Don't Just Blame Over-parametrization for Over-confidence: Theoretical Analysis of Calibration in Binary Classification

Yu Bai [1] **Song Mei** [2] **Huan Wang** [1] **Caiming Xiong** [1]

Abstract

Modern machine learning models with high accuracy are often miscalibrated—the predicted top probability does not reflect the actual accuracy, and tends to be *over-confident*. It is commonly believed that such over-confidence is mainly due to *over-parametrization*, in particular when the model is large enough to memorize the training data and maximize the confidence.

In this paper, we show theoretically that over-parametrization is not the only reason for over-confidence. We prove that *logistic regression is inherently over-confident*, in the realizable, under-parametrized setting where the data is generated from the logistic model, and the sample size is much larger than the number of parameters. Further, this over-confidence happens for general well-specified binary classification problems as long as the activation is symmetric and concave on the positive part. Perhaps surprisingly, we also show that over-confidence is not always the case—there exists another activation function (and a suitable loss function) under which the learned classifier is *under-confident* at some probability values. Overall, our theory provides a precise characterization of calibration in realizable binary classification, which we verify on simulations and real data experiments.

1. Introduction

Modern machine learning models such as deep neural networks with high accuracy tend to be miscalibrated: The predicted top probability (*confidence*) does not reflect the actual accuracy of the model, and tends to be *over-confident*. For example, a WideResNet 32 on CIFAR100 has on average a predicted top probability of 87%, while the actual test accuracy is only 72% (Guo et al., 2017). As the confidence is often comprehended as an estimate of the true accuracy, such over-confidence could be dangerous, especially in risk-sensitive domains such as medical AI (Begoli et al., 2019), self-driving cars (Michelmore et al., 2018), and so on. To address this issue, there is a growing line of research on improving the calibration of models, by either performing *recalibration* of well-trained models to adjust the confidence scores (Platt et al., 1999; Zadrozny & Elkan; Naeini et al., 2015; Guo et al., 2017), or by averaging the predictions over multiple models to make the confidence scores more accurate (Lakshminarayanan et al., 2016; Gal & Ghahramani, 2016). These methods in general can reduce the over-confidence and improve the calibration of the model, while preserving (or even improving) the model's accuracy (Ovadia et al., 2019).

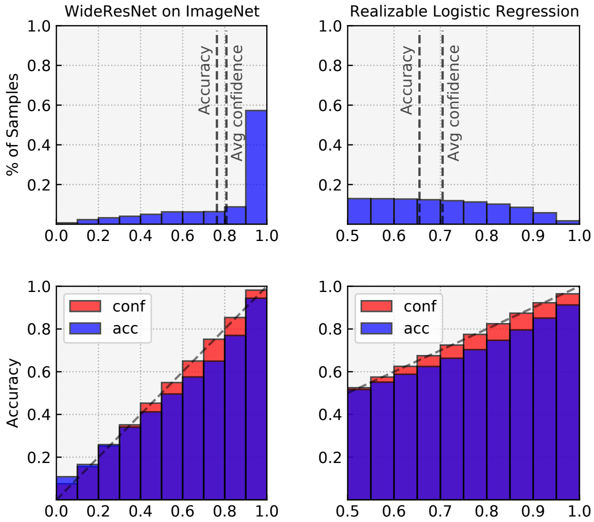


Figure 1. Reliability diagrams for calibration: Over-parametrized deep network vs. well-specified, under-parametrized logistic regression. The x-axes denote the confidences (predicted top probabilities) of the models. Left: 50-layer WideResNet on ImageNet. Right: Binary logistic regression on simulated data with $n = 2000$ and $d = 100$.

Despite these progresses, the more fundamental question of *why* such over-confidence happens for vanillaly trained

[1] Salesforce Research [2] University of California, Berkeley. Correspondence to: Yu Bai <yu.bai@salesforce.com>, Song Mei <songmei@berkeley.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

models remains not satisfactorily understood. One common understanding is that over-confidence is a result of *over-parametrization*: Models such as deep neural networks are large enough to memorize the entire training dataset, and are encouraged to magnify its weights and maximize the confidence so as to minimize the training loss (Mukhoti et al., 2020). Guo et al. (2017) also observed that increasing the depth and width makes the over-confident more severe, even when this improves the accuracy. However, so far it is unclear whether over-parametrization is the only reason, or whether there are other intrinsic reasons leading to over-confidence.

In this paper, we show that over-confidence is not just a result of over-parametrization and is more inherent. We conduct a precise theoretical study on the calibration in binary classification problems. Our main result shows that *standard logistic regression is also over-confident*, even in the well-specified, under-parametrized scenario where the model is correct (data generated from a linear logistic model), and there is abundant data (number of samples n much greater than number of parameters d).

Figure 1 illustrates our main finding via simulation: Similar to an over-parametrized neural network, the empirical risk minimizer of logistic regression is also over-confident at all confidence levels. Note that these two models have rather different behaviors in terms of the distribution of confidences, yet their over-confidence behaviors are similar.

Our contributions are summarized as follows:

- We show that *well-specified logistic regression is inherently over-confident*: Conditioned on the model predicting $p > 0.5$, the actual probability of the label being one is lower by an amount of $\Theta(d/n)$, in the limit of $n, d \to \infty$ proportionally and n/d is large (Section 3). In other words, the calibration error is always in the over-confident direction. We also show that the overall Calibration Error (CE) of the logistic model is $\Theta(d/n)$ in this limiting regime.

