. From Equation (23a), we have the following sequence of implications for any value of $\alpha > 0$

$$\begin{aligned}
&\operatorname{diag}(\boldsymbol{y})\, \boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}} = \mu h(\bar{\boldsymbol{q}}) \\
&\iff \boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{q}}_{\boldsymbol{y}} = \mu \operatorname{diag}(\boldsymbol{y})\, h(\bar{\boldsymbol{q}}) \\
&\iff \left(\boldsymbol{X}\boldsymbol{X}^\top - \alpha \boldsymbol{I}\right) \bar{\boldsymbol{q}}_{\boldsymbol{y}} = \mu \operatorname{diag}(\boldsymbol{y})\, h(\bar{\boldsymbol{q}}) - \alpha \bar{\boldsymbol{q}}_{\boldsymbol{y}} \\
&\Rightarrow \left\|\left(\boldsymbol{X}\boldsymbol{X}^\top - \alpha \boldsymbol{I}\right) \bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2 = \left\|\mu \operatorname{diag}(\boldsymbol{y})\, h(\bar{\boldsymbol{q}}) - \alpha \bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2 \\
&\iff \epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) = \left\|\mu \operatorname{diag}(\boldsymbol{y}) \frac{h(\bar{\boldsymbol{q}})}{\left\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2} - \alpha \frac{\bar{\boldsymbol{q}}_{\boldsymbol{y}}}{\left\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2}\right\|_2 \\
&\iff \epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) = \left\|\mu \frac{h(\bar{\boldsymbol{q}})}{\|\bar{\boldsymbol{q}}\|_2} - \alpha \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2}\right\|_2 \\
&\iff \epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) = \left\|\frac{\left\|\operatorname{diag}(\boldsymbol{y})\, \boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}}\right\|_2}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} \frac{h(\bar{\boldsymbol{q}})}{\|\bar{\boldsymbol{q}}\|_2} - \alpha \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2}\right\|_2.
\end{aligned} \tag{24}$$

The last implication follows because Equation (23a) implies $\left\|\operatorname{diag}(\boldsymbol{y})\, \boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}}\right\|_2 = \mu \left\|h(\bar{\boldsymbol{q}})\right\|_2$. We proceed from Equation (24). By the reverse triangle inequality, we have

$$\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) = \left\| \left(\alpha \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} - \alpha \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} \right) - \left(\alpha \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} - \frac{\left\|\mathrm{diag}(\boldsymbol{y}) \boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{y}) \bar{\boldsymbol{q}}\right\|_2}{\|\bar{\boldsymbol{q}}\|_2} \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} \right) \right\|_2$$

$$\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) \geq \left| \left\| \alpha \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} - \alpha \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} \right\|_2 - \left\| \alpha \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} - \frac{\left\|\mathrm{diag}(\boldsymbol{y}) \boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{y}) \bar{\boldsymbol{q}}\right\|_2}{\|\bar{\boldsymbol{q}}\|_2} \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} \right\|_2 \right|$$

$$= \left| \left\| \alpha \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} - \alpha \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} \right\|_2 - \underbrace{\left\| \left(\alpha - \frac{\left\|\mathrm{diag}(\boldsymbol{y}) \boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{y}) \bar{\boldsymbol{q}}\right\|_2}{\|\bar{\boldsymbol{q}}\|_2} \right) \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} \right\|_2}_{C} \right|$$

$$\Rightarrow \left\| \alpha \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} - \alpha \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} \right\|_2 \leq \epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) + C. \tag{25}$$

Therefore, it suffices to upper-bound the term C. We get

$$\begin{aligned} C &= \left\| \left(\alpha - \frac{\left\|\mathrm{diag}(\boldsymbol{y}) \boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{y}) \bar{\boldsymbol{q}}\right\|_2}{\|\bar{\boldsymbol{q}}\|_2} \right) \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} \right\|_2 \\ &= \left| \alpha - \frac{\left\|\mathrm{diag}(\boldsymbol{y}) \boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{y}) \bar{\boldsymbol{q}}\right\|_2}{\|\bar{\boldsymbol{q}}\|_2} \right| \left\| \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} \right\|_2 \\ &= \left| \alpha - \frac{\left\|\boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2}{\left\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2} \right| \\ &\leq \frac{\left\|\boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{q}}_{\boldsymbol{y}} - \alpha \bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2}{\left\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2} := \epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}), \end{aligned}$$

where the last inequality follows by again applying the reverse triangle inequality. Hence, Equation (25) together with the upper bound on C gives us

$$\left\| \frac{h(\bar{\boldsymbol{q}})}{\left\|h(\bar{\boldsymbol{q}})\right\|_2} - \frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} \right\|_2 \leq \frac{2\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}})}{\alpha}. \tag{26}$$

Step 3. Next, we show that the angle between $\bar{\boldsymbol{q}}$ and $\mathbf{1}$ is less than or equal to the angle between $h(\bar{\boldsymbol{q}})$ and $\bar{\boldsymbol{q}}$. We introduce the following key lemma, which critically utilizes the convexity of $g(\cdot)$.

Lemma 22 *For every non-negative, strictly increasing, convex function* $g : [0,1] \rightarrow [0,1]$, *where* $g\left(0\right) = 0$ *and* $g\left(1\right) = 1$, *we have* $h\left(q\right) := \left[g^{-1}\right]'\left(q\right)$ *is a decreasing function satisfying*

$$\frac{\sqrt{n}\sum_{i=1}^{n} h\left(q_i\right) q_i}{\left(\sum_{i=1}^{n} q_i\right)\left\|h\left(\boldsymbol{q}\right)\right\|_2} \leq 1 \tag{27}$$

for all $0 \leq q_i \leq 1$ *and* $\boldsymbol{q} = \left(q_1, \cdots, q_n\right)^\top$.

Proof (of Lemma 22) Without loss of generality, we assume $\{q_i\}$ is an increasing sequence, where $q_i \leq q_j$ if index $i \leq j$. Next, since $g\left(\cdot\right)$ is convex and strictly increasing, $g^{-1}\left(\cdot\right)$ is a concave function, and then $h\left(\cdot\right) := \left[g^{-1}\right]'\left(\cdot\right)$ is a decreasing function. Hence, $\left\{h\left(q_i\right)\right\}$ is a decreasing sequence, where $h\left(q_i\right) \geq h\left(q_j\right)$ for $i \leq j$. Then, according to Chebyshev's Sum Inequality (Hardy et al., 1952), we can have

$$\frac{\sqrt{n}\sum_{i=1}^{n} h\left(q_i\right) q_i}{\left(\sum_{i=1}^{n} q_i\right)\left\|h\left(\boldsymbol{q}\right)\right\|_2} \leq \frac{\frac{\sqrt{n}}{n}\left(\sum_{i=1}^{n} h\left(q_i\right)\right)\left(\sum_{i=1}^{n} q_i\right)}{\left(\sum_{i=1}^{n} q_i\right)\left\|h\left(\boldsymbol{q}\right)\right\|_2} = \frac{\left\|h\left(\boldsymbol{q}\right)\right\|_1}{\sqrt{n}\left\|h\left(\boldsymbol{q}\right)\right\|_2} \leq 1,$$

where the last inequality holds because $\left\|\cdot\right\|_1 \leq \sqrt{n}\left\|\cdot\right\|_2$. ∎

Then, Equation (26) yields

$$\begin{aligned}
\frac{2\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}})}{\alpha} &\geq \left\|\frac{h\left(\bar{\boldsymbol{q}}\right)}{\left\|h\left(\bar{\boldsymbol{q}}\right)\right\|_2} - \frac{\bar{\boldsymbol{q}}}{\left\|\bar{\boldsymbol{q}}\right\|_2}\right\|_2 \\
&= \sqrt{\sum_{i=1}^{n}\frac{1}{n} - \frac{2\sum_{i=1}^{n} h\left(\bar{q}_i\right)\bar{q}_i}{\left\|h\left(\bar{\boldsymbol{q}}\right)\right\|_2\left\|\bar{\boldsymbol{q}}\right\|_2} + \sum_{i=1}^{n}\frac{\bar{q}_i^2}{\left\|\bar{\boldsymbol{q}}\right\|_2^2}} \\
&= \sqrt{\sum_{i=1}^{n}\frac{1}{n} - \frac{2\sum_{i=1}^{n}\bar{q}_i}{\left\|\bar{\boldsymbol{q}}\right\|_2\sqrt{n}} \times \frac{\sqrt{n}\sum_{i=1}^{n} h\left(\bar{q}_i\right)\bar{q}_i}{\left(\sum_{i=1}^{n}\bar{q}_i\right)\left\|h\left(\bar{\boldsymbol{q}}\right)\right\|_2} + \sum_{i=1}^{n}\frac{\bar{q}_i^2}{\left\|\bar{\boldsymbol{q}}\right\|_2^2}} \\
&\geq \sqrt{\sum_{i=1}^{n}\frac{1}{n} - \frac{2\sum_{i=1}^{n}\bar{q}_i}{\left\|\bar{\boldsymbol{q}}\right\|_2\sqrt{n}} \times 1 + \sum_{i=1}^{n}\frac{\bar{q}_i^2}{\left\|\bar{\boldsymbol{q}}\right\|_2^2}} \\
&= \left\|\frac{\bar{\boldsymbol{q}}}{\left\|\bar{\boldsymbol{q}}\right\|_2} - \frac{1}{\sqrt{n}}\right\|_2,
\end{aligned} \tag{28}$$

where the last inequality follows by applying Lemma 22. To complete the proof of dual variable convergence, we relate $\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}})$ to $\epsilon_\alpha(\boldsymbol{y})$. Denote the unit-normalization of a vector $\boldsymbol{u}(\boldsymbol{q}) := \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2}$ as shorthand. Note that for any vector $\boldsymbol{q}$, we have

$$\epsilon_\alpha(\boldsymbol{q}) := \left\|(\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I})\boldsymbol{u}(\boldsymbol{q})\right\|_2.$$

Therefore, we get

$$\begin{aligned}\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) &= \left\|(\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I})\boldsymbol{u}(\bar{\boldsymbol{q}}_{\boldsymbol{y}})\right\|_2 \\ &\leq \left\|(\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I})\boldsymbol{u}(\boldsymbol{y})\right\|_2 + \left\|(\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I})(\boldsymbol{u}(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) - \boldsymbol{u}(\boldsymbol{y})\right\|_2 \\ &\leq \epsilon_\alpha(\boldsymbol{y}) + \left\|\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I}\right\|_2 \cdot \frac{2\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}})}{\alpha},\end{aligned}$$

where the last inequality follows by noting that $\left\|\boldsymbol{u}(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) - \boldsymbol{u}(\boldsymbol{y})\right\|_2 = \left\|\frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{\mathbf{1}}{\sqrt{n}}\right\|_2$ (owing to the binary labels $y_i = \pm 1$) and substituting Equation (28). Next, we utilize the assumption made in the statement of Theorem 6 that $\alpha > 0$ is chosen such that $\frac{\|\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I}\|_2}{\alpha} \leq \frac{1}{3} < \frac{1}{2}$. This assumption yields

$$\begin{aligned}&\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) \leq \epsilon_\alpha(\boldsymbol{y}) + 2c \cdot \epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) \\ \Rightarrow\ &\epsilon_\alpha(\bar{\boldsymbol{q}}_{\boldsymbol{y}}) \leq \frac{\epsilon_\alpha(\boldsymbol{y})}{(1-2c)} =: \frac{C\epsilon_\alpha(\boldsymbol{y})}{2},\end{aligned}$$

where $C := \frac{2}{(1-2c)} \in (0, \infty)$ is a universal positive constant. Thus, we ultimately get

$$\left\|\frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{1}{\sqrt{n}}\right\|_2 \leq \frac{C\epsilon_\alpha(\boldsymbol{y})}{\alpha} := \frac{C\left\|\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y} - \alpha\boldsymbol{y}\right\|_2}{\alpha\,\|\boldsymbol{y}\|_2}, \tag{29}$$

which completes our dual convergence proof.

Step 4. We complete the proof with the following lemma, which relates the primal variables to the dual variables.

Lemma 23 *Under the assumptions of Theorem 6, we have*

$$\left\|\frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} - \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2}\right\|_2 \leq 4\left\|\frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{1}{\sqrt{n}}\right\|_2 + \frac{12\epsilon_\alpha(\boldsymbol{y})}{\alpha}. \tag{30}$$

Lemma 23 essentially shows that the statement of dual convergence in Equation (29) can be converted into a statement of primal closeness with only the loss of a multiplicative constant factor. The proof of Lemma 23 follows via a series of algebraic manipulations and is listed below. Putting Lemma 23 together with Equation (29) completes the proof of Theorem 6.

Proof (of Lemma 23) Recall from Lemma 2 that the primal implicit bias is defined as $\bar{\boldsymbol{w}} \propto \boldsymbol{X}^\top \mathrm{diag}\,(\boldsymbol{y})\,\bar{\boldsymbol{q}}$. Also, recall the definition of the primal MNI as $\boldsymbol{w}_{\mathsf{MNI}} = \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}$. We define $\boldsymbol{u}_{\mathsf{MNI}} := \frac{\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}}{\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2}$ and $\bar{\boldsymbol{u}} := \frac{\mathrm{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}}{\left\|\mathrm{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}\right\|_2}$. Then, a simple normalization shows that

$$\left\|\frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} - \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2}\right\|_2 = \left\|\frac{\boldsymbol{X}^\top\bar{\boldsymbol{u}}}{\left\|\boldsymbol{X}^\top\bar{\boldsymbol{u}}\right\|_2} - \frac{\boldsymbol{X}^\top\boldsymbol{u}_{\mathsf{MNI}}}{\left\|\boldsymbol{X}^\top\boldsymbol{u}_{\mathsf{MNI}}\right\|_2}\right\|_2.$$

Now, we denote $\boldsymbol{u}_1 \coloneqq \bar{\boldsymbol{u}}$, $\boldsymbol{u}_2 \coloneqq \boldsymbol{u}_{\mathsf{MNI}}$ and $\boldsymbol{\Delta} \coloneqq \boldsymbol{u}_1 - \boldsymbol{u}_2$ as shorthand. We have

$$\begin{aligned}\left\| \frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} - \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2} \right\|_2 &= \left\| \frac{\boldsymbol{X}^\top \boldsymbol{u}_1}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2} - \frac{\boldsymbol{X}^\top \boldsymbol{u}_2}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_2\right\|_2} \right\|_2 \\ &\leq \underbrace{\left\| \frac{\boldsymbol{X}^\top \boldsymbol{u}_1}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2} - \frac{\boldsymbol{X}^\top \boldsymbol{u}_2}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2} \right\|_2}_{T_1} + \underbrace{\left\| \frac{\boldsymbol{X}^\top \boldsymbol{u}_2}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2} - \frac{\boldsymbol{X}^\top \boldsymbol{u}_2}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_2\right\|_2} \right\|_2}_{T_2}.\end{aligned}$$

We first show that both T_1 and T_2 are upper bounded by $2\|\boldsymbol{\Delta}\|_2$. We denote $\epsilon \coloneqq \left\|\boldsymbol{X}\boldsymbol{X}^\top - \alpha \boldsymbol{I}\right\|_2$ as shorthand, and note that by the assumptions of Theorem 6 we have $\epsilon \leq \frac{\alpha}{3}$. Beginning with T_1, note that

$$T_1 = \frac{\left\|\boldsymbol{X}^\top (\boldsymbol{u}_1 - \boldsymbol{u}_2)\right\|_2}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2} \leq \frac{\sigma_{\max}\left(\boldsymbol{X}^\top\right) \cdot \|\boldsymbol{\Delta}\|_2}{\sigma_{\min}\left(\boldsymbol{X}^\top\right)} \leq \sqrt{\frac{\alpha+\epsilon}{\alpha-\epsilon}} \cdot \|\boldsymbol{\Delta}\|_2 \leq 2\|\boldsymbol{\Delta}\|_2 .$$