- We identify sufficient conditions for over- and under-confidence in general binary classification problems, where the data is generated from an arbitrary nonlinear activation, and we solve a well-specified empirical risk minimization (ERM) problem with a suitable loss function (Section 4). Our conditions imply that any symmetric, monotone activation $\sigma : \mathbb{R} \to [0, 1]$ that is *concave* at all $z > 0$ will yield a classifier that is over-confident at any confidence level.

- Another perhaps surprising implication is that *over-confidence is not universal*: We prove that there exists an activation function for which under-confidence can happen for a certain range of confidence levels.

- We perform simulation and real data experiments to test our theory (Section 5). Our experiments suggest that the over-confidence of logistic regression happens broadly in a variety of under-parametrized settings, within or beyond our theory's assumptions. We also verify that under-confidence can indeed happen in simulations with the activation function constructed above.

- On the technical end, our analysis develops a precise understanding of the high-dimensional proportional limit of ERM in the *sufficient data* regime (n/d is large) by rigorously establishing the first-order behavior of the solution to the characterizing system of non-linear equations (Section 6), which may be of broader interest.

1.1. Related work

Algorithms for model calibration Practitioners have observed and dealt with the over-confidence of logistic regression long ago. *Recalibration algorithms* fix this by adjusting the output of a well-trained model, and dates back to the classical methods of Platt scaling (Platt et al., 1999), histogram binning (Zadrozny & Elkan) and isotonic regression (Zadrozny & Elkan, 2002). Platt et al. (1999) also uses a particular kind of label smoothing as a way of mitigating the over-confidence in logistic regression. Guo et al. (2017) show that temperature scaling, a simple method that learns a rescaling factor for the logits, is a competitive method for calibrating neural networks. A number of recent recalibration methods further improve the performances over these approaches (Kull et al., 2017; 2019; Ding et al., 2020; Rahimi et al., 2020; Zhang et al., 2020).

Another line of work improves calibration by aggregating the probabilisitic predictions over multiple models, using either an ensemble of models (Lakshminarayanan et al., 2016; Malinin et al., 2019; Wen et al., 2020; Tran et al., 2020), or randomized predictions such as Bayesian neural networks (Gal & Ghahramani, 2016; Gal et al., 2017; Maddox et al., 2019; Dusenberry et al., 2020). Finally, there are techniques for improving the calibration of a single neural network during training (Thulasidasan et al., 2019; Mukhoti et al., 2020; Liu et al., 2020).

Theoretical analysis of calibration Kumar et al. (2019) show that continuous rescaling methods such as temperature scaling is less calibrated than reported, and proposed a method that combines temperature scaling and histogram binning. Gupta et al. (2020) study the relationship between calibration and other notions of uncertainty such as confidence intervals. Shabat et al. (2020); Jung et al. (2020) study the sample complexity of estimating the multicalibration error (group calibration). A related theoretical result to ours is (Liu et al., 2019) which shows that the calibration error of any classifier is upper bounded by its square root excess logistic loss over the Bayes classifier. This result can be translated to a $O(\sqrt{d/n})$ upper bound for well-specified logistic regression, whereas our main result implies $\Theta(d/n)$ calibration error in our high-dimensional limiting regime (with input distribution assumptions).

High-dimensional behaviors of empirical risk minimization There is a rapidly growing literature on limiting characterizations of convex optimization-based estimators in the $n \propto d$ regime (Donoho et al., 2009; Bayati & Montanari, 2011; El Karoui et al., 2013; Karoui, 2013; Stojnic, 2013; Thrampoulidis et al., 2015; 2018; Mai et al., 2019; Sur & Candès, 2019; Candès et al., 2020). Our analysis builds on the characterization for unregularized convex risk minimization problems (including logistic regression) derived in Sur & Candès (2019).

2. Preliminaries

In this paper we consider binary classification problems, where we observe n data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} P$ for some distribution P on $\mathbb{R}^d \times \{0,1\}$.

2.1. Calibration

Let $\widehat{f} : \mathbb{R}^d \to [0,1]$ be a (probabilistic) classifier. $\widehat{f}$ is said to be perfectly calibrated if $\mathbb{P}(Y = 1 | \widehat{f}(\mathbf{X}) = p) = p$ for all $p \in [0,1]$, that is, the actual probability of $Y = 1$ conditioned on $\widehat{f}$ predicting p is exactly p. In reality, we cannot hope for obtaining perfect calibration, and would rather desire ways of measuring the calibration error.

A standard metric is the Calibration Error (CE), which measures the difference between the prediction and the conditional mean of Y given the prediction (Guo et al., 2017):

$$\text{CE}(\widehat{f}) := \mathbb{E}_{(\mathbf{X},Y)\sim P}\left[\left|\widehat{f}(\mathbf{X}) - \mathbb{E}[Y \mid \widehat{f}(\mathbf{X})]\right|\right]. \tag{1}$$

Notably, CE is the population (unbinned) version of the Expected Calibration Error (ECE), a commonly used calibration metric in recent work (Naeini et al., 2015; Guo et al., 2017; Ovadia et al., 2019; Nixon et al., 2019).