Above, the last inequality uses that $\sqrt{\frac{\alpha+\epsilon}{\alpha-\epsilon}} \leq \frac{\alpha+\epsilon}{\alpha-\epsilon} \leq 2$ as long as $\epsilon \leq \frac{\alpha}{3}$. Proceeding to T_2, we have

$$\begin{aligned} T_2 &= \left(\frac{1}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2} - \frac{1}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_2\right\|_2} \right) \left\|\boldsymbol{X}^\top \boldsymbol{u}_2\right\|_2 \\ &\leq \sqrt{\alpha+\epsilon} \cdot \frac{\left\|\boldsymbol{X}^\top \boldsymbol{u}_2\right\|_2 - \left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2 \cdot \left\|\boldsymbol{X}^\top \boldsymbol{u}_2\right\|_2} \\ &\leq \frac{\sqrt{\alpha+\epsilon} \cdot \left\|\boldsymbol{X}^\top (\boldsymbol{u}_1 - \boldsymbol{u}_2)\right\|_2}{\left\|\boldsymbol{X}^\top \boldsymbol{u}_1\right\|_2 \cdot \left\|\boldsymbol{X}^\top \boldsymbol{u}_2\right\|_2} \\ &\leq \frac{(\alpha+\epsilon)\|\boldsymbol{\Delta}\|_2}{(\alpha-\epsilon)} \\ &\leq 2\|\boldsymbol{\Delta}\|_2 , \end{aligned}$$

where in the above we have repeatedly used the inequality $\sqrt{\alpha-\epsilon} \leq \sigma_{\min}\left(\boldsymbol{X}^\top\right) \leq \sigma_{\max}\left(\boldsymbol{X}^\top\right) \leq \sqrt{\alpha+\epsilon}$. The second inequality above uses the reverse triangle inequality, and the last inequality again uses $\frac{\alpha+\epsilon}{\alpha-\epsilon} \leq 2$ as long as $\epsilon \leq \frac{\alpha}{3}$. Combining the upper bounds on T_1 and T_2 thus yields

$$\left\| \frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} - \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2} \right\|_2 \leq 4\|\boldsymbol{\Delta}\|_2 . \tag{31}$$

It remains to show that $\|\boldsymbol{\Delta}\|_2 \leq \frac{C\epsilon_\alpha(\boldsymbol{y})}{\alpha}$ for some universal constant C. We will use the statement of Equation (29) as a starting point to upper-bound $\|\boldsymbol{\Delta}\|_2$. Recall that $\boldsymbol{\Delta} = \bar{\boldsymbol{u}} - \boldsymbol{u}_{\mathsf{MNI}}$ and $\bar{\boldsymbol{u}} := \frac{\mathrm{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}}{\left\|\mathrm{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}\right\|_2}$. Then, applying the triangle inequality gives us

$$\|\boldsymbol{\Delta}\|_2 \leq \left\|\bar{\boldsymbol{u}} - \frac{\boldsymbol{y}}{\sqrt{n}}\right\|_2 + \left\|\boldsymbol{u}_{\mathsf{MNI}} - \frac{\boldsymbol{y}}{\sqrt{n}}\right\|_2 .$$

Consequently, it suffices to show that $\boldsymbol{u}_{\mathsf{MNI}}$ is sufficiently close to the label vector $\boldsymbol{y}$; in other words, to upper bound $\left\|\boldsymbol{u}_{\mathsf{MNI}} - \frac{\boldsymbol{y}}{\sqrt{n}}\right\|_2$. We use a similar algebraic technique as in the preceding steps. First, we write

$$\begin{aligned}
\left\|\boldsymbol{u}_{\mathsf{MNI}} - \frac{\boldsymbol{y}}{\sqrt{n}}\right\|_2 &= \left\| \frac{\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}}{\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2} - \frac{\alpha^{-1}\boldsymbol{y}}{\left\|\alpha^{-1}\boldsymbol{y}\right\|_2} \right\|_2 \\
&\leq \underbrace{\left\| \frac{\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}}{\left\|\alpha^{-1}\boldsymbol{y}\right\|_2} - \frac{\alpha^{-1}\boldsymbol{y}}{\left\|\alpha^{-1}\boldsymbol{y}\right\|_2} \right\|_2}_{T_1} + \underbrace{\left\| \frac{\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}}{\left\|\alpha^{-1}\boldsymbol{y}\right\|_2} - \frac{\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}}{\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2} \right\|_2}_{T_2} .
\end{aligned}$$

It remains to upper bound T_1 and T_2. Beginning with T_1, we have

$$\begin{aligned}
T_1 &= \frac{\left\|\left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \alpha^{-1}\boldsymbol{I}\right)\boldsymbol{y}\right\|_2}{\left\|\alpha^{-1}\boldsymbol{y}\right\|_2} \\
&\leq \left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2 \cdot \frac{\left\|\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y} - \alpha\boldsymbol{y}\right\|_2}{\|\boldsymbol{y}\|_2} \\
&= \frac{\left\|\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y} - \alpha\boldsymbol{y}\right\|_2}{\lambda_{\min}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \cdot \|\boldsymbol{y}\|_2} .
\end{aligned}$$

Now, we note that $\lambda_{\min}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \geq \alpha - \epsilon \geq \frac{2\alpha}{3}$ because $\epsilon \leq \frac{\alpha}{3}$. Consequently, we get

$$T_1 \leq \frac{3\left\|\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y} - \alpha\boldsymbol{y}\right\|_2}{2\alpha \cdot \|\boldsymbol{y}\|_2} =: \frac{3\epsilon_\alpha(\boldsymbol{y})}{2\alpha} .$$

Proceeding to T_2, an identical series of arguments to the previous term T_2 yields

$$\begin{aligned}
T_2 &= \frac{\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2}{\left\|\alpha^{-1}\boldsymbol{y}\right\|_2 \cdot \left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2} \cdot \left(\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2 - \left\|\alpha^{-1}\boldsymbol{y}\right\|_2\right) \\
&\leq \frac{\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2}{\left\|\alpha^{-1}\boldsymbol{y}\right\|_2 \cdot \left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}\right\|_2} \cdot \left\|\left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \alpha^{-1}\boldsymbol{I}\right)\boldsymbol{y}\right\|_2 \\
&= \frac{\left\|\left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \alpha^{-1}\boldsymbol{I}\right)\boldsymbol{y}\right\|_2}{\left\|\alpha^{-1}\boldsymbol{y}\right\|_2} =: T_1 \leq \frac{3\epsilon_\alpha(\boldsymbol{y})}{2\alpha}.
\end{aligned}$$

Consequently, we have $\left\|\boldsymbol{u}_{\mathsf{MNI}} - \frac{\boldsymbol{y}}{\sqrt{n}}\right\|_2 \leq T_1 + T_2 \leq \frac{3\epsilon_\alpha(\boldsymbol{y})}{\alpha}$, and so we ultimately get $\|\boldsymbol{\Delta}\|_2 \leq \left\|\frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{1}{\sqrt{n}}\right\|_2 + \frac{3\epsilon_\alpha(\boldsymbol{y})}{\alpha}$. Combining this with Equation (31) yields

$$\left\|\frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|_2} - \frac{\boldsymbol{w}_{\mathsf{MNI}}}{\|\boldsymbol{w}_{\mathsf{MNI}}\|_2}\right\|_2 \leq 4\|\boldsymbol{\Delta}\|_2 \leq 4\left\|\frac{\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|_2} - \frac{1}{\sqrt{n}}\right\|_2 + \frac{12\epsilon_\alpha(\boldsymbol{y})}{\alpha},$$

completing the desired proof of our theorem.

■

A.5 Proof of Corollary 7 (Upper Bound in Effective Dimensions)

In this section, we prove Corollary 7.

Proof (of Corollary 7) We consider the setting of independent sub-Gaussian covariates described in Corollary 7 and set $\alpha := \|\boldsymbol{\lambda}\|_1$. It suffices to show the following with high probability:

1. $\frac{\|\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I}\|_2}{\alpha} \leq \frac{1}{3}$, and
2. $\frac{C\epsilon_\alpha(\boldsymbol{y})}{\alpha} \leq \max\left\{\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right\}$.

To prove both statements, we will use Hsu et al. (2021, Lemma 8), restated below.

Lemma 24 (Hsu et al., 2021) *For any $\tau > 0$ and a universal constant $c > 0$, we have*

$$\mathbb{P}\left[\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right\|_2 \geq \tau\right] \leq 2 \cdot 9^n \cdot \exp\left(-c \cdot \min\left\{\frac{\tau^2}{v^2\|\boldsymbol{\lambda}\|_2^2}, \frac{\tau}{v\|\boldsymbol{\lambda}\|_\infty}\right\}\right).$$

To prove the first statement, we select $\tau = \frac{1}{3}\|\boldsymbol{\lambda}\|_1$, so that the upper bound on the probability becomes $2 \cdot 9^n \cdot \exp\left(-\frac{c}{9} \cdot \min\left\{\frac{d_2}{v^2}, \frac{d_\infty}{v}\right\}\right)$. Because $d_2 \gg v^2 n$ and $d_\infty \gg vn$, there exists a large enough constant $C > 0$ such that $\min\left\{\frac{d_2}{v^2}, \frac{d_\infty}{v}\right\} \geq Cn$ and $C \cdot c > 9 \ln 9$. Therefore, we get $\frac{\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2}{\|\boldsymbol{\lambda}\|_1} \leq \frac{1}{3}$ with probability at least $1 - 2 \cdot \exp(-n(C \cdot c/9 - \ln 9))$. To prove the second statement, we instead select $\tau = C \cdot v \cdot \max(\|\boldsymbol{\lambda}\|_2 \sqrt{n}, \|\boldsymbol{\lambda}\|_\infty n)$, where $C > 1$ is picked to be large enough so that $C \cdot c > \ln 9$. This ensures that $\min\left\{\frac{\tau^2}{v^2 \|\boldsymbol{\lambda}\|_2^2}, \frac{\tau}{v \|\boldsymbol{\lambda}\|_\infty}\right\} \geq Cn$, and in turn that

$$\mathbb{P}\left[\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \geq C \cdot v \cdot \max(\|\boldsymbol{\lambda}\|_2 \sqrt{n}, \|\boldsymbol{\lambda}\|_\infty n)\right] \leq 2 \cdot \exp(-n(C \cdot c - \ln 9)).$$

Thus, for the choice $\alpha \coloneqq \|\boldsymbol{\lambda}\|_1$, we have, with probability at least $1 - 2e^{-c'n}$,

$$\begin{aligned}
\frac{C\epsilon_\alpha(\boldsymbol{y})}{\alpha} &= \frac{\left\|\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y} - \|\boldsymbol{\lambda}\|_1 \boldsymbol{y}\right\|_2}{\|\boldsymbol{\lambda}\|_1 \|\boldsymbol{y}\|_2} \\
&\leq \frac{\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2}{\|\boldsymbol{\lambda}\|_1} \\
&\leq C \cdot v \cdot \max\left\{\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right\}
\end{aligned}$$

which completes the proof of the second statement. ∎

A.6 Proof of Proposition 8 (Popular Loss Functions)

In this section, we prove Proposition 8.

Proof (of Proposition 8) In Ji and Telgarsky (2021, Theorem 11), it has already shown that ℓ_{exp}, ℓ_{log} and ℓ_{poly} satisfy both Assumption 1 and Lemma 1. In the following, we demonstrate their respective $g(\cdot)$ functions.

Exponential loss: Here $\ell(z) = \exp(z)$, $\ell'(z) = \exp(z)$, $\ell''(z) = \exp(z)$, $\ell^{-1}(z) = \ln(z)$, and $\ell'\big(\ell^{-1}(z)\big) = z$. Therefore, we have $\ell'(\ell^{-1}(z)) = z$, which directly gives $\frac{\ell'(\ell^{-1}(a \cdot z))}{\ell'(\ell^{-1}(a))} = z$ and yields the function $g_{\text{exp}}(d) = d$.

Logistic loss: Here $\ell(z) = \ln\big(1 + \exp(z)\big)$, $\ell'(z) = \frac{\exp(z)}{1+\exp(z)}$, $\ell''(z) = \frac{\exp(z)}{\big(1+\exp(z)\big)^2}$, $\ell^{-1}(z) = \ln\big(\exp(z) - 1\big)$, and $\ell'\big(\ell^{-1}(z)\big) = 1 - \exp(-z)$. Consequently, we have $\ell'(\ell^{-1}(z)) = 1 - \exp(-z)$ and so $\frac{\ell'(\ell^{-1}(a \cdot z))}{\ell'(\ell^{-1}(a))} = \frac{1-\exp(-a \cdot z)}{1-\exp(-a)}$. Applying l'Hospital's rule yields

$$\lim_{a \to 0} \frac{\ell'(\ell^{-1}(a \cdot z))}{\ell'(\ell^{-1}(a))} = \lim_{a \to 0} \frac{1 - \exp(-a \cdot z)}{1 - \exp(-a)} = \lim_{a \to 0} \frac{\exp(-a \cdot z) \cdot z}{\exp(-a)} = z.$$

As a result, we get $g_{\text{log}}(d) = d$.

Polynomial loss (degree $m > 0$): Here we use the continuation of the polynomial loss to $z > 0$ used in Ji and Telgarsky (2021); Wang et al. (2021b) to ensure convexity.

$$\ell(z) = \begin{cases} \frac{1}{(1-z)^m} & z \leq 0 \\ 2mz + \frac{1}{(1+z)^m} & z > 0, \end{cases} \qquad \ell'(z) = \begin{cases} \frac{m}{(1-z)^{m+1}} & z \leq 0 \\ 2m - \frac{m}{(1+z)^{m+1}} & z > 0, \end{cases}$$

$$\ell''(z) = \begin{cases} \frac{m(m+1)}{(1-z)^{m+2}} & z \leq 0 \\ \frac{m(m+1)}{(1+z)^{m+2}} & z > 0. \end{cases}$$

For $z \leq \ell(0)$, we have $\ell^{-1}(z) = 1 - z^{-1/m}$, and $\ell'\left(\ell^{-1}\left(z\right)\right) = mz^{\frac{m+1}{m}}$. Hence, we have

$$\lim_{a \to 0} \frac{\ell'(\ell^{-1}(a \cdot z))}{\ell'(\ell^{-1}(a))} = \left(\frac{a \cdot z}{a}\right)^{\frac{m+1}{m}} = z^{\frac{m+1}{m}},$$

and so we get $g_{\mathrm{poly}}(d) = d^{\frac{m+1}{m}}$.

■

Appendix B. Derivations for Lemma 9 (Generalizing Primal-dual Analysis to the Multiclass Setting)

In this section, we provide the derivations for Lemma 9, which generalizes the primal-dual analysis of Ji and Telgarsky (2021) to the multiclass setting. In particular, we state several lemmas for the multiclass that are analogous to the lemmas in Ji and Telgarsky (2021) for the binary case. We first introduce these analogous lemmas, and then we show Lemma 9 follows as a direct result of them. Note that we require these analogous lemmas because the generalized sum $\psi(\cdot)$ is slightly different in the multiclass setting for loss functions satisfying Assumption 3, and completely different for the cross-entropy loss under Assumption 4. Some of these lemmas are direct extensions of those in Ji and Telgarsky (2021), and so we do not provide proofs for these particular lemmas.

Lemma 25 *Under Assumption 3, (or 4), for all $\boldsymbol{Q} \in \operatorname{dom} \psi^*$, if $\hat{\eta}_t \leq 1/\beta$, then the following results hold:*

1. *Dual convergence: for all $t \geq 0$,*

$$F\left(\boldsymbol{Q}_{t+1}\right) \leq F\left(\boldsymbol{Q}_t\right), \text{ and } \hat{\eta}_t\Big(F\left(\boldsymbol{Q}_{t+1}\right) - F\left(\boldsymbol{Q}\right)\Big) \leq D_{\psi^*}\left(\boldsymbol{Q}, \boldsymbol{Q}_t\right) - D_{\psi^*}\left(\boldsymbol{Q}, \boldsymbol{Q}_{t+1}\right).$$

As a result, for all $t > 0$,

$$F\left(\boldsymbol{Q}_t\right) - F\left(\boldsymbol{Q}\right) \leq \frac{D_{\psi^*}\left(\boldsymbol{Q}, \boldsymbol{Q}_0\right) - D_{\psi^*}\left(\boldsymbol{Q}, \boldsymbol{Q}_t\right)}{\sum_{j<t} \hat{\eta}_j} \leq \frac{D_{\psi^*}\left(\boldsymbol{Q}, \boldsymbol{Q}_0\right)}{\sum_{j<t} \hat{\eta}_j}$$

2. *Primal convergence: for all $t \geq 0$,*

$$\psi\left(\boldsymbol{P}_t\right) - \psi\left(\boldsymbol{P}_{t+1}\right) \geq \hat{\eta}_t\Big(F\left(\boldsymbol{Q}_t\right) + F\left(\boldsymbol{Q}_{t+1}\right)\Big) = \frac{\hat{\eta}_t}{2}\left(\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}_t\right\|^2 + \left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}_{t+1}\right\|^2\right),$$

and thus if $\hat{\eta}_t$ is nonincreasing, we have

$$\psi\left(\boldsymbol{P}_0\right) - \psi\left(\boldsymbol{P}_t\right) \geq \sum_{j<t} \hat{\eta}_j \left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}_j\right\|^2 - \frac{\hat{\eta}_0}{2}\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}_0\right\|^2 + \frac{\hat{\eta}_t}{2}\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}_t\right\|^2$$

This lemma is analogous to and a direct application of Ji and Telgarsky (2021, Theorem 1) in our notation, since the primal and dual setup is identical to the binary case and the generalized sum $\psi(\cdot)$ was verified to be β-smooth for loss functions satisfying either Assumption 3, or 4.