In this paper, we consider the *calibration error of $\widehat{f}$ at level p*:

$$\Delta_p^{\mathsf{cal}}(\widehat{f}) := p - \mathbb{P}_{(X,Y)\sim P}\Big(Y = 1 \mid \widehat{f}(\mathbf{X}) = p\Big) \tag{2}$$

for all $p \in (0,1)$. Note that $\Delta_p^{\mathsf{cal}}(\widehat{f})$ is the quantity inside the expectation in (1), and provides a more fine-grained characterization of the calibration error by specifying which p we are interested in.

Over-confidence and under-confidence The *confidence* of $\widehat{f}$ at $\mathbf{x}$ is the predicted top probability, i.e. $\max\{\widehat{f}(\mathbf{x}), 1 - \widehat{f}(\mathbf{x})\}$ for binary problems. In particular, when $\widehat{f}(\mathbf{x}) > 0.5$, the confidence is equal to $\widehat{f}(\mathbf{x})$. We say that the model is *over-confident* when the confidence is higher than the actual accuracy: For example, when the model predicts $\widehat{f}(\mathbf{x}) = 0.9$, but we have $\mathbb{E}[Y | \widehat{f}(\mathbf{x}) = 0.9] = 0.8$, then $\widehat{f}$ is over-confident at level $p = 0.9$. Note that in this case the calibration error at level 0.9 is positive: $\Delta_{0.9}^{\mathsf{cal}}(\widehat{f}) = 0.1 > 0$. In other words, over- or under-confidence is determined by the **sign** of the calibration error $\Delta_p^{\mathsf{cal}}(\widehat{f})$ in definition (2):

For any $p \in (0.5, 1)$:

- $\Delta_p^{\mathsf{cal}}(\widehat{f}) > 0$: $\widehat{f}$ is **over-confident** at level p;
- $\Delta_p^{\mathsf{cal}}(\widehat{f}) < 0$: $\widehat{f}$ is **under-confident** at level p.

We remark that we only state results for $p > 0.5$ in this paper; all the results also hold for $p \in (0, 0.5)$ by symmetry.

Extension to multi-class problems In our experiments we also consider multi-class classification problems, for which there is a standard generalization of definitions (2) and (1) (Guo et al., 2017): Given a multi-class predictor $\widehat{F} : \mathbb{R}^d \to \Delta_K$ where $K \geq 2$ is the number of classes, we replace Y with the indicator of correct prediction: $\mathbf{1}\left\{Y = \arg\max_k \widehat{F}(x)_k\right\}$, and replace $\widehat{f}(x)$ with the confidence $\max_k \widehat{F}(x)_k$. Thus the calibration error of $\widehat{F}$ at level $p \in [1/K, 1]$ is $\Delta_p^{\mathsf{cal}}(\widehat{F}) := p - \mathbb{P}\Big(Y = \arg\max_k \widehat{F}(\mathbf{X})_k \mid \max_k \widehat{F}(\mathbf{X})_k = p\Big)$.

2.2. Model and data distribution

We consider the following data distribution where $\mathbf{X}$ is standard Gaussian and $Y|\mathbf{X}$ follows a binary linear model with activation function $\sigma : \mathbb{R} \to [0,1]$:

$$P: \quad \mathbf{X} \sim \mathsf{N}(0, \mathbf{I}_d), \quad \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \sigma(\mathbf{w}_\star^\top \mathbf{x}), \tag{3}$$

where $\mathbf{w}_\star \in \mathbb{R}^d$ is the ground truth coefficient vector. (This is also known as generalized linear models with link function σ (McCullagh, 2018)). We make the Gaussian input assumption as our analysis requires a precise limiting calculation; however, our real data experiments in Section 5.2 suggest that the implications of our theory may hold more broadly without such distributional assumptions.

Realizable logistic regression Our primary focus is *realizable logistic regression*, in which $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic (sigmoid) activation, and we solve the unregularized

ERM (empirical risk minimization) problem

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \widehat{R}_n(\mathbf{w}) \\ := \frac{1}{n}\sum_{i=1}^{n}\left[\log(1+\exp(\mathbf{w}^\top \mathbf{x}_i)) - y_i \mathbf{w}^\top \mathbf{x}_i\right]. \tag{4}$$

Let $R(\mathbf{w}) := \mathbb{E}[\widehat{R}_n(\mathbf{w})]$ denote the expected (population) risk. It is a classical result that $\arg\min_{\mathbf{w}} R(\mathbf{w}) = \mathbf{w}_\star$, i.e. logistic regression is well-specified when data comes from the logistic model (Hastie et al., 2009).

Extension to general activations We also consider generalizations where σ is a general monotone activation function, and we wish to learn a linear classifier $\widehat{\mathbf{w}}$ that is close to $\mathbf{w}_\star$. In this case, we consider solving the general ERM

$$\text{minimize } \widehat{R}_n(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{w}^\top \mathbf{x}_i) - y_i \mathbf{w}^\top \mathbf{x}_i, \tag{5}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is a loss function. Let $R(\mathbf{w}) := \mathbb{E}[\widehat{R}_n(\mathbf{w})]$ denote the expected (population) risk.

To make sure the problem is well-specified, we choose ρ to be the (integrated) convex loss associated with σ: $\rho(z) = \int_0^z \sigma(u)du + C$ for some constant C; in other words $\rho'(z) = \sigma(z)$. It is known that for such a choice of ρ we have $\arg\min_{\mathbf{w}} R(\mathbf{w}) = \mathbf{w}_\star$ (Kakade et al., 2011). (For completeness we also provide a proof in Appendix A.3.)

We require the following assumption on the activation function σ along with the loss function ρ, which only requires the activation to be smooth along with some basic properties, such as monotonicity and symmetry around 0.