Lemma 26 *Under Assumption 3, (or 4) and Assumption 2, suppose $\hat{\eta}_t \leq 1/\beta$ is nonincreasing, and $\sum_{t=0}^{\infty} \hat{\eta}_t = \infty$.*

1. *The set $\left\{\boldsymbol{Q} | \psi^*\left(\boldsymbol{Q}\right) \leq 0\right\}$ is nonempty, compact and convex. Moreover, $\min_{\psi^*(\boldsymbol{Q}) \leq 0} F\left(\boldsymbol{Q}\right) > 0$, and $\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}$ is the same for all $\bar{\boldsymbol{Q}} \in \arg\min_{\psi^*(\boldsymbol{Q}) \leq 0} F\left(\boldsymbol{Q}\right)$.*

2. *For $\bar{\boldsymbol{Q}} \in \arg\min_{\psi^*(\boldsymbol{Q})\leq 0} F(\boldsymbol{Q})$, and all t with $\psi\left(\boldsymbol{C}\tilde{\boldsymbol{Z}}\boldsymbol{W}_t\right) \leq 0$ (which holds for all large enough t), we have*

$$\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}_t - \tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}\right\|^2 \leq \frac{2D_{\psi^*}\left(\bar{\boldsymbol{Q}}, \boldsymbol{Q}_0\right)}{\sum_{j<t}\hat{\eta}_j}, \text{ and } \left\langle \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}, \frac{\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}}{\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}\right\|}\right\rangle \geq 1 - \frac{\delta\left(\boldsymbol{Q}_0, \bar{\boldsymbol{Q}}\right)}{\sum_{j<t}\hat{\eta}_j},$$

where

$$\delta\left(\boldsymbol{W}_0, \bar{\boldsymbol{Q}}\right) := \frac{\psi\left(\boldsymbol{Q}_0\right) + \hat{\eta}_0 F\left(\boldsymbol{Q}_0\right) + \|\boldsymbol{W}_0\|\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}\right\|}{2F\left(\bar{\boldsymbol{Q}}\right)}$$

is a constant, depending only on $\boldsymbol{W}_0$ and $\bar{\boldsymbol{Q}}$. In particular, it holds that the implicit bias is

$$\bar{\boldsymbol{W}} := \lim_{t\to\infty}\frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|} = \frac{\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}}{\left\|\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}\right\|}, \tag{32}$$

where $\bar{\boldsymbol{Q}} = \left(\bar{\boldsymbol{q}}_1^\top, \cdots, \bar{\boldsymbol{q}}_K^\top\right)^\top$ for $\bar{\boldsymbol{q}}_k \in \mathbb{R}^n$ for all $k \in [K]$.

This lemma is analogous to Ji and Telgarsky (2021, Theorem 5), and almost all the steps in its proof are a direct extension of their proof. We only reproduce the parts of the proof that need to be done from scratch. We begin with the following lemma, which shows the feasibility of the convex conjugate constraint $\psi^*(\boldsymbol{Q}) \leq 0$. This admits a different proof from the binary case due to the differing formulations of the generalized sum $\psi(\cdot)$ in the multiclass case.

Lemma 27 *Under Assumption 3, (or 4), for $\boldsymbol{\Xi} \in \mathbb{R}^{Kn}$ such that $\psi(\boldsymbol{\Xi}) \leq 0$, it holds that $\psi^*\left(\nabla\psi(\boldsymbol{\Xi})\right) \leq 0$. This lemma is analogous to Ji and Telgarsky (2021, Lemma 6).*

Before we prove Lemma 27, we introduce Lemma 28 as an auxiliary lemma.

Lemma 28 *For loss functions under Assumption 1, we have $\sigma(s) := \ell'\big(\ell^{-1}(s)\big)\ell^{-1}(s)$ which is a super-additive function on $\big(0, \ell(0)\big)$.*

Proof (of Lemma 28) The proof follows the proof of Ji and Telgarsky (2021, Lemma 6). Note that by Assumption 1, we have

$$\lim_{s\to 0}\sigma(s) = \ell'\big(\ell^{-1}(s)\big)\ell^{-1}(s) = 0 \quad \text{and} \quad \sigma(s)/s \text{ is increasing on } \big(0, \ell(0)\big),$$

by letting $s = \ell(z)$ and $z = \ell^{-1}(s)$. Next, for some $t \in \mathbb{R}$ and $0 < t \leq 1$, we assume $s_1 = tx$ and $s_2 = x$ for some $x \in \mathbb{R}$ and $0 < x \leq \ell(0)$. Then we have

$$\frac{\sigma(s_1)}{s_1} \leq \frac{\sigma(s_2)}{s_2} \iff \frac{\sigma(tx)}{tx} \leq \frac{\sigma(x)}{x} \iff \sigma(tx) \leq t\sigma(x).$$

For $a, b > 0$ and $a + b < \ell(0)$,

$$\begin{aligned}\sigma(a) + \sigma(b) &= \sigma\left(\frac{a}{a+b} \times (a+b)\right) + \sigma\left(\frac{b}{a+b}\right) \times (a+b)\\ &\leq \frac{a}{a+b}\sigma(a+b) + \frac{b}{a+b}\sigma(a+b)\\ &= \sigma(a+b).\end{aligned}$$

■

Armed with Lemma 28 we can prove Lemma 27.

Proof (Proof of Lemma 27) Recalling the definition of ψ and its convex conjugate ψ^*, we have

$$\psi^*\left(\nabla\psi(\boldsymbol{\Xi})\right) = \left\langle \boldsymbol{\Xi}, \nabla\psi(\boldsymbol{\Xi})\right\rangle - \psi(\boldsymbol{\Xi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{\xi_{k,i}\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}}}{\ell'\left(\psi(\boldsymbol{\Xi})\right)} - \psi(\boldsymbol{\Xi}).$$

Multiplying $\ell'\left(\psi(\boldsymbol{\Xi})\right)$ on both sides, we get

$$\begin{aligned}\ell'\left(\psi(\boldsymbol{\Xi})\right)\psi^*\left(\nabla\psi(\boldsymbol{\Xi})\right) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\xi_{k,i}\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}} - \ell'\left(\psi(\boldsymbol{\Xi})\right)\psi(\boldsymbol{\Xi})\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\xi_{k,i}\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}}\\ &\quad - \ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)\right)\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\xi_{k,i}\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}} - \sigma\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right),\end{aligned} \tag{33}$$

since we substitute $\sigma(s) = \ell'\left(\ell^{-1}(s)\right)\ell^{-1}(s)$ by Lemma 28. Hereafter, we handle the situations under Assumption 3 and 4 separately (as the generalized sum ψ is distinct in each case). Under Assumption 3, we have $\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right) = \sum_{k=1}^{K}\ell\left(\xi_{k,i}\right)$, and Equation (33) becomes

$$\begin{aligned}\ell'\left(\psi(\boldsymbol{\Xi})\right)\psi^*\left(\nabla\psi(\boldsymbol{\Xi})\right) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\xi_{k,i}\ell'\left(\xi_{k,i}\right) - \sigma\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\ell\left(\xi_{k,i}\right)\right)\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\sigma\left(\ell\left(\xi_{k,i}\right)\right) - \sigma\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\ell\left(\xi_{k,i}\right)\right) \leq 0,\end{aligned}$$

where the last inequality uses the super-additivity property of $\sigma(s)$ on $\big(0, \ell(0)\big)$ (Lemma 28, together with the assumption $\psi(\boldsymbol{\Xi}) \leq 0$ or equivalently $\sum_{i=1}^{n} \sum_{k=1}^{K} \ell\left(\xi_{k,i}\right) \leq \ell(0)$). Finally, since $\ell' > 0$, we have $\psi^*\left(\nabla\psi(\boldsymbol{\Xi})\right) \leq 0$.

Under Assumption 4, we have

$$\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right) = \ln\left(1 + \sum_{k\neq y_i}^{K} \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right)\right) = \ln\left(1+\delta_i\right),$$

where we denote $\delta_i := \sum_{k\neq y_i}^{K} \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right)$, $\delta_{k,i} := \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right)$. Direct calculations verify that

$$\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}} = \begin{cases} \frac{c_{y_i,i}\delta_i}{1+\delta_i} & k = y_i \\ \frac{-c_{k,i}\delta_{k,i}}{1+\delta_i} & k \neq y_i. \end{cases}$$

Hence, Equation (33) becomes

$$\ell'\left(\psi(\boldsymbol{\Xi})\right)\psi^*\left(\nabla\psi(\boldsymbol{\Xi})\right) = \underbrace{\sum_{i=1}^{n} \frac{c_{y_i,i}\xi_{y_i,i}\delta_i + \sum_{k\neq y_i}^{K} -c_{k,i}\xi_{k,i}\delta_{k,i}}{1+\delta_i}}_{T} - \sigma\left(\sum_{i=1}^{n} \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right). \tag{34}$$

Next, we show T is upper bounded by $\sum_{i=1}^{n} \sigma\Big(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\Big)$ in a series of calculations below

$$\begin{aligned}
T &= \sum_{i=1}^{n} \frac{c_{y_i,i}\xi_{y_i,i}\delta_i + \sum_{k\neq y_i}^{K} -c_{k,i}\xi_{k,i}\delta_{k,i}}{1+\delta_i} \\
&= \sum_{i=1}^{n} \frac{\sum_{k\neq y_i}^{K} (c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i})\delta_{k,i}}{1+\delta_i} \\
&= \sum_{i=1}^{n} \frac{\sum_{k\neq y_i}^{K} \ln\left(\delta_{k,i}\right)\delta_{k,i}}{1+\delta_i} \\
&\leq \sum_{i=1}^{n} \frac{\sum_{k\neq y_i}^{K} \ln\left(\delta_i\right)\delta_{k,i}}{1+\delta_i} \\
&= \sum_{i=1}^{n} \frac{\delta_i}{1+\delta_i} \ln\left(\delta_i\right) \\
&= \sum_{i=1}^{n} \frac{\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right) - 1}{\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)} \left(\ln\left(\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right) - 1\right)\right) \\
&= \sum_{i=1}^{n} \ell'\left(\ell^{-1}\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)\right)\ell^{-1}\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right) = \sum_{i=1}^{n} \sigma\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right).
\end{aligned}$$

Above, the inequality holds because $\ln(z)$ is an increasing function. We also use the property that $\ell'\big(\ell^{-1}(z)\big)= \frac{\exp(z)-1}{\exp(z)}$ and $\ell^{-1}(z) = \ln\big(\exp(z) - 1\big)$ for logistic loss in the last equality. Proceeding from Equation (34), we then get

$$\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)\psi^*\left(\nabla\psi\left(\boldsymbol{\Xi}\right)\right) \leq \sum_{i=1}^{n}\sigma\Big(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\Big)-\sigma\Big(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\Big)\leq 0,$$

where the last inequality uses the super-additivity property of $\sigma(s)$ on $\big(0, \ell(0)\big)$ (Lemma 28, together with the assumption $\psi(\boldsymbol{\Xi}) \leq 0$ or equivalently $\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right) \leq \ell(0)$). Since $\ell' > 0$, we have $\psi^*\big(\nabla\psi(\boldsymbol{\Xi})\big) \leq 0$. This completes the proof for both types of losses. ■

Finally, one key step that is utilized in the proofs of the binary analogs Lemma 25 and 26 (specifically, the proof of Ji and Telgarsky, 2021, Lemma 4 and Ji and Telgarsky, 2021, Theorem 5 part 1) is the statement that $\boldsymbol{Q} \in \operatorname{dom}\psi^* \implies \boldsymbol{Q} = \nabla\psi(\boldsymbol{P})$ for some $\boldsymbol{P}$; or, equivalently, $\boldsymbol{P} \in \partial\psi^*(\boldsymbol{Q}) \implies \boldsymbol{Q} = \nabla\psi(\boldsymbol{P})$. This fact appears from Rockafellar (1970, Theorems 23.5) along with the reverse implication and implicitly assumes the joint convexity of ψ, but we show below that the forward implication continues to hold under individual convexity. (Note that the reverse implications no longer hold under individual convexity, but are not required for these proofs.)

Lemma 29 *For any individually convex and differentiable function $\psi(\cdot)$ and its convex conjugate $\psi^*(\cdot)$, we have that $\boldsymbol{Q} \in dom\,\psi^* \implies \boldsymbol{Q} = \nabla\psi(\boldsymbol{P}^*)$ for some $\boldsymbol{P}^*$ that achieves $\sup_{\boldsymbol{P}\in\mathbb{R}^{Kn}}\langle\boldsymbol{P},\boldsymbol{Q}\rangle - \psi(\boldsymbol{P})$. Equivalently, $\boldsymbol{P}^* \in \partial\psi^*(\boldsymbol{Q}) \implies \boldsymbol{Q} = \nabla\psi(\boldsymbol{P}^*)$.*

Proof (of Lemma 29) Recall the definition of the convex conjugate $\psi^*(\boldsymbol{Q}) = \sup\limits_{\boldsymbol{P}\in\mathbb{R}^{Kn}}\langle\boldsymbol{P},\boldsymbol{Q}\rangle - \psi(\boldsymbol{P})$. Because $\psi(\boldsymbol{P})$ is individually convex, the function $\langle\boldsymbol{P},\boldsymbol{Q}\rangle - \psi(\boldsymbol{P})$ is individually concave. Consider any $\boldsymbol{P}^*$ that achieves the supremum of this function over $\boldsymbol{P}$. Because of the property of individual concavity, it is *necessary* (but not sufficient) for $\boldsymbol{P}^*$ to satisfy the first-order condition $\boldsymbol{Q} = \nabla\psi(\boldsymbol{P}^*)$. Moreover, for any $\boldsymbol{P}^*$ that achieves the supremum of this function over $\boldsymbol{P}$, we have $\psi^*(\boldsymbol{Q}) = \langle\boldsymbol{P}^*,\boldsymbol{Q}\rangle - \psi(\boldsymbol{P}^*)$. Taking the subdifferential with respect to $\boldsymbol{Q}$ directly gives $\boldsymbol{P}^* \in \partial\psi^*(\boldsymbol{Q})$. This completes the proof. ■

Now that we have established Lemmas 25, 26, 27 and 29, Lemma 9 now directly follows as a result.

Appendix C. Proofs of Multiclass Results

In this section, we present the proofs of all of our results for the multiclass case. In order to prove Theorem 10 and Theorem 12, we introduce their respective auxiliary convex programs in multiclass which is analogous to the binary case.

C.1 Proof of Theorem 10 (Approximate Equivalence to One-vs-all MNI Upper Bound)

Similar to the strategy in the binary case, before we prove Theorem 10, we introduce an auxiliary convex program that will ultimately provide a simpler characterization of the solution in (9).

Lemma 30 *Under Assumptions 3 and 2, any solution to the auxiliary convex program*

$$\bar{\boldsymbol{Q}} \in \underset{\boldsymbol{Q}\in\mathbb{R}^{Kn}}{\arg\min} \underbrace{\frac{1}{2}\boldsymbol{Q}^\top \boldsymbol{C}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}}_{F(\boldsymbol{Q})} \tag{35}$$

$$\text{subject to} \qquad -q_{k,i} < 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K],$$

$$\text{and} \qquad 1 - \sum_{i=1}^{n}\sum_{k=1}^{K} g^{-1}\left(q_{k,i}\right) \leq 0.$$

is also an optimal solution to the original convex program (9)*. Above,* $g(\cdot)$ *is a convex function depending on the loss function, defined in Lemma 1.*

In order to prove Lemma 30, we first show that the new equality constraints are sufficient to imply the original equality constraint. Second, we show that any solution to the original program also satisfies the new equality constraint; therefore, the solution sets of both programs coincide. The following lemma demonstrates the first part.