Assumption A (Smooth activation). *The loss function $\rho : \mathbb{R} \to \mathbb{R}$ is strictly convex and four-times continuously differentiable with uniformly bounded $\{1, 2, 3, 4\}$-th derivatives. The activation function $\sigma = \rho'$ is strictly increasing, and satisfies $\sigma(0) = 1/2$, $\lim_{z\to-\infty}\sigma(z) = 0$, $\lim_{z\to\infty}\sigma(z) = 1$, and $\sigma'(z) = \sigma'(-z) > 0$ for all $z \in \mathbb{R}$.*

3. Logistic regression is over-confident

As a warm-up, consider running unregularized (linear) logistic regression in the over-parametrized setting where $n < d$ and the data is separable. In this case, it is known that the (ERM) solution to the logistic regression (4) does not exist (Albert & Anderson, 1984; Candès et al., 2020); the gradient descent path will also diverge to infinity norm (Soudry et al., 2018). Using an approximate solution $\widehat{\mathbf{w}}$ with a high norm will cause the learned classifier $\sigma(\widehat{\mathbf{w}}^\top \mathbf{x})$ to be nearly a step function (outputs are close to either 0 or 1). Such classifiers are clearly over-confident whenever the true conditional distribution $Y|\mathbf{X}$ is not approximately deterministic.

We are now ready to present our main result, which states that even in the most vanilla setting (well-specified, under-parametrized), logistic regression is still over-confident.

Theorem 1 (Well-specified logistic regression is over-confident). *Consider the classifier $\widehat{f}(\mathbf{x}) = \sigma(\widehat{\mathbf{w}}^\top \mathbf{x})$ obtained from logistic regression (4), where the data is generated from the logistic model (3). Then we have the following.*

- *In the limit of $n, d \to \infty$[1] and $d/n \to \kappa$, where $\kappa \in (0, \kappa_0]$ for some constant $\kappa_0 > 0$ (which only depends on $\|\mathbf{w}_\star\|$), for any $p \in (0.5, 1)$, almost surely, we have*

  $$\Delta_p^{\mathsf{cal}}(\widehat{f}) \to C_{p,\kappa} \quad \textit{for some } C_{p,\kappa} > 0.$$

  In words, logistic regression gives inherently over-confident estimates of the actual probabilities.

- *We have, for small enough $\kappa > 0$,*

  $$C_{p,\kappa} = C_p \cdot \kappa + o(\kappa).$$

  In words, as the sample size $n/d = 1/\kappa$ becomes large, the over-confidence effect becomes weaker. The scaling of this over-confidence effect is roughly $C_p \cdot d/n$.

Over-confidence is inherent for logistic regression Theorem 1 considers the under-parametrized setting, as we allow $d/n = \kappa$ to be any *small* value, thus the sample size n can be arbitrarily higher than the dimension d. It thus suggests that over-confidence of logistic regression a rather fundamental property, and challenges the common belief that over-confidence mostly comes from over-parametrization. Furthermore, even though $\Delta_p^{\mathsf{cal}}(\widehat{f})$ becomes smaller as the sample size increases (κ becomes lower), Theorem 1 still asserts the sign of $\Delta_p^{\mathsf{cal}}(\widehat{f})$ being always positive in the proportional limit of $n, d \to \infty$, $d/n \to \kappa$. This result perhaps unveils another source of over-confidence in real-world machine learning models beyond linear logistic models.

Furthermore, Theorem 1 shows that logistic regression is over-confident at all $p \in (0.5, 1)$. This suggests that the over-confidence in every confidence bin, as an empirical observation in well-trained neural networks (Guo et al., 2017), holds for logistic regression as well and is not unique to large over-parametrized models.

Regularization; comparison with classical asymptotics We remark that our result only holds for *unregularized* logistic regression, while it is known that various regularization can improve calibration (Gal & Ghahramani, 2016; Thulasidasan et al., 2019). Indeed, in our model, applying regularization (e.g. an L2 regularizer) will in general reduce the calibration error, as long as the regularization reduces

[1] We assume $\|\mathbf{w}_\star\|$ is the same for all (n, d).

the norm of $\widehat{\mathbf{w}}$ and does not hurt its correlation with $\mathbf{w}_\star$ too much. However, we intentionally focus on unregularized logistic regression which resembles practical setups such as neural networks in the memorizing regime. We also note that, in general, the best regularization strength for the optimal accuracy and the optimal calibration may be different.

We also briefly remark that our setting of $d, n \to \infty$, $d/n \to \infty$ is different from classical asymptotic statistics (which considers fixed d and $n \to \infty$) (Van der Vaart, 2000). Classical asymptotics would imply $\sqrt{n}\Delta_p^{\mathsf{cal}}(\widehat{f}) \xrightarrow{d} \mathsf{N}(0, V^2)$ for some V^2, and thus $\Delta_p^{\mathsf{cal}}(\widehat{f})$ has about equal chance to be positive or negative; in contrast, we show that $\Delta_p^{\mathsf{cal}}(\widehat{f})$ has a positive bias in the proportional limit, a regime arguably more realistic than classical asymptotics.

CE of logistic regression Theorem 1 further implies a result on the calibration error (CE) of logistic regression.

Corollary 2 (Asymptotics of calibration error)**.** *In the same setting as Theorem 1, as $d, n \to \infty$, $d/n \to \kappa$, the* CE *of the logistic regression solution $\widehat{f}$ satisfies*

$$\mathrm{CE}(\widehat{f}) \to C_\kappa,$$

almost surely, where for small enough κ we have $C_\kappa = C\kappa + o(\kappa)$ for some absolute constant $C > 0$.

Corollary 2 implies that, in the limiting regime, the CE of logistic regression is $O(\kappa) = O(d/n)$. This improves over the results of Liu et al. (2019) in certain aspects. First, Liu et al. (2019, Corollary 2.4) showed that the CE of any classifier is bounded by the square root excess logistic loss over the Bayes classifier. This implies the CE of well-specified logistic regression is bounded by $\sqrt{d/n}$. Here we show the CE has a better rate $\Theta(d/n)$ at small d/n in our limiting regime[2]. Second, our Theorem 1 determines the sign of the calibration error (confidence > accuracy), which is not implied by their results.

The proof of Corollary 2 follows directly from Theorem 1 by integrating $\Delta_p^{\mathsf{cal}}(\widehat{f})$ over $p \in (0, 1)$ (with p distributed as $\widehat{f}(\mathbf{x})$ for $\mathbf{x} \sim P$). The proof can be found in Appendix D.3.

3.1. Proof sketch of Theorem 1

We now provide a high-level overview of the proof of Theorem 1. A more detailed overview of the most technical steps is deferred to Section 6, and the full proofs can be found in Appendix C & D.

Closed-form expression for calibration error Recall that

$$\Delta_p^{\mathsf{cal}}(\widehat{f}) = p - \mathbb{E}_{\mathbf{x}}\left[\sigma(\mathbf{w}_\star^\top \mathbf{x}) \mid \sigma(\widehat{\mathbf{w}}^\top \mathbf{x}) = p\right].$$

As $\mathbf{x}$ is standard Gaussian, the conditional distribution of $\mathbf{x}|\sigma(\widehat{\mathbf{w}}^\top \mathbf{x}) = p$ can be characterized precisely in terms of the projection of $\mathbf{x}$ onto the direction $\widehat{\mathbf{w}}$ and its orthogonal complement subspace. Standard calculation then yields the closed form expression

$$\Delta_p^{\mathsf{cal}}(\widehat{f}) = p - \mathbb{E}_Z\left[\sigma\left(\frac{\|\mathbf{w}_\star\|}{\|\widehat{\mathbf{w}}\|}\cos\widehat{\theta} \cdot \sigma^{-1}(p) + \sin\widehat{\theta}\,\|\mathbf{w}_\star\|\,Z\right)\right], \tag{6}$$

where $\cos\widehat{\theta} = \frac{\widehat{\mathbf{w}}^\top \mathbf{w}_\star}{\|\widehat{\mathbf{w}}\|\|\mathbf{w}_\star\|}$ is the angle between $\widehat{\mathbf{w}}$ and $\mathbf{w}_\star$, and $Z \sim \mathsf{N}(0, 1)$. (See Lemma B.1 for the detailed statement and proof.)

Concentration of $\widehat{\mathbf{w}}$ In the second step, we apply results from recent advances in high-dimensional convex risk minimization (Sur & Candès, 2019; Taheri et al., 2020) to show that $\widehat{\mathbf{w}}$ concentrates around fixed values in the high-dimensional limit, in terms of its norm and cosine angle with $\mathbf{w}_\star$. These results show that, in the limit of $d, n \to \infty$ and $d/n \to \kappa$, the following concentration happens almost surely:

$$\begin{aligned} \|\widehat{\mathbf{w}}\| &\to R_\star = R_\star(\kappa, \|\mathbf{w}_\star\|), \\ \cos\widehat{\theta} &\to c_\star = c_\star(\kappa, \|\mathbf{w}_\star\|), \end{aligned} \tag{7}$$

Above, $R_\star$ and $c_\star$ are determined by the solutions of a system of nonlinear equations with three variables $(\alpha, \sigma, \lambda)$ (see Section 6 and Theorem C.1 for the formal statement).

The concentration directly implies that $\Delta_p^{\mathsf{cal}}(\widehat{f})$ converges to the following limiting calibration error (Corollary C.1):

$$\Delta_p^{\mathsf{cal}}(\widehat{f}) \to C_{p,\kappa} := p - \mathbb{E}_Z\left[\sigma\left(\frac{\|\mathbf{w}_\star\|}{R_\star}c_\star \cdot \sigma^{-1}(p) + \sqrt{1 - c_\star^2}\,\|\mathbf{w}_\star\|\,Z\right)\right]. \tag{8}$$

This expression hints on potential sources of over- or under-confidence: (1) $R_\star$ and $c_\star$ will affect the "multiplier" $\|\mathbf{w}_\star\|\,c_\star/R_\star$ in front of $\sigma^{-1}(p)$, drifting the expectation away from p; (2) $c_\star$ also affects the expectation over the $\sqrt{1 - c_\star^2}\,\|\mathbf{w}_\star\|\,Z$ term. This term itself has mean zero, but can affect the overall expectation through the nonlinear activation function σ.

Calculating the limiting calibration error The final part, as a technical crux of the proof, calculates the limiting calibration error (8) by precisely analyzing the interplay between the concentration values $R_\star$, $c_\star$ and the activation function σ. This is achieved by a novel analysis on the solutions of the aformentioned system of equations at small

[2] We remark that Corollary 2 does not readily imply a $\Theta(d/n)$ result in the non-asymptotic setting. However, we believe a similar result (with additional terms such as $1/\sqrt{n}$) holds and can be established via a more refined analysis.