Lemma 31 *Under Assumption 3, for any* $\boldsymbol{Q} \in \mathbb{R}^{Kn}$ *such that* $q_{k,i} = g\left(z_{k,i}\right) = \lim_{a\to 0}\frac{\ell'\left(\ell^{-1}\left(a\cdot z_{k,i}\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)} > 0$*, for all* $i \in [n]$ *and* $k \in [K]$ *with* $z_{k,i} \in (0,1]$ *and* $\sum_{i=1}^{n}\sum_{k=1}^{K} z_{k,i} = 1$*, it implies* $\psi^*(\boldsymbol{Q}) = 0$*.*

Proof (of Lemma 31) Analogous to Lemma 20, we once again apply the convex analysis in the astral space introduced in Dudík et al. (2022). We show that $\boldsymbol{Q}$ is a subdifferential of the astral extension of ψ for some $\overline{\boldsymbol{P}} \in \overline{\mathbb{R}^{Kn}}$, and then the astral convex conjugate (which is equivalent to the original convex conjugate) is equal to zero. We start with writing $\boldsymbol{Q}$ in the astral format. Since $q_{k,i}$ is finite and continuous in a, we can take the limit inside the function, obtaining

$$q_{k,i} = \lim_{a\to 0}\frac{\ell'\left(\ell^{-1}\left(a\cdot z_{k,i}\right)\right)}{\ell'\left(\ell^{-1}\left(a\right)\right)} = \frac{\ell'\left(\ell^{-1}\left(\lim\limits_{a\to 0} a\cdot z_{k,i}\right)\right)}{\ell'\left(\ell^{-1}\left(\lim\limits_{a\to 0} a\right)\right)} = \frac{\overline{\ell}'\left(\overline{\ell}^{-1}\left(\lim\limits_{a\to 0} a\cdot z_{k,i}\right)\right)}{\overline{\ell}'\left(\overline{\ell}^{-1}\left(\lim\limits_{a\to 0} a\right)\right)}, \tag{36}$$

where in the last equality, we replace the original functions with their astral extensions. Next, we define an astral point $\overline{\boldsymbol{P}} \in \overline{\mathbb{R}^{Kn}}$ such that $\overline{p}_{k,i} = \lim_{a \to 0} \overline{\ell}^{-1}\left(a \cdot z_{k,i}\right) = \overline{\ell}^{-1}\left(\lim_{a \to 0} a \cdot z_{k,i}\right)$ for all $i \in [n]$ and $k \in [K]$. This also implies $\sum_{i=1}^{n}\sum_{k=1}^{K} \overline{\ell}(\overline{p}_{k,i}) = \lim_{a \to 0} a$. Substituting these values in Equation (36), we can write $q_{k,i}$ as

$$q_{k,i} = \frac{\overline{\ell}'(\overline{p}_{k,i})}{\overline{\ell}'\left(\overline{\ell}^{-1}\left(\sum_{i=1}^{n} \overline{\ell}\left(\overline{p}_{k,i}\right)\right)\right)}$$

for all $i \in [n]$ and $k \in [K]$.

On the other hand, according to Lemma 1, we can also define $q_{k,i}$ in the limit of a different co-convergent sequence as

$$q_{k,i} = \lim_{a \to 0} \frac{\ell^{-1}(a) \cdot z_{k,i}}{\ell^{-1}\left(a \cdot z_{k,i}\right)} = \frac{\ell^{-1}\left(\lim_{a \to 0} a\right) \cdot z_{k,i}}{\ell^{-1}\left(\lim_{a \to 0} a \cdot z_{k,i}\right)} = \frac{\overline{\ell}^{-1}\left(\lim_{a \to 0} a\right) \cdot z_{k,i}}{\overline{\ell}^{-1}\left(\lim_{a \to 0} a \cdot z_{k,i}\right)} = \frac{\overline{\ell}^{-1}\left(\sum_{i=1}^{n} \overline{\ell}(\overline{p}_{k,i})\right) \cdot z_{k,i}}{\overline{p}_{k,i}} \tag{37}$$

for all $i \in [n]$ and $k \in [K]$. Next, we show that $\boldsymbol{Q}$ is a subdifferential of $\overline{\psi}\left(\overline{\boldsymbol{P}}\right)$. By the definition of $\overline{\psi} : \overline{\mathbb{R}^{Kn}} \to \overline{\mathbb{R}}$ and for any $\boldsymbol{P} \in \overline{\mathbb{R}^{Kn}}$, we have

$$\overline{\psi}(\boldsymbol{P}) = \overline{\ell}^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \overline{\ell}\left(p_{k,i}\right)\right) \text{ and } \frac{\partial}{\partial p_{k,i}} \overline{\psi}(\boldsymbol{P}) = \frac{\overline{\ell}'\left(p_{k,i}\right)}{\overline{\ell}'\left(\overline{\ell}^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \overline{\ell}\left(p_{k,i}\right)\right)\right)},$$

for all $i \in [n]$ and $k \in [K]$. Therefore, according to Equation (36), it implies that $\boldsymbol{Q}$ is in the subdifferential of $\overline{\psi}\left(\overline{\boldsymbol{P}}\right)$ such that $\boldsymbol{Q} \in \partial\overline{\psi}\left(\overline{\boldsymbol{P}}\right)$. Hence, we can further apply the property of Fenchel–Young inequality in the convex conjugate, obtaining

$$\overline{\psi}^*(\boldsymbol{Q}) = \sup_{\boldsymbol{P} \in \overline{\mathbb{R}^{Kn}}} \langle \boldsymbol{P}, \boldsymbol{Q} \rangle - \overline{\psi}(\boldsymbol{P}) = \left\langle \overline{\boldsymbol{P}}, \boldsymbol{Q} \right\rangle - \overline{\psi}\left(\overline{\boldsymbol{P}}\right).$$

As a result, by Equation (37) and the definition of $\overline{\psi}(\cdot)$, we can write

$$\begin{aligned} \overline{\psi}^*(\boldsymbol{Q}) &= \left\langle \overline{\boldsymbol{P}}, \boldsymbol{Q} \right\rangle - \overline{\psi}\left(\overline{\boldsymbol{P}}\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \overline{p}_{k,i} \cdot \frac{\overline{\ell}^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \overline{\ell}(\overline{p}_{k,i})\right) \cdot z_{k,i}}{\overline{p}_{k,i}} - \overline{\ell}^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \overline{\ell}(\overline{p}_{k,i})\right) = 0. \end{aligned}$$

Finally, by Dudík et al. (2022, Proposition 8.5), we have $\psi^*(\boldsymbol{Q}) = \overline{\psi}^*(\boldsymbol{Q}) = 0$. This completes the proof of the lemma. ■

Next, we show in the following lemma that $\bar{\boldsymbol{Q}} = \lim_{t \to \infty} \boldsymbol{Q}_t$ under Assumption 2.

Lemma 32 *Under Assumption 3 and 2, the gradient descent dual variable $\boldsymbol{Q}_t$ converges to $\bar{\boldsymbol{Q}}$ when $t \to \infty$. It gives $\bar{q}_{k,i} = \lim_{t \to \infty} q_{t,k,i} = g(z_{k,i}) > 0$ for all $i \in [n]$ and $k \in [K]$ with some $z_{k,i} \in (0, 1]$ and $\sum_{i=1}^{n}\sum_{k=1}^{K} z_{k,i} = 1$.*

Proof (of Lemma 32) Analogous to Lemma 21, according to Lemma 26, we have $\lim_{t\to\infty} \tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}_t = \tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}$, and $\tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}$ is the same for all $\bar{\boldsymbol{Q}} \in \arg\min_{\psi^*(\boldsymbol{Q})\le 0} F(\boldsymbol{Q})$. Based on the definition of $q_{t,k,i}$ such that

$$q_{t,k,i} = \frac{\partial}{\partial p_{t,k,i}}\psi\left(\boldsymbol{P}_t\right) = \frac{\ell'\left(p_{t,k,i}\right)}{\ell'\Big(\ell^{-1}\Big(\sum_{i=1}^n \sum_{k=1}^K \ell\left(p_{t,k,i}\right)\Big)\Big)} = \frac{\ell'\left(p_{t,k,i}\right)}{\ell'\left(\psi\left(\boldsymbol{P}_t\right)\right)}, \tag{38}$$

and considering $\ell'(\cdot)$ is an increasing function with $p_{t,k,i} \le \psi\left(\boldsymbol{P}_t\right)$ for all $i \in [n]$ and $k \in [K]$ and $t \ge 0$, it follows that $0 < q_{t,k,i} \le 1$; hence, we can conclude that $\lim_{t\to\infty} \tilde{\boldsymbol{X}}^\top \boldsymbol{C}\boldsymbol{Q}_t = \tilde{\boldsymbol{X}}^\top \boldsymbol{C} \lim_{t\to\infty} \boldsymbol{Q}_t = \tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}$. Next, by Assumption 2, $\boldsymbol{X}^\top$ has full column rank, and $\boldsymbol{X}\boldsymbol{X}^\top \succ \boldsymbol{0}$. Therefore, we can multiply the pseudo-inverse of $\tilde{\boldsymbol{X}}^\top \boldsymbol{C}$ on both sides of $\tilde{\boldsymbol{X}}^\top \boldsymbol{C} \lim_{t\to\infty} \boldsymbol{Q}_t = \tilde{\boldsymbol{X}}^\top \boldsymbol{C}\bar{\boldsymbol{Q}}$, which implies that $\lim_{t\to\infty} \boldsymbol{Q}_t = \bar{\boldsymbol{Q}}$. Finally, by the definition of $q_{t,k,i}$ in Equation (38) and the primal convergence in Lemma 25 such that $\lim_{t\to\infty} \sum_{i=1}^n \sum_{k=1}^K \ell\left(p_{t,k,i}\right) = 0$, we have

$$\lim_{t\to\infty} q_{t,k,i} = \lim_{t\to\infty} \frac{\ell'\Big(\ell^{-1}\Big(\ell\left(p_{t,k,i}\right)\Big)\Big)}{\ell'\Big(\ell^{-1}\Big(\sum_{i=1}^n \sum_{k=1}^K \ell\left(p_{t,k,i}\right)\Big)\Big)} = \lim_{a\to 0} \frac{\ell'\big(\ell^{-1}\big(a\cdot z_{k,i}\big)\big)}{\ell'\big(\ell^{-1}\big(a\big)\big)} = g(z_{k,i}),$$

where we let $a = \sum_{i=1}^n \sum_{k=1}^K \ell\left(p_{t,k,i}\right)$, $\ell\left(p_{t,k,i}\right) = a\cdot z_{k,i}$, and $z_{k,i} = \lim_{t\to\infty} \frac{\ell\left(p_{t,k,i}\right)}{\sum_{i=1}^n \sum_{k=1}^K \ell\left(p_{t,k,i}\right)}$. Finally, Assumption 1 guarantees that $\bar{q}_{k,i} = g(z_{k,i}) > 0$. This completes the proof of the lemma. ■

Armed with Lemma 31 and 32, we can prove Lemma 30.

Proof (of Lemma 30) We start with the original convex program in (9)

$$\bar{\boldsymbol{Q}} \in \arg\min_{\psi^*(\boldsymbol{Q})\le 0} F\left(\boldsymbol{Q}\right).$$

By complementary slackness in KKT conditions, if the constraint is inactive such that $\psi^*\left(\boldsymbol{Q}\right) < 0$, we get an invalid solution $\bar{\boldsymbol{Q}} = \boldsymbol{0}$. Therefore, the constraint is active and $\bar{\boldsymbol{Q}}$ satisfies $\psi^*\left(\boldsymbol{Q}\right) = 0$. In other words, we can write

$$\bar{\boldsymbol{Q}} \in \arg\min_{\psi^*(\boldsymbol{Q})= 0} F\left(\boldsymbol{Q}\right). \tag{39}$$

Now, Lemma 32 directly implies that $\bar{\boldsymbol{Q}}$ must satisfy $\sum_{i=1}^n \sum_{k=1}^K g^{-1}(\bar{q}_{k,i}) = 1$ and $\bar{q}_{k,i} > 0$ for all $i \in [n]$ and $k \in [K]$. This means that we can further tighten (39) to obtain

$$\begin{aligned} &\bar{\boldsymbol{Q}} \in \arg\min_{\boldsymbol{Q}\in\mathbb{R}^{Kn}} F\left(\boldsymbol{Q}\right) \\ &\text{subject to} \quad \psi^*\left(\boldsymbol{Q}\right) = 0, \\ &\qquad -q_{k,i} < 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K], \\ &\text{and} \quad 1 - \sum_{i=1}^n \sum_{k=1}^K g^{-1}\left(q_{k,i}\right) = 0. \end{aligned} \tag{40}$$

Next, Lemma 31 tells us that $1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}(q_i) = 0 \implies \psi^*(\boldsymbol{Q}) = 0$, meaning that the constraint $\psi^*(\boldsymbol{Q}) = 0$ is redundant and can simply be omitted, leading to the simplified program

$$\bar{\boldsymbol{Q}} \in \arg\min_{\boldsymbol{Q} \in \mathbb{R}^{Kn}} F(\boldsymbol{Q}) \tag{41}$$

$$\text{subject to} \qquad -q_{k,i} < 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K],$$

$$\text{and} \quad 1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}\right) = 0.$$

The final step is to derive an auxiliary convex program

$$\boldsymbol{Q}^\star \in \arg\min_{\boldsymbol{Q} \in \mathbb{R}^{Kn}} \underbrace{\frac{1}{2} \boldsymbol{Q}^\top \boldsymbol{C} \tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}}^\top \boldsymbol{C} \boldsymbol{Q}}_{F(\boldsymbol{Q})} \tag{42}$$

$$\text{subject to} \qquad -q_{k,i} < 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K],$$

$$\text{and} \quad 1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}\right) \leq 0.$$

Note that in the above, we have relaxed the equality constraint $1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}\right) = 0$ to an inequality constraint, $1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}\right) \leq 0$.

To complete the proof, we need to show that any optimal solution to (42) satisfies $\sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}\right) = 1$. From (41), this directly implies that the set of optima of (9) and (42) are identical. We now show this final step. It is necessary and sufficient for any optimal solution $\boldsymbol{Q}^\star$ to the auxiliary convex program (42) to satisfy its KKT conditions, listed below

$$-q_{k,i} < 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K], \tag{43a}$$

$$1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}\right) \leq 0, \tag{43b}$$

$$\lambda_{k,i} \geq 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K], \tag{43c}$$

$$\mu \geq 0, \tag{43d}$$

$$-\lambda_{k,i} q_{k,i} = 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K], \tag{43e}$$

$$\mu \left(1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}\right) \right) = 0, \tag{43f}$$

$$\operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right) \boldsymbol{X} \boldsymbol{X}^\top \operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right) \boldsymbol{q}_k - \boldsymbol{\lambda}_k - \mu \left[g^{-1}\right]'(\boldsymbol{q}_k) = \boldsymbol{0} \qquad \text{for all } k \in [K]. \tag{43g}$$

First, we claim that any optimal solution $\boldsymbol{Q}^\star$ needs to satisfy $1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}^\star\right) = 0$. This follows because we need to set $\mu > 0$ for a valid solution; together with Equation (43f), this implies that we need $1 - \sum_{i=1}^{n} \sum_{k=1}^{K} g^{-1}\left(q_{k,i}^\star\right) = 0$. To see why we need to set $\mu > 0$, consider the alternative choice $\mu = 0$ for all $k \in [K]$. Note that Equations (43a) and (43e) together also require $\boldsymbol{\lambda}_k = \mathbf{0}$. Equation (43g) would then become

$$\boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}\big(\boldsymbol{c}_k^{-1}\big)\boldsymbol{q}_k^\star = \mathbf{0} \iff \mathrm{diag}\big(\boldsymbol{c}_k^{-1}\big)\boldsymbol{q}_k^\star = \mathbf{0} \iff \boldsymbol{q}_k^\star = \mathbf{0},$$

where the first *iff* statement follows because we have assumed that $\boldsymbol{X}\boldsymbol{X}^\top \succ \mathbf{0}$. However, this $\boldsymbol{q}_k^\star$ is not a valid solution as it violates Equation (43a). Hence, we can conclude both $\bar{\boldsymbol{Q}}$ and $\boldsymbol{Q}^\star$ satisfy $1 - \sum_{i=1}^n \sum_{k=1}^K g^{-1}\left(q_{k,i}\right) = 0$, and $\boldsymbol{Q}^\star = \bar{\boldsymbol{Q}}$. This completes the proof of the lemma. ■

With the auxiliary convex program, we can now prove Theorem 10. In the proof, we show that we have the exact characteristic equations in $\bar{\boldsymbol{q}}_k$ for each k. Therefore, the primal and dual rate for each $k \in [K]$ is the same as Theorem 6.

Proof (Proof of Theorem 10) Our proof starts with the auxiliary convex program (35) and identifies a necessary set of characteristic equations that the optimal solution $\bar{\boldsymbol{Q}}$ needs to satisfy. The KKT conditions for this convex program are given in Equation (43). Lemma 30 postulates that any optimal solution must satisfy $\boldsymbol{\lambda}_k = \mathbf{0}$ and $\mu > 0$ for all $k \in [K]$; therefore, it is necessary for $\bar{\boldsymbol{Q}}$ to satisfy the following characteristic equations for each k

$$\mathrm{diag}\big(\boldsymbol{c}_k^{-1}\big)\boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}\big(\boldsymbol{c}_k^{-1}\big)\bar{\boldsymbol{q}}_k = \mu\left[g^{-1}\right]'(\bar{\boldsymbol{q}}_k), \tag{44a}$$

$$\mu > 0, \tag{44b}$$

$$\sum_{i=1}^n \sum_{k=1}^K g^{-1}\left(\bar{q}_{k,i}\right) = 1. \tag{44c}$$

It is easy to see that for each value of k, the characteristic equations in Equation (44) are identical to the characteristic equations for the binary case (23). Therefore, the rates of convergence of the dual and primal solutions are identical to the binary case for every value of k. This completes the proof. ■

C.2 Proof of Theorem 12 (Exact Equivalence to Simplex MNI for Cross-entropy Loss under Assumption 4)

Before we prove Theorem 12 for cross-entropy loss under Assumption 4, we state and prove two lemmas that we need to analyze the constraint $\psi^*(\boldsymbol{Q}) \le 0$. Note that since the ψ function for cross-entropy loss is different from other multiclass losses, we apply a different proof technique for the proof of this part. First, we utilize the following lemma to analyze the domain of $\psi^*(\boldsymbol{Q})$ under Assumption 4.