κ. In particular, we show that $C_{p,\kappa} = C_p\kappa + o(\kappa)$ for small κ, and $C_p > 0$ is positive, thereby establishing Theorem 1. We present a more detailed description of this analysis in Section 6.

4. Over-confidence is not universal

It is natural to ask—based on Theorem 1—whether over-confidence is true in other well-specified problems as well, or is due to some specific property about logistic regression. This section makes steps towards this by looking at the generalized problem (5) where σ is an arbitrary activation and we solve the corresponding convex ERM.

Our main result in this section is the following characterization of sufficient conditions for whether over- or under-confidence happens in the general convex ERM (5). The proof of this result can be found in Appendix D.1.

Theorem 3 (Sufficient conditions for over- and under-confidence)**.** *In the same setting as Theorem 1 except that the activation function σ is general and satisfies Assumption A, let $\widehat{f}(\mathbf{x}) = \sigma(\widehat{\mathbf{w}}^\top \mathbf{x})$ be the classifier obtained from the convex ERM (5). We have simultaneously for any $p \in (0.5, 1)$ that, almost surely in the limit of $d, n \to \infty$, $d/n \to \kappa$,*

$$\Delta_p^{\mathsf{cal}}(\widehat{f}) \to C_{p,\kappa}(\sigma) = C_p(\sigma)\kappa + o(\kappa). \tag{9}$$

Further, we have the following sufficient conditions for the sign of $C_p(\sigma)$: For any $p \in (0.5, 1)$,

(a) If σ is concave at $\sigma^{-1}(p)$, i.e.,

$$\sigma''(\sigma^{-1}(p)) \le 0, \tag{10}$$

then $C_p(\sigma) > 0$, and $\widehat{f}$ is over-confident at this p for all sufficiently small κ.

(b) Conversely, if

$$\mathbb{E}_{Q_1 \sim \mathsf{N}(0, \|\mathbf{w}_\star\|^2)}[Q_1 \sigma''(Q_1)] > 0, \quad \text{and} \tag{11}$$

$$\sigma''(\sigma^{-1}(p)) - 2\sigma'(\sigma^{-1}(p)) \cdot \sigma^{-1}(p) / \|\mathbf{w}_\star\|^2 > 0, \tag{12}$$

then $C_p(\sigma) < 0$, and $\widehat{f}$ is under-confident at this p for all sufficiently small κ.

Interpretations Theorem 3 suggests that the *curvature* of the activation function σ is critical for determining its over- or under-confidence. We parse these sufficient conditions as follows:

- *Concavity of $\sigma(z)|_{z>0}$ implies over-confidence*: By part (a), at any p where $\sigma''(\sigma^{-1}(p)) \ge 0$, $\widehat{f}$ will be over-confident at that p. Moreover, any σ that is concave on the entire positive part ($\sigma''(z) \le 0$ for all $z > 0$) will result in over-confident at every $p > 0.5$. This strictly generalizes Theorem 1, and suggests that over-confidence is a common mode, as any σ that is monotone and bounded must have some concave regions on the positive part.
- *Under-confidence is possible but cannot hold at every p*: Part (b) suggests that under-confidence may be possible, provided we design σ that is sufficiently *convex* at $\sigma^{-1}(p)$ (to counteract the other term in (12)), and that additional condition (11) holds. However, as $\sigma''(z) > 0$ cannot happen for all $z > 0$, under-confidence cannot happen at every $p \in (0.5, 1)$.

Is the sufficient condition for under-confidence in Theorem 3(b) indeed possible? We give an affirmative answer.

Corollary 4 (Under-confidence can happen)**.** *There exists an activation function σ satisfying Assumption A, such that $C_p(\sigma) < 0$ for some $p \in (0.5, 1)$ and $\|\mathbf{w}_\star\| > 0$. At these p, the convex ERM (5) is under-confident in the limit of $d/n \to \kappa$ for all small κ.*

The activation we find in Corollary 4 is very close to the following activation function (up to minor tweaks in order to satisfy Assumption A):

$$\sigma_{\rm underconf}(z) = \begin{cases} 0, & z < -2\pi, \\ \dfrac{1}{2} + \dfrac{1}{4\pi}(z - \sin z), & |z| \le 2\pi, \\ 1, & z > 2\pi. \end{cases} \tag{13}$$

(See Figure 2a for the plot of this activation function.) The unique feature about this $\sigma_{\rm underconf}$ is that, unlike the logistic activation, this function is convex at all small values of $z > 0$. This leads to both the convexity condition (12) as well as the expectation condition (11) (which roughly requires the positive part in the expectation of $Q_1\sigma''(Q_1)$ to supercede the negative part).

To the best of our knowledge, this is the first known case of under-confidence for a well-specified classification problem, though we remark this under-confidence effect is weak and restricted to only a small region of p (see Figure 2c for simulation results using this activation).

5. Experiments

5.1. Simulations

We test our theories via simulations on well-specified under-parametrized logistic regression, as well as general convex ERM with the under-confident activation $\sigma_{\rm underconf}$ (13).