Lemma 33 *Under Assumption 4, for any* $\boldsymbol{Q} = \left(\boldsymbol{q}_1^\top, \cdots, \boldsymbol{q}_K^\top\right)^\top \in \mathrm{dom}\,\psi^*$, *where* $\boldsymbol{q}_k \in \mathbb{R}^n$ *for all* $k \in [K]$, *we have* $\boldsymbol{Q} = \nabla\psi(\boldsymbol{P}^*)$ *satisfying* $0 < q_{k,i} < 1$, $c_{y_i,i}^{-1} q_{y_i,i} = -\sum_{k \neq y_i}^K c_{k,i}^{-1} q_{k,i}$ *for all* $i \in [n]$ *and* $k \in [K]$, *and* $\sum_{k=1}^K \mathbf{1}^\top \boldsymbol{q}_k \ge 1$, *for some* $\boldsymbol{P}^* = \left(\boldsymbol{p}_1^{*\top}, \cdots, \boldsymbol{p}_K^{*\top}\right)^\top \in \mathbb{R}^{Kn}$, *where* $\boldsymbol{p}_k^* \in \mathbb{R}^n$ *for all* $k \in [K]$.

Proof (of Lemma 33) Under Assumption 4, we have the definition of ψ such that

$$\psi\left(\boldsymbol{P}\right)=\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{p_{k,i}\right\}_{k=1}^{K}\right)\right)=\ell^{-1}\left(\sum_{i=1}^{n}\ln\left(1+\sum_{k\neq y_i}^{K}\exp\left(c_{y_i,i}p_{y_i,i}-c_{k,i}p_{k,i}\right)\right)\right).$$

For simplicity, we denote

$$\delta_i\coloneqq\sum_{k\neq y_i}^{K}\exp\left(c_{y_i,i}p_{y_i,i}-c_{k,i}p_{k,i}\right),\ \text{and}\ \delta_{k,i}\coloneqq\exp\left(c_{y_i,i}p_{y_i,i}-c_{k,i}p_{k,i}\right),\tag{45}$$

for all $i\in[n]$ and $k\in[K]$. By the definition of $\mathfrak{L}$, we also have $\delta_i=\exp\left(\mathfrak{L}\left(\left\{p_{k,i}\right\}_{k=1}^{K}\right)\right)-1$ for all $i\in[n]$. For any $\boldsymbol{Q}=\left(\boldsymbol{q}_1^{\top},\cdots,\boldsymbol{q}_K^{\top}\right)^{\top}\in\operatorname{dom}\psi^*$, where $\boldsymbol{q}_k\in\mathbb{R}^n$ for all $k\in[K]$, we have $\psi^*\left(\boldsymbol{Q}\right)=\sup_{\boldsymbol{P}\in\mathbb{R}^{Kn}}\langle\boldsymbol{P},\boldsymbol{Q}\rangle-\psi\left(\boldsymbol{P}\right)=\langle\boldsymbol{P}^*,\boldsymbol{Q}\rangle-\psi\left(\boldsymbol{P}^*\right)$, where $\boldsymbol{Q}=\nabla\psi\left(\boldsymbol{P}^*\right)$ for some $\boldsymbol{P}^*=\left(\boldsymbol{p}_1^{*\top},\cdots,\boldsymbol{p}_K^{*\top}\right)^{\top}\in\mathbb{R}^{Kn}$, where $\boldsymbol{p}_k^*\in\mathbb{R}^n$. Therefore, we can conclude that

$$q_{k,i}=\frac{\frac{\partial\mathfrak{L}\left(\left\{p_{k,i}^*\right\}_{k=1}^{K}\right)}{\partial p_{k,i}^*}}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{p_{k,i}^*\right\}_{k=1}^{K}\right)\right)\right)}=\begin{cases}\frac{c_{y_i,i}\delta_i}{(1+\delta_i)\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{p_{k,i}^*\right\}_{k=1}^{K}\right)\right)\right)} & k=y_i\\ \frac{-c_{k,i}\delta_{k,i}}{(1+\delta_i)\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{p_{k,i}^*\right\}_{k=1}^{K}\right)\right)\right)} & k\neq y_i,\end{cases}\tag{46}$$

for all $i\in[n]$ and $k\in[K]$. Since we have $\ell'>0$ and the simplex labeling that $c_{k,i}=\begin{cases}\frac{K-1}{K} & k=y_i\\ -\frac{1}{K} & k\neq y_i\end{cases}$ in Assumption 4, we can conclude that $q_{y_i,i}\geq q_{k,i}$, $q_{k,i}>0$, $c_{y_i,i}^{-1}q_{y_i,i}=-\sum_{k\neq y_i}^{K}c_{k,i}^{-1}q_{k,i}$ and $\sum_{k=1}^{K}q_{k,i}=\frac{\delta_i}{(1+\delta_i)\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{p_{k,i}^*\right\}_{k=1}^{K}\right)\right)\right)}$ for all $i\in[n]$ and $k\in[K]$.

Next, since logistic loss is used for ℓ in Assumption 4, we get $\ell'\left(\ell^{-1}\left(z\right)\right)=\frac{\exp(z)-1}{\exp(z)}$ which is an increasing sub-additive function (Ji and Telgarsky, 2021, Proof of Lemma 14). Hence, we can have implication from Equation (46) that

$$q_{y_i,i}=\frac{c_{y_i,i}\ell'\left(\ell^{-1}\left(\mathfrak{L}\left(\left\{p_{k,i}\right\}_{k=1}^{K}\right)\right)\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{p_{k,i}\right\}_{k=1}^{K}\right)\right)\right)}\ \text{for all}\ i\in[n].\tag{47}$$

Moreover, since we know $\mathfrak{L}\left(\left\{p_{k,i}\right\}_{k=1}^{K}\right)\leq\sum_{i=1}^{n}\mathfrak{L}\left(\left\{p_{k,i}\right\}_{k=1}^{K}\right)$, $c_{y_i,i}\leq 1$, and $q_{y_i,i}\geq q_{k,i}$ for all $i\in[n]$ and $k\in[K]$, these conditions imply $q_{k,i}\leq q_{y_i,i}\leq c_{y_i,i}<1$ for all $i\in[n]$ and $k\in[K]$. Next, by Equation (47), we let $\boldsymbol{A}\coloneqq\sum_{i=1}^{n}c_{y_i,i}^{-1}q_{y_i,i}=\sum_{i=1}^{n}\sum_{k\neq y_i}^{K}\left(-c_{k,i}^{-1}q_{k,i}\right)=\frac{\sum_{i=1}^{n}\ell'\left(\ell^{-1}\left(\mathfrak{L}\left(\left\{p_{k,i}\right\}_{k=1}^{K}\right)\right)\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{p_{k,i}\right\}_{k=1}^{K}\right)\right)\right)}\geq 1$, where we reuse the property that $\ell'\left(\ell^{-1}\left(z\right)\right)$ is an increasing sub-additive function. Followed by the operation in Wang et al. (2021a, Equation 31), we have

$$\begin{aligned}\boldsymbol{A} = \frac{K-1}{K}\boldsymbol{A} + \frac{1}{K}\boldsymbol{A} &= \frac{K-1}{K}\sum_{i=1}^{n} c_{y_i,i}^{-1} q_{y_i,i} + \frac{1}{K}\sum_{i=1}^{n}\sum_{k\neq y_i}^{K}\left(-c_{k,i}^{-1} q_{k,i}\right)\\ &= \sum_{i=1}^{n} q_{y_i,i} + \sum_{i=1}^{n}\sum_{k\neq y_i}^{K} q_{k,i}\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} q_{k,i} = \sum_{k=1}^{K}\mathbf{1}^T \boldsymbol{q}_k \geq 1.\end{aligned}$$

■

Next, we introduce Lemma 34 that shows that $\sum_{k=1}^{K}\mathbf{1}^\top \boldsymbol{q}_k \leq 1$ implies the feasibility of the convex conjugacy feasibility constraint, i.e. $\psi^*(\boldsymbol{Q}) \leq 0$.

Lemma 34 *Under Assumption 4, for any* $\boldsymbol{Q} \in \mathbb{R}^{Kn} \in \mathrm{dom}\,\psi^*$ *that satisfies* $\sum_{k=1}^{K}\mathbf{1}^\top \boldsymbol{q}_k \leq 1$ *also satisfies the convex conjugacy feasibility constraint* $\psi^*(\boldsymbol{Q}) \leq 0$.

Proof (of Lemma 34) Under Assumption 4, we have

$$\begin{aligned}\psi(\boldsymbol{P}) &= \ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\{p_{k,i}\}_{k=1}^{K}\right)\right)\\ &= \ell^{-1}\left(\sum_{i=1}^{n}\ln\left(1 + \sum_{k\neq y_i}^{K}\exp\left(c_{y_i,i}p_{y_i,i} - c_{k,i}p_{k,i}\right)\right)\right).\end{aligned}$$

By Lemma 33, for any $\boldsymbol{Q} = \left(\boldsymbol{q}_1^\top, \cdots, \boldsymbol{q}_K^\top\right)^\top \in \mathrm{dom}\,\psi^*$, where $\boldsymbol{q}_k \in \mathbb{R}^n$ for all $k \in [K]$, we have $\boldsymbol{Q} = \nabla\psi(\boldsymbol{P}^*)$ for some $\boldsymbol{P}^* = \left(\boldsymbol{p}_1^{*\top}, \cdots, \boldsymbol{p}_K^{*\top}\right)^\top \in \mathbb{R}^{Kn}$. Also, we reuse the setup in Equations (45) and (46) and have

$$q_{k,i} = \begin{cases}\frac{c_{y_i,i}\delta_i}{(1+\delta_i)\ell'\left(\psi(\boldsymbol{P}^*)\right)} & k = y_i\\ \frac{-c_{k,i}\delta_{k,i}}{(1+\delta_i)\ell'\left(\psi(\boldsymbol{P}^*)\right)} & k \neq y_i,\end{cases} \tag{48}$$

for all $i \in [n]$ and $k \in [K]$. Next, in Lemma 27, we already show that $\psi^*(\boldsymbol{Q}) \leq 0$ if $\psi(\boldsymbol{P}^*) \leq 0$. Therefore, we only need to check the case when $\psi(\boldsymbol{P}^*) > 0$, and we have

$$\begin{aligned}\psi^*(\boldsymbol{Q}) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{k,i}^* q_{k,i} - \psi(\boldsymbol{P}) &= \sum_{i=1}^{n}\frac{c_{y_i,i}p_{y_i,i}^*\delta_i + \sum_{k\neq y_i}^{K}\left(-c_{k,i}p_{k,i}^*\right)\delta_{k,i}}{(1+\delta_i)\,\ell'\left(\psi(\boldsymbol{P}^*)\right)} - \psi\left(\boldsymbol{P}^*\right)\\ &= \sum_{i=1}^{n}\frac{\sum_{k\neq y_i}^{K}\left(c_{y_i,i}p_{y_i,i}^* - c_{k,i}p_{k,i}^*\right)\delta_{k,i}}{(1+\delta_i)\,\ell'\left(\psi(\boldsymbol{P}^*)\right)} - \psi\left(\boldsymbol{P}^*\right)\end{aligned} \tag{49}$$

On the other hand, we have

$$
\begin{aligned}
q_{y_i,i} &= \frac{c_{y_i,i}\delta_i}{(1+\delta_i)\,\ell'\left(\psi\left(\boldsymbol{P}^*\right)\right)} \\
\iff q_{y_i,i}\ell'\Big(\psi\left(\boldsymbol{P}^*\right)\Big) &= c_{y_i,i}\ell'\Bigg(\ell^{-1}\Bigg(\mathfrak{L}\left(\left\{p^*_{k,i}\right\}_{k=1}^{K}\right)\Bigg)\Bigg) \\
\iff c^{-1}_{y_i,i}q_{y_i,i}\ell'\Big(\psi\left(\boldsymbol{P}^*\right)\Big) &= \ell'\Bigg(\ln\Bigg(\sum_{k\neq y_i}^{K}\delta_{k,i}\Bigg)\Bigg) \\
\iff \left[\ell'\right]^{-1}\Big(c^{-1}_{y_i,i}q_{y_i,i}\ell'\Big(\psi\left(\boldsymbol{P}^*\right)\Big)\Big) &= \ln\Bigg(\sum_{k\neq y_i}^{K}\delta_{k,i}\Bigg) \\
\Rightarrow \left[\ell'\right]^{-1}\Big(\ell'\Big(\psi\left(\boldsymbol{P}^*\right)\Big)\Big) &\geq \ln\Bigg(\sum_{k\neq y_i}^{K}\delta_{k,i}\Bigg) \\
\Rightarrow \psi\left(\boldsymbol{P}^*\right) \geq \ln\left(\delta_{k,i}\right) &= c_{y_i,i}p^*_{y_i,i} - c_{k,i}p^*_{k,i} \quad \text{for all } i\in[n],\ k\neq y_i,
\end{aligned} \tag{50}
$$

where the second equality comes from $\frac{\delta_i}{1+\delta_i} = \ell'\Bigg(\ell^{-1}\Bigg(\mathfrak{L}\left(\left\{p^*_{k,i}\right\}_{k=1}^{K}\right)\Bigg)\Bigg)$, the first inequality derives from $c^{-1}_{y_i,i}q_{y_i,i} < 1$ in Lemma 33, and the last inequality holds because $\ln(\cdot)$ is an increasing function and $\delta_{k,i} > 0$. Next, by introducing Equation (50) into Equation (49), we get

$$
\begin{aligned}
\psi^*\left(\boldsymbol{Q}\right) &= \sum_{i=1}^{n}\Big(\sum_{k\neq y_i}^{K}\Big(c_{y_i,i}p^*_{y_i,i} - c_{k,i}p^*_{k,i}\Big)\delta_{k,i}\Big)\left(1+\delta_i\right)\ell'\Big(\psi\left(\boldsymbol{P}^*\right)\Big) - \psi\left(\boldsymbol{P}^*\right) \\
&\leq \sum_{i=1}^{n}\frac{\sum_{k\neq y_i}^{K}\psi\left(\boldsymbol{P}^*\right)\delta_{k,i}}{(1+\delta_i)\,\ell'\left(\psi\left(\boldsymbol{P}^*\right)\right)} - \psi\left(\boldsymbol{P}^*\right) \\
&= \psi\left(\boldsymbol{P}^*\right)\left(\sum_{i=1}^{n}\frac{\sum_{k\neq y_i}^{K}\delta_{k,i}}{(1+\delta_i)\,\ell'\left(\psi\left(\boldsymbol{P}^*\right)\right)} - 1\right) \\
&= \psi\left(\boldsymbol{P}^*\right)\left(\sum_{i=1}^{n}\sum_{k=1}^{K}q_{k,i} - 1\right) = \psi\left(\boldsymbol{P}^*\right)\left(\sum_{k=1}^{K}\boldsymbol{1}^\top\boldsymbol{q}_k - 1\right) \leq 0,
\end{aligned}
$$

where we reuse $\frac{\delta_i}{(1+\delta_i)\ell'\left(\psi\left(\boldsymbol{P}^*\right)\right)} = \sum_{k=1}^{K}q_{k,i}$ in the second to the last equality. The last inequality derives from the assumption in the lemma statement, $\sum_{k=1}^{K}\boldsymbol{1}^\top\boldsymbol{q}_k \leq 1$, and because we are in the case where $\psi\left(\boldsymbol{P}^*\right) > 0$. This completes the proof of the lemma. ■

Armed with Lemma 33 and 34, we can prove the Part 2 of Theorem 12.

Proof (of Theorem 12 Part 2) Note that the constraint in Equation (9) ($\psi^*\left(\bar{\boldsymbol{Q}}\right) \leq 0$) implicitly implies that $\bar{\boldsymbol{Q}} \in \operatorname{dom}\psi^*$. Therefore, by Lemma 33 the following constraints are implied

$$\bar{q}_{k,i} > 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K],$$
$$c_{y_i,i}^{-1} q_{y_i,i} = -\sum_{k \neq y_i}^{K} c_{k,i}^{-1} q_{k,i} \qquad \text{for all } i \in [n], \text{ and}$$
$$\sum_{k=1}^{K} \mathbf{1}^\top \boldsymbol{q}_k \geq 1.$$

We now show that a particular solution from the following convex program is also a solution in the original convex program (9). We define a reformulated convex program

$$\tilde{\boldsymbol{Q}} \in \arg\min_{\boldsymbol{Q} \in \mathbb{R}^{Kn}} \underbrace{\frac{1}{2} \boldsymbol{Q}^\top \boldsymbol{C} \tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}}^\top \boldsymbol{C} \boldsymbol{Q}}_{F(\boldsymbol{Q})} \tag{51}$$
$$\text{subject to} \qquad -q_{k,i} < 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K],$$
$$c_{y_i,i}^{-1} q_{y_i,i} = -\sum_{k \neq y_i}^{K} c_{k,i}^{-1} q_{k,i} \qquad \text{for all } i \in [n],$$
$$\text{and} \quad 1 - \sum_{k=1}^{K} \mathbf{1}^\top \boldsymbol{q}_k \leq 0.$$

It is necessary and sufficient for any optimal solution $\tilde{\boldsymbol{Q}}$ to this reformulated convex program (51) to satisfy its KKT conditions, listed below

$$-q_{k,i} < 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K], \tag{52a}$$
$$c_{y_i,i}^{-1} q_{y_i,i} + \sum_{k \neq y_i}^{K} c_{k,i}^{-1} q_{k,i} = 0 \qquad \text{for all } i \in [n], \tag{52b}$$
$$1 - \sum_{k=1}^{K} \mathbf{1}^\top \boldsymbol{q}_k \leq 0, \tag{52c}$$
$$\lambda_{k,i} \geq 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K], \tag{52d}$$
$$\delta_i \in \mathbb{R} \qquad \text{for all } i \in [n], \tag{52e}$$
$$\mu \geq 0, \tag{52f}$$
$$-\lambda_{k,i} q_{k,i} = 0 \qquad \text{for all } i \in [n] \text{ and } k \in [K], \tag{52g}$$
$$\mu \left(1 - \sum_{k=1}^{K} \mathbf{1}^\top \boldsymbol{q}_k \right) = 0, \tag{52h}$$
$$\operatorname{diag}\big(\boldsymbol{c}_k^{-1}\big) \boldsymbol{X} \boldsymbol{X}^\top \operatorname{diag}\big(\boldsymbol{c}_k^{-1}\big) \boldsymbol{q}_k - \boldsymbol{\lambda}_k + \operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right) \boldsymbol{\delta} - \mu \mathbf{1} = \mathbf{0} \qquad \text{for all } k \in [K]. \tag{52i}$$

Then we can pick a candidate solution $\tilde{\boldsymbol{q}}_k = \frac{\operatorname{diag}(\boldsymbol{c}_k)\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{c}_k}{\sum_{k=1}^{K} \boldsymbol{c}_k^\top \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{c}_k}$ satisfying all KKT conditions such that

- The primal feasibility equations, Equation (52a), is satisfied by theorem statement that $c_{k,i}\beta_{k,i} > 0$, and Equation (52b), is satisfied because of the following: Followed by Wang et al. (2021a, Theorem 1 Step 2), we let $\boldsymbol{g}_i \in \mathbb{R}^n$ denote the ith row of $(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$ for all $i \in [n]$. Then for ith element of $\tilde{q}_k$, we have $\tilde{q}_{k,i} = \frac{c_{k,i}\boldsymbol{g}_i^T \boldsymbol{c}_k}{\sum_{k=1}^K \boldsymbol{c}_k^T \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{c}_k}$. Thus, for all $i \in [n]$, we have

$$c_{y_i,i}^{-1}\tilde{q}_{y_i,i} + \sum_{k \neq y_i}^{K} c_{k,i}^{-1}\tilde{q}_{k,i} = \frac{\boldsymbol{g}_i^T(\boldsymbol{c}_{y_i} + \sum_{k\neq y_i}^K \boldsymbol{c}_k)}{\sum_{k=1}^K \boldsymbol{c}_k^T \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k} = \frac{\boldsymbol{g}_i^T\left(\sum_{k=1}^K \boldsymbol{c}_k\right)}{\sum_{k=1}^K \boldsymbol{c}_k^T \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k} = 0,$$

  where the last equality followed by the simplex definition of $\boldsymbol{c}_k$. Equation (52c) is satisfied such that $1 - \sum_{k=1}^K \mathbf{1}^\top \tilde{\boldsymbol{q}}_k = 0$.

- The dual feasibility equations, Equations (52d) and (52f), are satisfied by setting $\boldsymbol{\lambda}_k = \mathbf{0}$, and $\mu = \frac{1}{\sum_{k=1}^K \boldsymbol{c}_k^\top \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{c}_k}$.

- The complementary slackness equations, Equations (52g) and (52h), are satisfied because $\boldsymbol{\lambda}_k = \mathbf{0}$ and $1 - \sum_{k=1}^K \mathbf{1}^\top \tilde{\boldsymbol{q}}_k = 0$.

- Stationary condition is satisfied because we choose $\boldsymbol{\delta} = \mathbf{0}$, and then

$$\begin{aligned}&\operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right)\boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right)\tilde{\boldsymbol{q}}_k - \boldsymbol{\lambda}_k + \operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right)\boldsymbol{\delta} - \mu\mathbf{1} \\ &= \frac{1}{\sum_{k=1}^K \boldsymbol{c}_k^\top \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k}\mathbf{1} - \mathbf{0} + \mathbf{0} - \frac{1}{\sum_{k=1}^K \boldsymbol{c}_k^\top \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k}\mathbf{1} = 0,\end{aligned}$$

  for all $k \in [K]$.

Lastly, by Lemma 34, we have $\psi^*\left(\tilde{\boldsymbol{Q}}\right) \leq 0$ since $\sum_{k=1}^K \mathbf{1}^\top \tilde{\boldsymbol{q}}_k = 1$. Therefore, $\tilde{\boldsymbol{Q}}$ is also a solution in the original convex program (9) that satisfies $\psi^*\left(\tilde{\boldsymbol{Q}}\right) \leq 0$, and we end up with $\bar{\boldsymbol{Q}} = \tilde{\boldsymbol{Q}}$. As a result, by Lemma 9, we have $\bar{\boldsymbol{w}}_k = \lim_{t\to\infty} \frac{\boldsymbol{w}_{k,t}}{\left\|\boldsymbol{w}_{k,t}\right\|_2} = \frac{\boldsymbol{X}^\top \operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right)\bar{\boldsymbol{q}}_k}{\left\|\boldsymbol{X}^\top \operatorname{diag}\left(\boldsymbol{c}_k^{-1}\right)\bar{\boldsymbol{q}}_k\right\|_2} = \frac{\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k}{\left\|\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k\right\|_2}$ for each k. Therefore, $\bar{\boldsymbol{w}}_k$ parallel to simplex version MNI $\boldsymbol{w}_{\mathsf{simplex},k} = \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{c}_k$ for each k. The proof is complete. ∎

C.3 Convexity and Smoothness Proof

We can directly apply loss functions defined in Assumption 1 in the binary case with the following lemmas that ensure the properties of convexity and β-smoothness with respect to the ℓ_∞ norm of ψ in the multiclass case.

Lemma 35 (From Ji and Telgarsky 2021, Lemma 12) *If $\ell'^2/\left(\ell\ell''\right)$ is increasing on $(-\infty, \infty)$, then ψ is jointly convex under Assumption 3, and is individually convex toward each $\boldsymbol{\xi}_k$ under Assumption 4.*

Proof (of Lemma 35) We discuss the situation under Assumption 3 and Assumption 4 separately. Under Assumption 3, we have the definition of ψ such that

$$\psi\left(\boldsymbol{\Xi}\right) = \ell^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\ell\left(\xi_{k,i}\right)\right),$$

and the gradient $\nabla\psi\left(\boldsymbol{\Xi}\right)$ is defined by

$$\nabla\psi\left(\boldsymbol{\Xi}\right)_{k,i} = \frac{\partial\psi\left(\boldsymbol{\Xi}\right)}{\partial\xi_{k,i}} = \frac{\ell'\left(\xi_{k,i}\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\ell\left(\sum_{k=1}^{K}\xi_{k,i}\right)\right)\right)} = \frac{\ell'\left(\xi_{k,i}\right)}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)},$$

for all $i \in [n]$ and $k \in [K]$. Next, the Hessian $\nabla^2\psi\left(\boldsymbol{\Xi}\right) \in \mathbb{R}^{Kn\times Kn}$ is

$$\nabla^2\psi\left(\boldsymbol{\Xi}\right) = \operatorname{diag}\left(\frac{\ell''\left(\xi_{1,1}\right)}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)},\cdots,\frac{\ell''\left(\xi_{K,n}\right)}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\right) - \frac{\ell''\left(\psi\left(\boldsymbol{\Xi}\right)\right)}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\nabla\psi\left(\boldsymbol{\Xi}\right)\nabla\psi\left(\boldsymbol{\Xi}\right)^{\top}. \tag{53}$$

Note that the Hessian is identical to Ji and Telgarsky (2021, Lemma 12, Equation 24) with additional K dimensions; therefore, the convexity for ψ holds for loss functions under Assumption 3.

Next, we show the convexity proof for cross-entropy loss under Assumption 4. According to the definition of ψ function under Assumption 4, we have

$$\psi\left(\boldsymbol{\Xi}\right) = \ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right),$$

and the gradient $\nabla_{\boldsymbol{\xi}_k}\psi\left(\boldsymbol{\Xi}\right)$ is defined by

$$\frac{\partial\psi\left(\boldsymbol{\Xi}\right)}{\partial\xi_{k,i}} = \frac{\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}}}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)\right)} = \left(\frac{1}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\right)\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}}.$$

Next, the second order of the partial derivatives are

$$\begin{aligned}
\frac{\partial^2\psi\left(\boldsymbol{\Xi}\right)}{\partial\xi_{k,i}^2} &= \left(\frac{1}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\right)\frac{\partial^2\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}^2} \\
&\quad - \frac{\ell''\left(\psi\left(\boldsymbol{\Xi}\right)\right)}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\left(\frac{1}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\right)\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}}\left(\frac{1}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\right)\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}}, \\
\frac{\partial^2\psi\left(\boldsymbol{\Xi}\right)}{\partial\xi_{k,i}\partial\xi_{k,j}} &= -\frac{\ell''\left(\psi\left(\boldsymbol{\Xi}\right)\right)}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\left(\frac{1}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\right)\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}}\left(\frac{1}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\right)\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,j}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,j}}
\end{aligned}$$

for all $i \neq j$ and $k \in [K]$. Hence, we can write the Hessian $\nabla^2_{\boldsymbol{\xi}_k} \psi(\boldsymbol{\Xi}) \in \mathbb{R}^{n \times n}$ as

$$\nabla^2_{\boldsymbol{\xi}_k} \psi(\boldsymbol{\Xi}) = \operatorname{diag}\left(\frac{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,1}\right\}_{k=1}^K\right)}{\partial \xi_{k,1}^2}}{\ell'\left(\psi(\boldsymbol{\Xi})\right)}, \cdots, \frac{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,n}\right\}_{k=1}^K\right)}{\partial \xi_{k,n}^2}}{\ell'\left(\psi(\boldsymbol{\Xi})\right)} \right) - \frac{\ell''\left(\psi(\boldsymbol{\Xi})\right)}{\ell'\left(\psi(\boldsymbol{\Xi})\right)} \nabla_{\boldsymbol{\xi}_k} \psi(\boldsymbol{\Xi}) \nabla_{\boldsymbol{\xi}_k} \psi(\boldsymbol{\Xi})^\top,$$

for all $k \in [K]$. Therefore, it remains to show that for any $v \in \mathbb{R}^n$

$$\sum_{i=1}^n \frac{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)}{\partial \xi_{k,i}^2}}{\ell'\left(\psi(\boldsymbol{\Xi})\right)} v_i \geq \frac{\ell''\left(\psi(\boldsymbol{\Xi})\right)}{\ell'\left(\psi(\boldsymbol{\Xi})\right)} \left(\sum_{i=1}^n \frac{\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)}{\partial \xi_{k,i}}}{\ell'\left(\psi(\boldsymbol{\Xi})\right)} v_i \right)^2. \tag{54}$$

By the Cauchy-Schwarz inequality, we can write

$$\left(\sum_{i=1}^n \frac{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)}{\partial \xi_{k,i}^2}}{\ell'\left(\psi(\boldsymbol{\Xi})\right)} v_i^2 \right) \left(\sum_{i=1}^n \frac{\left[\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)}{\partial \xi_{k,i}}\right]^2}{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)}{\partial \xi_{k,i}^2} \ell'\left(\psi(\boldsymbol{\Xi})\right)} \right) \geq \left(\sum_{i=1}^n \frac{\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)}{\partial \xi_{k,i}}}{\ell'\left(\psi(\boldsymbol{\Xi})\right)} v_i \right)^2 \tag{55}$$

Next, we can show that Equation (54) is satisfied for each $k \in [K]$ by showing

$$\frac{\ell'\left(\psi(\boldsymbol{\Xi})\right)^2}{\ell''\left(\psi(\boldsymbol{\Xi})\right)} = \frac{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)\right)\right)^2}{\ell''\left(\ell^{-1}\left(\sum_{i=1}^n \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)\right)\right)} \geq \left(\sum_{i=1}^n \frac{\left[\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)}{\partial \xi_{k,i}}\right]^2}{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)}{\partial \xi_{k,i}^2}} \right). \tag{56}$$

For cross-entropy loss under Assumption 4, we have

$$\begin{aligned} \psi(\boldsymbol{\Xi}) &= \ell^{-1}\left(\sum_{i=1}^n \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right) \right) \\ &= \ln\left(\exp\left(\sum_{i=1}^n \ln\left(1 + \sum_{k \neq y_i}^K \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right) \right) \right) - 1 \right). \end{aligned}$$

We start from the LHS of Equation (56), we have

$$\frac{\ell'\left(\psi(\boldsymbol{\Xi})\right)^2}{\ell''\left(\psi(\boldsymbol{\Xi})\right)} = \frac{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)\right)\right)^2}{\ell''\left(\ell^{-1}\left(\sum_{i=1}^n \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right)\right)\right)} = \exp\left(\sum_{i=1}^n \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^K\right) \right) - 1, \tag{57}$$

by direct expansion with $\ell(z) = \ln\big(1+\exp(z)\big)$. Next, we work on RHS of Equation (56). For simplicity, we denote

$$\delta_i \coloneqq \sum_{k\neq y_i}^{K} \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right) \text{ and } \delta_{k,i} \coloneqq \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right).$$

Since $\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right) = \ln\left(1 + \sum_{k\neq y_i}^{K}\exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right)\right) = \ln(1+\delta_i)$, we have the first derivative as

$$\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}} = \begin{cases} \frac{c_{y_i,i}\delta_i}{1+\delta_i} & k = y_i \\ \frac{-c_{k,i}\delta_{k,i}}{1+\delta_i} & k \neq y_i \end{cases}, \tag{58}$$

and the second derivative as

$$\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}^2} = \begin{cases} \frac{c_{y_i,i}^2\delta_i}{[1+\delta_i]^2} & k = y_i \\ \frac{c_{k,i}^2\delta_{k,i}\left(1+\sum_{k\neq y_i,k\neq k}^{K}\delta_{k,i}\right)}{[1+\delta_i]^2} & k \neq y_i \end{cases}. \tag{59}$$

Next, by substituting Equations (58) and (59) into the RHS of Equation (56), we get

$$\frac{\left[\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}}\right]^2}{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}^2}} = \begin{cases} \delta_i & k = y_i \\ \frac{\delta_{k,i}}{1+\sum_{k\neq y_i,k\neq k}^{K}\delta_{k,i}} & k \neq y_i \end{cases}.$$

Based on this, we can also derive

$$\frac{\left[\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{y_i,i}}\right]^2}{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{y_i,i}^2}} \geq \frac{\left[\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}}\right]^2}{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}^2}}, \tag{60}$$

for all $k \neq y_i$ by a direct comparison of the two conditions. Moreover, we can also write

$$\frac{\left[\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{y_i,i}}\right]^2}{\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{y_i,i}^2}} = \delta_i = \exp\Big(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\Big) - 1. \tag{61}$$

Therefore, starting from Equation (57), we can show

$$\begin{aligned}
\frac{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)^2}{\ell''\left(\psi\left(\boldsymbol{\Xi}\right)\right)} &= \exp\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right) - 1 \\
&\geq \sum_{i=1}^{n}\left(\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right) - 1\right) \\
&= \left(\sum_{i=1}^{n}\frac{\left[\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{y_i,i}}\right]^2}{\frac{\partial^2\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{y_i,i}^2}}\right) \geq \left(\sum_{i=1}^{n}\frac{\left[\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}}\right]^2}{\frac{\partial^2\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}^2}}\right),
\end{aligned}$$

where the first inequality holds because $f(z) = \exp(z) - 1$ is a super-additive function, the second equality comes from Equation (61), and the last inequality derives from Equation (60). Therefore, Equation (56) holds for all $k \in [K]$. This completes the proof of this lemma. ∎

Lemma 36 (From Ji and Telgarsky 2021, Lemma 13) *Under Assumption 3, if $\ell'' \leq c\ell'$ for some constant $c > 0$, then the smoothness constant $\beta \leq cnK$ for ψ. Particularly, for exponentially-tailed loss, the smoothness constant is $\beta = 1$ for ψ. Under Assumption 4, the cross-entropy loss has the smoothness constant $\beta = 2K^2$ for ψ.*

Proof (of Lemma 36) We follow the proof strategy in Ji and Telgarsky (2021, Lemma 13) and Shalev-Shwartz (2007, Lemma 14), to check the β smoothness of ψ with respect to ℓ_∞ norm. Note that it is sufficient to show for any $\boldsymbol{\Xi}$, $\boldsymbol{v} \in \mathbb{R}^{Kn}$, it holds that $\boldsymbol{v}^\top \nabla^2 \psi(\boldsymbol{\Xi})\,\boldsymbol{v} \leq \beta\left\|\boldsymbol{v}\right\|_\infty^2$. We discuss the situations under Assumption 3 and Assumption 4 separately.