For both activations, we generate data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from the realizable model (3), where we fix $d = 100$, $\|\mathbf{w}_\star\| = 1$, and vary $d/n \in \{0.01, 0.05, 0.10, 0.25\}$. For each (d, n),

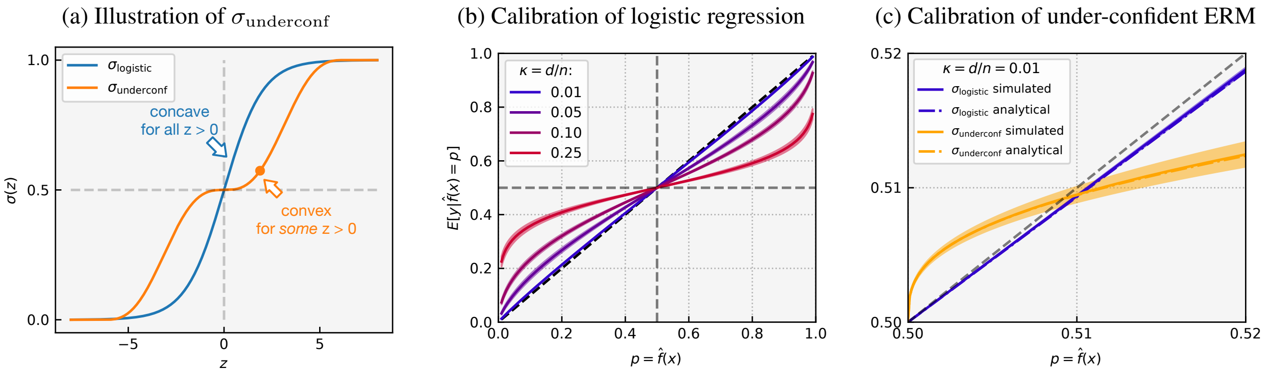


Figure 2. Binary classification simulations on realizable data. **(a)** Illustration of the activation function $\sigma_{\text{underconf}}$ constructed in Corollary 4 (cf. (13)), against the logistic (sigmoid) activation σ_{logistic}. **(b)** Calibration curves for simulated logistic regression, with $d = 100$ and $d/n \in \{0.01, 0.05, 0.1, 0.25\}$. **Logistic regression is over-confident** (prediction is higher than actual probability when prediction> 0.5) at all d/n. **(c)** Zoomed-in calibration curves for simulated realizable ERM with the $\sigma_{\text{underconf}}$ activation. In contrast to logistic regression, $\sigma_{\text{underconf}}$ **leads to under-confidence** for $p \in (0.5, 0.51)$, verifying our Theorem 3 and Corollary 4. Here "analytical" refers to our theoretical prediction $p - C_{p,\kappa}(\sigma)$ from Theorem 3. (b)(c): Shaded area are one-std error bars over 5 runs.

we generate 5 problem instances, solve the ERM problem on each instance, and plot the "calibration curves" (where the x-axis is p and y-axis is the average probability given the prediction: $\mathbb{E}[y|\widehat{f}^{(i)}(\mathbf{x}) = p] = p - \Delta_p^{\mathsf{cal}}(\widehat{f}^{(i)})$), visualizing their mean and one-standard-deviation error bar. Notice that by the closed-form expression (6), we are able to compute $\Delta_p^{\mathsf{cal}}(\widehat{f})$ exactly (using Gaussian integration) without needing to introduce a test set.

In addition to the simulated calibration curves, we also plot the limiting calibration curve suggested by Theorem 1 & 3, in which we compute the concentration values $R_\star, c_\star$ analytically by solving its defining equations (cf. Appendix C.1), and plug these values into the closed-form expression (8). This yields a curve of p against $p - C_{p,\kappa}(\sigma)$, which we compare against our simulated curves.

Results Figure 2 shows the results of our simulations. We find logistic regression indeed yields over-confident calibration curves (Figure 2b): $\mathbb{E}[y|\widehat{f}(\mathbf{x}) = p] = p - \Delta_p^{\mathsf{cal}}(\widehat{f})$ is less than p for $p > 0.5$ (and greater than p for $p < 0.5$). Further, notice that the gap Δ_p^{cal} increases as we increase κ. This agrees with our intuition that over-confidence is more severe when d/n increases (effective sample size gets lower), and further suggests that the conclusion of our theory holds more broadly than its assumptions: κ can be as large as $\kappa = 0.25$ and d can be as low as 100, both being realistic values for modeling practice.

We also find that the under-confidence shown in Corollary 4 does show up in the simulations: With the activation $\sigma_{\text{underconf}}$, $\mathbb{E}[y|\widehat{f}(\mathbf{x}) = p]$ is *higher* than p for $p \in (0.5, 0.51)$, although this range of p is fairly narrow (Figure 2c).

Finally, we observe that our theoretical prediction $C_{p,\kappa}$ closely matches the simulation: the analytical calibration curve $p - C_{p,\kappa}(\sigma)$ and the mean simulated curve are almost identical for both activations, which further confirms our theory even at a realistic $d = 100$.