Under Assumption 3, by the definition of ψ, we have

$$\psi\left(\boldsymbol{\Xi}\right) = \ell^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\ell\left(\xi_{k,i}\right)\right).$$

According to the Hessian we derived in Equation (53), it is enough to show that for any $\boldsymbol{v} = \left(v_1, \cdots, v_K\right)^\top \in \mathbb{R}^{Kn}$, where $v_k \in \mathbb{R}^n$ for all $k \in [K]$ and $\boldsymbol{\Xi} \in \mathbb{R}^{Kn}$, we have

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\ell''\left(\xi_{k,i}\right)}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}v_{k,i}^2 \leq \beta\max_{1\leq k\leq K}\max_{1\leq i\leq n}v_{k,i}^2.$$

Note that the condition is identical to Ji and Telgarsky (2021, Lemma 13, Equation 27) with additional summation in K; therefore, the smoothness constant conclusion for ψ is the same as in the binary case for loss functions under Assumption 3.

Next, for cross-entropy loss under Assumption 4, by the definition of ψ, we have

$$\psi(\mathbf{\Xi}) = \ell^{-1}\left(\sum_{i=1}^{n} \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)$$
$$= \ln\left(\exp\left(\sum_{i=1}^{n} \ln\left(1 + \sum_{k\neq y_i}^{K} \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right)\right)\right) - 1\right).$$

For simplicity, we again denote

$$\delta_i := \sum_{k\neq y_i}^{K} \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right), \text{ and } \delta_{k,i} := \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right),$$

and since $\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right) = \ln\left(1 + \sum_{k\neq y_i}^{K} \exp\left(c_{y_i,i}\xi_{y_i,i} - c_{k,i}\xi_{k,i}\right)\right) = \ln(1+\delta_i)$, and we have the first derivative as

$$\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}} = \begin{cases} \frac{c_{y_i,i}\delta_i}{1+\delta_i} & k = y_i \\ \frac{-c_{k,i}\delta_{k,i}}{1+\delta_i} & k \neq y_i \end{cases},$$

and the second derivative as

$$\frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}\partial \xi_{h,i}} = \begin{cases} \frac{c_{y_i,i}^2\delta_i}{[1+\delta_i]^2} & k = h = y_i \\ \frac{-c_{k,i}c_{h,i}\delta_{h,i}}{[1+\delta_i]^2} & k = y_i \text{ and } h \neq y_i \\ \frac{c_{k,i}^2\delta_{k,i}\left(1+\sum_{k\neq y_i, k\neq k}^{K}\delta_{k,i}\right)}{[1+\delta_i]^2} & k = h \neq y_i \\ \frac{-c_{k,i}c_{h,i}\delta_{k,i}\delta_{h,i}}{[1+\delta_i]^2} & k \neq y_i \text{ and } h \neq y_i \end{cases}.$$

By direct comparison in value, we can conclude that the first derivative w.r.t $\xi_{y_i,i}$ upper-bounds all the second derivatives such that $\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{y_i,i}} \geq \frac{\partial^2 \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{k,i}\partial \xi_{h,i}}$ for all $k, h \in [K]$. Also, we have $\frac{\partial \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial \xi_{y_i,i}} = \frac{c_{y_i,i}\left(\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)-1\right)}{\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)}$ and

$$\ell'\left(\psi\left(\mathbf{\Xi}\right)\right) = \ell'\left(\ell^{-1}\left(\sum_{i=1}^{n} \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)\right) = \frac{\exp\left(\sum_{i=1}^{n} \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)-1}{\exp\left(\sum_{i=1}^{n} \mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)},$$

since $\ell'\left(\ell^{-1}\left(z\right)\right)=\frac{\exp(z)-1}{\exp(z)}$. Hence, we can write

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n}\frac{\frac{\partial^2\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{k,i}\partial\xi_{h,i}}}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}v_{k,i}v_{h,i} &\leq \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n}\frac{\frac{\partial\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)}{\partial\xi_{y_i,i}}}{\ell'\left(\psi\left(\boldsymbol{\Xi}\right)\right)}\left|v_{k,i}v_{h,i}\right| \\
&= \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n}\frac{\frac{c_{y_i,i}\left(\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)-1\right)}{\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)}}{\frac{\exp\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)-1}{\exp\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)}}\left|v_{k,i}v_{h,i}\right| \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{K}2\max_{1\leq i\leq n}\left|v_{k,i}v_{h,i}\right| \leq 2K^2\left\|\boldsymbol{v}\right\|_{\infty}^{2},
\end{aligned}$$

where the second inequality holds because $\frac{\sum_{i=1}^{n}\frac{\left(\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)-1\right)}{\exp\left(\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)}}{\frac{\exp\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)-1}{\exp\left(\sum_{i=1}^{n}\mathfrak{L}\left(\left\{\xi_{k,i}\right\}_{k=1}^{K}\right)\right)}} \leq 2$ by Ji and Telgarsky (2021, Proof of Lemma 14). This completes the proof of the lemma. ■

Appendix D. Proofs of Converse Results

In this section, we collect the proofs of the converse results in Section 4.

D.1 Proof of Proposition 13

Proof (of Proposition 13) The proof is divided into two parts.

Proof of Part 1. For the proof of Part 1, we work from Part 1 of the proof of Theorem 6. There, we showed that we require $\bar{\boldsymbol{q}}$ to satisfy a series of characteristic equations; in particular, $\bar{\boldsymbol{q}}$ needs to satisfy the equation

$$\operatorname{diag}(\boldsymbol{y})\,\boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\,\bar{\boldsymbol{q}} = \mu h(\bar{\boldsymbol{q}}) \text{ for some } \mu > 0, \tag{62}$$

where we recall that we defined $h(\cdot) := \left[g^{-1}\right]'(\cdot)$ as shorthand. Our goal is to show that if $\boldsymbol{y}$ is not an exact eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$, then all candidate solutions in the family $\bar{\boldsymbol{q}} \propto \operatorname{diag}(\boldsymbol{y})\,(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$ cannot satisfy Equation (62) for any value of $\mu > 0$. We consider the candidate solution $\bar{\boldsymbol{q}} = \beta \operatorname{diag}(\boldsymbol{y})\,(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$ for some $\beta > 0$. Because we have assumed that $\boldsymbol{X}\boldsymbol{X}^\top$ is full-rank, this is the unique direction of the dual solution that would correspond to a primal solution that is proportional to the MNI. Then, Equation (62) being satisfied for some $\mu > 0$ implies

$$\begin{aligned}\beta \cdot \operatorname{diag}(\boldsymbol{y})\,\boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y}) \cdot \operatorname{diag}(\boldsymbol{y})\,(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} &= \mu \cdot h(\beta \cdot \operatorname{diag}(\boldsymbol{y})\,(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}) \\ \implies \mathbf{1} &= \frac{\mu}{\beta} \cdot h(\beta \operatorname{diag}(\boldsymbol{y})\,(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}).\end{aligned} \tag{63}$$

Now, we recall the properties of $h(\cdot)$ that arise when $g(d) \neq d$, i.e. when the mapping $g(\cdot)$ is not the identity. Because $g(\cdot)$ is strictly convex and increasing, we have that $g^{-1}(\cdot)$ is strictly concave and $h(\cdot) = \left[g^{-1}\right]'(\cdot)$ is therefore *strictly decreasing*. This means that for any $d \neq e$, we have $h(d) \neq h(e)$. Consequently, for Equation (63) to be true for any value of $\mu > 0$, we require all the entries of $\operatorname{diag}(\boldsymbol{y})\,(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$ to be equal. In other words, we need

$$\begin{aligned}&\operatorname{diag}(\boldsymbol{y})\,(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} \propto \mathbf{1} \\ \implies &\operatorname{diag}(\boldsymbol{y})\,(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \gamma\mathbf{1} \text{ for some } \gamma \neq 0 \\ \implies &(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \gamma\boldsymbol{y} \text{ for some } \gamma \neq 0 \\ \implies &\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y} = \frac{1}{\gamma}\boldsymbol{y} \text{ for some } \gamma \neq 0,\end{aligned}$$

implying that $\boldsymbol{y}$ needs to be an exact non-zero eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$. This completes the proof of the first part of the proposition.

Proof of Part 2. For the proof of Part 2, we recall the KKT conditions in Equation (21) in the proof of Lemma 3 for the auxiliary convex program (5), reproduced below for completeness.

$$
\begin{aligned}
-q_i &< 0 && \text{for all } i \in [n], && \text{(64a)}\\
1 - \sum_{i=1}^{n} g^{-1}\left(q_i\right) &\leq 0, && && \text{(64b)}\\
\lambda_i &\geq 0 && \text{for all } i \in [n], && \text{(64c)}\\
\mu &\geq 0, && && \text{(64d)}\\
-\lambda_i q_i &= 0 && \text{for all } i \in [n], && \text{(64e)}\\
\mu\left(1 - \sum_{i=1}^{n} g^{-1}\left(q_i\right)\right) &= 0, && && \text{(64f)}\\
\operatorname{diag}\left(\boldsymbol{y}\right)\boldsymbol{X}\boldsymbol{X}^\top\operatorname{diag}\left(\boldsymbol{y}\right)\boldsymbol{q} - \boldsymbol{\lambda} - \mu h\left(\boldsymbol{q}\right) &= \boldsymbol{0}, && && \text{(64g)}
\end{aligned}
$$

where we denote $h\left(\boldsymbol{q}\right) \coloneqq \left[g^{-1}\right]'\left(\boldsymbol{q}\right)$ as shorthand.

To write our candidate solution $\bar{\boldsymbol{q}}$ in the case where $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{D} = \operatorname{diag}\left(\boldsymbol{d}\right) \neq \boldsymbol{I}$, we define some additional notation. We define $f(d) \coloneqq \frac{h(d)}{d}$ on the domain $(0, 1]$. We note that, because $h(d)$ is strictly decreasing in d and $\frac{1}{d}$ is strictly decreasing in d, $f(d)$ is strictly decreasing in d as well, and is therefore invertible. Also note that $f(d) \in (0, \infty)$.

Then we can pick a candidate $\bar{\boldsymbol{q}}$ such that for every $i \in [n]$, we have

$$
\bar{q}_i = f^{-1}\left(\frac{d_i}{\mu}\right),
$$

where $\mu > 0$ satisfies

$$
\sum_{i=1}^{n} g^{-1}\left(f^{-1}\left(\frac{d_i}{\mu}\right)\right) = 1. \tag{65}
$$

Before verifying the KKT conditions for this candidate solution, let us confirm that it is possible to select a value of $\mu > 0$ satisfying Equation (65). Hiding the dependence on $\boldsymbol{d}$, we define $H(\mu) \coloneqq \sum_{i=1}^{n} g^{-1}\left(f^{-1}\left(\frac{d_i}{\mu}\right)\right)$. Note that $H(\mu)$ is a continuous function in $\mu > 0$. Moreover, it is easy to verify that $H(0) = \sum_{i=1}^{n} g^{-1}(0) = 0 < 1$. Assuming $h(1) \neq 0$, we can set $\mu = \frac{d_1}{f(1)}$ and get $H(\mu) > g^{-1}\left(f^{-1}(f(1))\right) = g^{-1}(1) = 1$. In the alternative case where $h(1) = 0$, we would still get $\lim_{\mu\to\infty} H(\mu) > 1$ by the same logic; meaning that there exists a value of $\mu > 0$ such that $H(\mu) > 1$ as well. In either case, the mean-value-theorem implies that there exists a $\mu \in \left[0, \frac{d_1}{f(1)}\right)$ such that Equation (65) is satisfied.

We verify that all the KKT conditions are satisfied by this candidate solution:

- *Primal feasibility:* Equation (64a) is satisfied because the domain of $f(\cdot)$, and therefore the range of $f^{-1}(\cdot)$, is $(0, 1]$. Equation (64b) is satisfied because Equation (65) implies that $\sum_{i=1}^{n} g^{-1}(q_i) = 1$.

- *Dual feasibility:* Equation (64c) is satisfied by setting $\boldsymbol{\lambda} = \boldsymbol{0}$, and Equation (64d) is satisfied by the choice of $\mu \in \left[0, \frac{d_1}{f(1)}\right)$.

- *Complementary slackness:* Equations (64e) and (64f) are satisfied because of the choices of $\boldsymbol{\lambda} = \mathbf{0}$ and Equation (65) respectively being satisfied.

- *Stationary condition:* We require $\operatorname{diag}(\boldsymbol{y})\, \boldsymbol{X}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\, \bar{\boldsymbol{q}} = \mu h(\bar{\boldsymbol{q}})$. This is equivalent to

$$\begin{aligned} d_i \bar{q}_i &= \mu h(\bar{q}_i) \\ \iff \mu &= \frac{d_i}{f(\bar{q}_i)} \text{ for all } i \in [n]. \end{aligned}$$

  Substituting our choice of $\bar{q}_i$ into the RHS above gives

$$\frac{d_i}{f(\bar{q}_i)} = \frac{d_i}{\frac{d_i}{\mu}} = \mu.$$

Thus, we have verified all the KKT conditions for this candidate solution. Ultimately, we get $\bar{\boldsymbol{w}} := \lim_{t\to\infty} \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2} = \frac{\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}}{\left\|\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}\right\|_2} = \frac{\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{D}\operatorname{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}}{\left\|\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{D}\operatorname{diag}(\boldsymbol{y})\bar{\boldsymbol{q}}\right\|_2}$. Therefore, the primal solution interpolates the adjusted binary levels given by

$$\tilde{y}_i = d_i y_i \bar{q}_i = d_i y_i f^{-1}\left(\frac{d_i}{\mu}\right).$$

This completes the proof. ■

D.2 Proof of Corollary 14

In this section, we prove Corollary 14.

Proof (of Corollary 14) It is easy to verify that minimizing the importance-weighted empirical risk with a polynomial loss function of degree $m > 0$ becomes equivalent to minimizing the unweighted empirical risk on the following per-example loss function

$$\ell_i(\tilde{z}_i; Q) := \frac{1}{(Q^{-\frac{\mathbb{I}[i\in S]}{m}} - \tilde{z}_i)^m}, \tag{66}$$

where $\tilde{z}_i := -y_i \langle \tilde{\boldsymbol{x}}_i, \boldsymbol{w}\rangle$ and $\tilde{\boldsymbol{x}}_i := Q^{-\frac{1}{m}\mathbb{I}[i\in S]}\boldsymbol{x}_i$. Clearly, the per-example loss function in Equation (66) continues to verify Assumption 1 for any fixed value of $Q > 0$. Specifically, it continues to satisfy $\ell_i'(\ell_i^{-1}(z)) = mz^{\frac{m+1}{m}}$ for $z \geq 0$ and so we get $g_i(d) = g(d) = d^{\frac{m+1}{m}}$ for each $i \in [n]$. Consequently, the convex program underlying the dual implicit bias is identical to (5) and the setting of Proposition 13, with adjusted diagonal matrix $\boldsymbol{D} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$, where we denote $\tilde{\boldsymbol{X}} := \operatorname{diag}\left(\boldsymbol{Q}^{-\frac{1}{m}\mathbb{I}[i\in S]}\right)\boldsymbol{X}$. To apply Proposition 13, we first calculate the functions $g(\cdot), g^{-1}(\cdot), h(\cdot), f(\cdot)$ and $f^{-1}(\cdot)$. Direct calculations yield $g(z) = z^{\frac{m+1}{m}}, g^{-1}(z) = z^{\frac{m}{m+1}}, h(z) = \frac{m}{m+1} \cdot z^{-\frac{1}{m+1}}, f(z) = \frac{m}{m+1} \cdot z^{-\frac{m+2}{m+1}}$, and $f^{-1}(z) = \left(\frac{(m+1)z}{m}\right)^{-\frac{m+1}{m+2}}$. Applying

Proposition 13 then gives us $\bar{q}_i = f^{-1}\left(\frac{d_i}{\mu}\right) \propto d_i^{-\frac{m+1}{m+2}}$. Next, we have

$$\bar{\boldsymbol{w}} := \lim_{t\to\infty} \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2} = \frac{\tilde{\boldsymbol{X}}^\top \mathrm{diag}\left(\boldsymbol{y}\right)\bar{\boldsymbol{q}}}{\left\|\tilde{\boldsymbol{X}}^\top \mathrm{diag}\left(\boldsymbol{y}\right)\bar{\boldsymbol{q}}\right\|_2} = \frac{\tilde{\boldsymbol{X}}^\top (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\boldsymbol{D}\mathrm{diag}\left(\boldsymbol{y}\right)\bar{\boldsymbol{q}}}{\left\|\tilde{\boldsymbol{X}}^\top (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\boldsymbol{D}\mathrm{diag}\left(\boldsymbol{y}\right)\bar{\boldsymbol{q}}\right\|_2}$$
$$= \frac{\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\mathrm{diag}\left(\boldsymbol{Q}^{\frac{1}{m}\mathbb{I}[i\in S]}\right)\boldsymbol{D}\mathrm{diag}\left(\boldsymbol{y}\right)\bar{\boldsymbol{q}}}{\left\|\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\mathrm{diag}\left(\boldsymbol{Q}^{\frac{1}{m}\mathbb{I}[i\in S]}\right)\boldsymbol{D}\mathrm{diag}\left(\boldsymbol{y}\right)\bar{\boldsymbol{q}}\right\|_2}.$$

Therefore, the adjusted labels that are interpolated are proportional to $Q^{\frac{1}{m}\mathbb{I}[i\in S]} y_i d_i \cdot d_i^{-\frac{m+1}{m+2}} = Q^{\frac{1}{m}\mathbb{I}[i\in S]} d_i^{\frac{1}{m+2}}$. It remains to calculate the value of d_i. Note that we have assumed $\boldsymbol{X}\boldsymbol{X}^\top = \alpha\boldsymbol{I}$, and so $\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle = \delta_{ij}$ where δ_{ij} denotes the Kronecker delta function. Because we have defined $\tilde{\boldsymbol{x}}_i := Q^{-\frac{1}{m}\mathbb{I}[i\in S]}\boldsymbol{x}_i$, we automatically get $\langle \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j\rangle = Q^{-\frac{1}{m}\cdot\left(\mathbb{I}[i\in S]+\mathbb{I}[j\in S]\right)}\delta_{ij}$, meaning that $d_i = Q^{-\frac{2}{m}\mathbb{I}[i\in S]}$. Putting all of this together results in interpolation of the per-example-adjusted labels $\tilde{y}_i \propto Q^{\frac{1}{m+2}\mathbb{I}[i\in S]} y_i$. This completes the proof. ∎

D.3 Proof of Proposition 17

In this section, we prove Proposition 17.

Proof (of Proposition 17) Our starting point lies in the proof of Theorem 6; the necessity for the dual implicit bias $\bar{\boldsymbol{q}}$ to satisfy the following characteristic equations, restated below.

$$\mathrm{diag}\left(\boldsymbol{y}\right)\boldsymbol{X}\boldsymbol{X}^\top \mathrm{diag}\left(\boldsymbol{y}\right)\bar{\boldsymbol{q}} = \mu h\left(\bar{\boldsymbol{q}}\right), \tag{67a}$$
$$\mu > 0, \qquad \text{and} \tag{67b}$$
$$\sum_{i=1}^{n} g^{-1}\left(\bar{q}_i\right) = 1. \tag{67c}$$

We consider in particular Equations (67a) and (67b). Recalling that we defined $\bar{\boldsymbol{q}}_{\boldsymbol{y}} := \mathrm{diag}\left(\boldsymbol{y}\right)\bar{\boldsymbol{q}}$ as shorthand, our equivalent goal is to lower-bound $\left\|\frac{\bar{\boldsymbol{q}}_{\boldsymbol{y}}}{\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\|_2} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2}\right\|_2$. Moreover, pre-multiplying both sides by $\mathrm{diag}\left(\boldsymbol{y}\right)$ means that the first and second characteristic equations imply

$$\boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{q}}_{\boldsymbol{y}} = \mu \mathrm{diag}\left(\boldsymbol{y}\right) h\left(\bar{\boldsymbol{q}}\right) \text{ for some } \mu > 0. \tag{68}$$

Next, we show that without loss of generality we can set $\mu = 1$ or any positive value, which greatly simplifies the proof exposition. The reason for this is as follows: consider a solution $\bar{\boldsymbol{q}}_{\boldsymbol{y}}$ that satisfies Equation (68) for some $\mu \neq 1$. Then, since $h\left(q\right)$ is a homogeneous function where $h\left(ab\right) = a^\gamma h\left(b\right)$ for $a, b \geq 0$ and $\gamma \in \mathbb{R}$, it is easy to verify that the modified solution $\mu^{\frac{1}{\gamma-1}}\bar{\boldsymbol{q}}_{\boldsymbol{y}}$ will satisfy Equation (68) for $\mu = 1$. Moreover, because the new solution is a scalar multiple of $\bar{\boldsymbol{q}}_{\boldsymbol{y}}$, it is identical in direction. Hence, for simplicity, we choose to solve the characteristic equations with a $\bar{\mu}$ where $\bar{\mu}h(1) = 1$, and we also define $\bar{h}(z) \coloneqq \bar{\mu}h(z)$, where $\bar{h}(z)$ is still a homogeneous function. Equation (68) becomes

$$\boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{q}}_{\boldsymbol{y}} = \operatorname{diag}(\boldsymbol{y})\, \bar{h}(\bar{\boldsymbol{q}}). \tag{69}$$

Then we denote that $\bar{\boldsymbol{q}} = \beta(\boldsymbol{1} + \boldsymbol{\Delta})$ for some $\beta > 0$ and some vector $\boldsymbol{\Delta}$ such that $\|1 + \boldsymbol{\Delta}\|_2 = \sqrt{n}$, and this ensures that $\left\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2 = \beta\sqrt{n}$. Note that

$$\frac{1}{\sqrt{n}}\|\boldsymbol{\Delta}\|_2 = \left\|\frac{\bar{\boldsymbol{q}}}{\sqrt{n}\beta} - \frac{\boldsymbol{1}}{\sqrt{n}}\right\|_2 = \left\|\frac{\bar{\boldsymbol{q}}_{\boldsymbol{y}}}{\left\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2}\right\|_2,$$

and so to obtain our desired lower bound on the set of candidate solutions $\bar{\boldsymbol{q}}_{\boldsymbol{y}}$ satisfying $\left\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2 = \beta\sqrt{n}$, it suffices to obtain a lower bound on $\|\boldsymbol{\Delta}\|_2$. By considering Equation (69), we write as shorthand $\boldsymbol{\Delta}_{\boldsymbol{y}} \coloneqq \operatorname{diag}(\boldsymbol{y}) \cdot \boldsymbol{\Delta}$; therefore, we have

$$\begin{aligned} &\beta \cdot \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y} + \beta \cdot \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Delta}_{\boldsymbol{y}} = \beta^\gamma \cdot \operatorname{diag}(\boldsymbol{y})\, \bar{h}(\boldsymbol{1} + \boldsymbol{\Delta}) \\ \iff &\beta \cdot \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Delta}_{\boldsymbol{y}} - \beta^\gamma \cdot \operatorname{diag}(\boldsymbol{y})\, \sigma(\boldsymbol{\Delta}) = \left(\beta^\gamma \boldsymbol{I} - \beta \boldsymbol{X}\boldsymbol{X}^\top\right)\boldsymbol{y}, \end{aligned} \tag{70}$$

where we define $\sigma(\Delta_i) \coloneqq \bar{h}(1 + \Delta_i) - 1$ for any $\Delta_i > -1$. This function is well-defined for our choice of $\boldsymbol{\Delta}$, because the constraint $\bar{\boldsymbol{q}} \succ 0$ necessitates $\boldsymbol{\Delta} \succ -\boldsymbol{1}$.

We now upper bound the norm of the LHS of Equation (70) above. Note that we assume $\|\boldsymbol{\Delta}\|_\infty \le \delta$ for some $\delta \in (0, 1)$, and $\bar{h}(z)$ is a decreasing function with $\bar{h}(1) = 1$. Hence, since we have $\sigma(0) = 0$, it is straightforward to upper bound $|\sigma(\Delta_i)|$ using an absolute linear function $|-k\Delta_i|$ with $k \ge 0$. If $\bar{h}(z)$ is a convex function, we can determine k using $\Delta_i = -\delta$; otherwise, if $\bar{h}(z)$ is a concave function, we can determine k using $\Delta_i = \delta$. Therefore, we have $|\sigma(\Delta_i)| \le \begin{cases} \frac{\bar{h}(1-\delta)-1}{\delta}|\Delta_i| & \bar{h}''(z) \ge 0 \\ \frac{1-\bar{h}(1+\delta)}{\delta}|\Delta_i| & \bar{h}''(z) < 0 \end{cases}$. As a result, we can choose $k = \max\left(\frac{\bar{h}(1-\delta)-1}{\delta}, \frac{1-\bar{h}(1+\delta)}{\delta}\right)$ such that $|\sigma(\Delta_i)| \le k|\Delta_i|$ and $\left\|\sigma(\boldsymbol{\Delta})\right\|_2 \le k\|\boldsymbol{\Delta}\|_2$. This leads to the upper bound

$$\begin{aligned} \left\|\beta \cdot \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Delta}_{\boldsymbol{y}} - \beta^\gamma \cdot \operatorname{diag}(\boldsymbol{y})\, \sigma(\boldsymbol{\Delta})\right\|_2 &\le \beta\left\|\boldsymbol{X}\boldsymbol{X}^\top\right\|_2 \left\|\boldsymbol{\Delta}_{\boldsymbol{y}}\right\|_2 + \beta^\gamma \left\|\sigma(\boldsymbol{\Delta})\right\|_2 \\ &\le \left(\beta\left\|\boldsymbol{X}\boldsymbol{X}^\top\right\|_2 + k\beta^\gamma\right)\|\boldsymbol{\Delta}\|_2 \\ &\le 2\max(\beta\left\|\boldsymbol{X}\boldsymbol{X}^\top\right\|_2, k\beta^\gamma)\,\|\boldsymbol{\Delta}\|_2. \end{aligned}$$

Plugging this upper bound into Equation (70) above and dividing numerator and denominator by $\beta > 0$ yields

$$\|\boldsymbol{\Delta}\|_2 \ge \min\left\{\frac{\left\|(\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I})\boldsymbol{y}\right\|_2}{2k\alpha}, \frac{\left\|(\boldsymbol{X}\boldsymbol{X}^\top - \alpha\boldsymbol{I})\boldsymbol{y}\right\|_2}{2\left\|\boldsymbol{X}\boldsymbol{X}^\top\right\|_2}\right\},$$

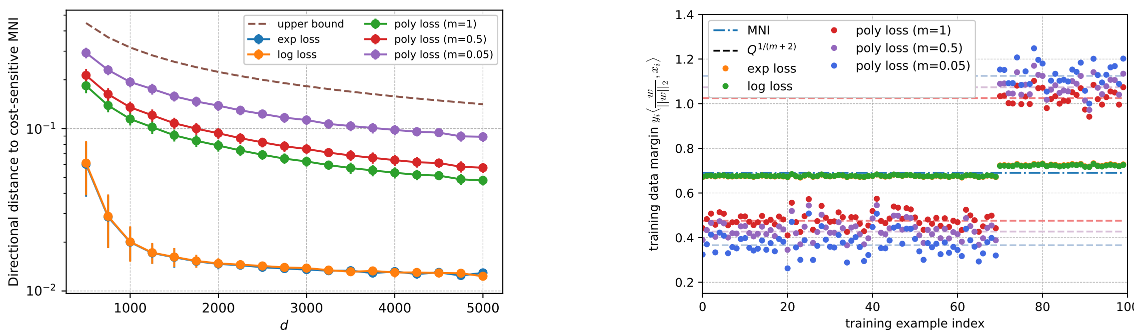
where we defined $\alpha \coloneqq \beta^{\gamma-1}$ as shorthand. Consequently, we have

$$\left\| \frac{\bar{\boldsymbol{q}}_{\boldsymbol{y}}}{\left\|\bar{\boldsymbol{q}}_{\boldsymbol{y}}\right\|_2} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} \right\|_2 = \frac{1}{\sqrt{n}} \|\boldsymbol{\Delta}\|_2 \geq \frac{1}{2\sqrt{n}} \min \left\{ \frac{\left\|(\boldsymbol{X}\boldsymbol{X}^\top - \alpha \boldsymbol{I})\boldsymbol{y}\right\|_2}{k\alpha}, \frac{\left\|(\boldsymbol{X}\boldsymbol{X}^\top - \alpha \boldsymbol{I})\boldsymbol{y}\right\|_2}{\left\|\boldsymbol{X}\boldsymbol{X}^\top\right\|_2} \right\},$$

Further minimizing over all $\alpha > 0$ then yields the desired result. ■

Appendix E. Additional Simulations for Importance Weighting under Random Data

In this section, we provide additional simulations on random data in order to evaluate how different loss functions influence the training data margins under importance weighting. These simulations are a more realistic complement to Figure 4b, which considered the idealized scenario where $\boldsymbol{X}\boldsymbol{X}^\top = \alpha \boldsymbol{I}$.



(a) Directional convergence of various loss functions with importance weighting.

(b) Importance weighting on different loss functions for random data.

Figure 5: Panel (a) compares the implicit bias of gradient descent to the *cost-sensitive MNI* (defined in Corollary 14), which is obtained by fitting the adjusted label $\operatorname{diag}\left(Q^{\frac{1}{m+2}\cdot \mathbb{I}[i\in S]}\right)\boldsymbol{y}$ with $Q = 2.0$ on $y_i = -1$. The results demonstrate that the directional distance to the cost-sensitive MNI follows a similar upper bound as in Theorem 6. The simulation setup is the same as Figure 3b. Panel (b) present the outcomes of importance weighting on different loss functions under random data for data dimensions $d = 5000$. The covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ are independently and identically distributed (IID) isotropic Gaussian with a fixed sample size $n = 100$. The first 70 examples are *majority examples* and labeled as $y_i = +1$, and the rest of the 30 examples are *minority examples* labeled as $y_i = -1$. Note that we apply the importance weighting factor $Q = 10.0$ only to the minority examples. We run gradient descent on all loss functions for the minimum of 10^4 iterations, or when the empirical risk falls below 10^{-12}. We also observe that heavier-tailed polynomial losses (i.e. smaller values of m) lead to a stronger importance weighting effect.

References

Navid Ardeshir, Clayton Sanford, and Daniel J Hsu. Support vector machines and linear regression coincide with very high-dimensional features. *Advances in Neural Information Processing Systems*, 34:4907–4918, 2021.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26 (5):1651–1686, 1998.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Tina Behnia, Ke Wang, and Christos Thrampoulidis. On how to avoid exacerbating spurious correlations when models are overparameterized. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 121–126. IEEE, 2022.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *The Journal of Machine Learning Research*, 22 (1):5721–5750, 2021.

Niladri S Chatterji, Philip M Long, and Peter L Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks. *The Journal of Machine Learning Research*, 23(263):1–48, 2022.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Miroslav Dudík, Ziwei Ji, Robert E Schapire, and Matus Telgarsky. Convex analysis at infinity: An introduction to astral space. *arXiv preprint arXiv:2205.03260*, 2022.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

Pavel Golik, Patrick Doetsch, and Hermann Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *Interspeech*, volume 13, pages 1756–1760, 2013.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information processing Systems*, 31, 2018b.

Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. Cambridge university press, 1952.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2): 949–986, 2022.

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99. PMLR, 2021.

Hanwen Huang. Asymptotic behavior of support vector machine for spiked population model. *The Journal of Machine Learning Research*, 18(1):1472–1492, 2017.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017.

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

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

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

Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

William Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34:18970–18983, 2021.

Douglas M Kline and Victor L Berardi. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing & Applications*, 14:310–318, 2005.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *The Journal of Machine Learning Research*, 21(1):6863–6878, 2020.

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34: 10144–10157, 2021.

Gábor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. *The Annals of statistics*, 32(1):30–55, 2004.

Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. IEEE, 2019.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. *Advances in Neural Information Processing Systems*, 21, 2008.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *The Journal of Machine Learning Research*, 22(1):10104–10172, 2021.

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

Arkadi Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

Hrithik Ravi, Clay Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. *Advances in Neural Information Processing Systems*, 37:81324–81359, 2024.

R Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. Hebrew University, 2007.

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

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE transactions on information theory*, 51(1):128–142, 2005.

Vignesh Subramanian, Rahul Arya, and Anant Sahai. Generalization for multiclass classification with overparameterized linear models. *arXiv preprint arXiv:2206.01399*, 2022.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116 (29):14516–14525, 2019.

Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *The Journal of Machine Learning Research*, 8(5), 2007.

Alexandre B Tsybakov. Nonparametric estimators. *Introduction to Nonparametric Estimation*, pages 1–76, 2009.

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, 4(1):260–284, 2022.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34:24164–24179, 2021a.

Ke Alexander Wang, Niladri S Chatterji, Saminul Haque, and Tatsunori Hashimoto. Is importance weighting incompatible with interpolating classifiers? *arXiv preprint arXiv:2112.12986*, 2021b.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.