5.2. CIFAR10 with pseudo labels

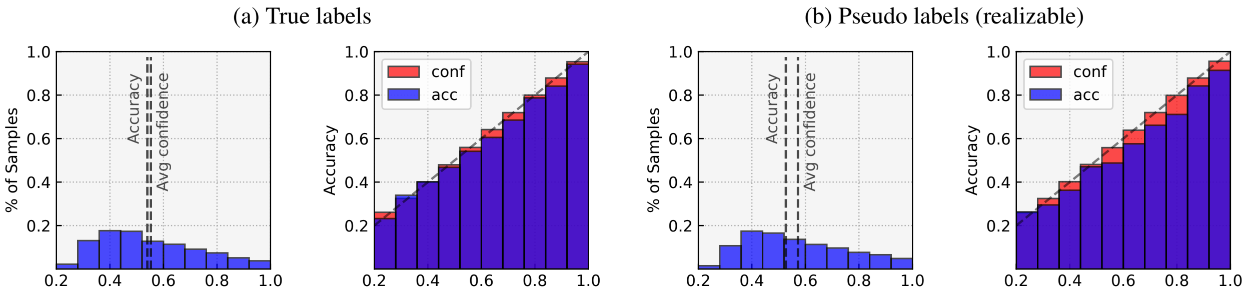
We further test the generality of our theory beyond the Gaussian input assumption and the binary classification setting. We run multi-class logistic regression on the first 5 classes of CIFAR10, which contains $n = 25000$ training images and 5000 test images, and each image has $d = 3072$ features. We perform logistic regression on two kinds of labels:

- The true label $y^{\mathsf{true}} \in \{0, 1, 2, 3, 4\}$.
- The pseudo-label $y^{\mathsf{pseudo}} \in \{0, 1, 2, 3, 4\}$ generated as follows: After fitting the logistic classifier $\widehat{\mathbf{W}} \in \mathbb{R}^{3072 \times 5}$ on the true labels, we generate pseudo-labels y_i^{pseudo} from the multi-class logistic (softmax) model

$$\mathbb{P}\left(y_i^{\mathsf{pseudo}} = k \mid \mathbf{x}_i\right) = \frac{\exp(\widehat{\mathbf{W}}_k^\top \mathbf{x}_i)}{\sum_{k'} \exp(\widehat{\mathbf{W}}_{k'}^\top \mathbf{x}_i)}.$$

The motivation for the pseudo-labels is to construct a well-specified problem (labels do come from a linear softmax model) and remove the potential effect of model-misspecification with the true labels. Note that this problem is still in the under-parametrized setting as $d < n$.

As the exact conditioning $\widehat{f}(\mathbf{x}) = p$ is no longer computable on finite data, we compute the average confidence and accuracy on the test set using binning (10 equally spaced confidence bins in $[0.2, 1.0]$), similar as in the standard practice for evaluating the ECE (Guo et al., 2017). Additional experimental details are provided in Appendix E.2.


(a) True labels

(b) Pseudo labels (realizable)

Figure 3. Calibration of multi-class logistic regression on CIFAR10's first 5-classes. The x-axes denote the confidences (predicted top probabilities) of the models. **(a)(b)**: Left: Confidence distribution across bins; Right: Average confidence against average accuracy within each bin (right); both evaluated on the test set. **(a)** Logistic regression on the true labels. **(b)** Logistic regression on pseudo-labels generated from the fitted logistic model (realizable setting) from step (a). Observe that **over-confidence happens for both the pseudo-labels generated from a multi-class logistic model, and the true labels.**

Results We find that logistic regression on (the 5-class subset of) CIFAR10 is over-confident on both the pseudo-labels and true labels (Figure 3). A closer look reveals that the over-confidence is more severe on the pseudo-labels than the true labels, yet both tasks exhibit a reasonable level of over-confidence (especially in the high confidence bins). This suggests our result that logistic regression is inherently over-confident may hold more broadly for other under-parametrized problems without strong assumptions on the input distribution, or even when the labels are not necessarily realizable by the model.

6. Overview of analysis

This section provides an overview of the two novel proof steps for our results in Section 3 & 4: (1) Characterization of the high-dimensional limit (concentration value) of logistic regression (4) and the general convex ERM (5) at small $\kappa = d/n$. (2) Determining the sign of the limiting calibration error based on the above characterization, filling in the (abbreviated) last part of the proof sketch in Section 3.1.

6.1. Local linear analysis at small κ

Let $\gamma := \|\mathbf{w}_\star\|$. By the results of Sur & Candès (2019), the values $R_\star, c_\star$ in (7) have the form $R_\star = \sqrt{\alpha_\star^2 + \kappa\sigma_\star^2}$ and $c_\star = (1 + \kappa\sigma_\star^2/\alpha_\star^2\gamma^2)^{-1/2}$, where $(\alpha_\star, \sigma_\star, \lambda_\star)$ are the solutions to the following system of nonlinear equations in three variables $(\alpha, \sigma, \lambda)$:

$$\begin{cases} \sigma^2 = \dfrac{1}{\kappa^2}\mathbb{E}\big[2\rho'(Q_1)\lambda^2\rho'(\mathsf{prox}_{\lambda\rho}(Q_2))^2\big], \\ 0 = \mathbb{E}\big[\rho'(Q_1)Q_1\lambda\rho'(\mathsf{prox}_{\lambda\rho}(Q_2))\big], \\ 1 - \kappa = \mathbb{E}\big[2\rho'(Q_1)/(1 + \lambda\rho''(\mathsf{prox}_{\lambda\rho}(Q_2)))\big]. \end{cases}$$

Above, (Q_1, Q_2) has a bivariate normal distribution with covariance depending on $(\alpha, \sigma, \kappa, \gamma)$, and prox is the prox operator. (See Theorem C.1 and Appendix C.1 for a formal statement.) These solutions are guaranteed to uniquely exist for small enough κ. However, they are only implicitly defined without closed-form expressions for these solutions, which prohibits us from analyzing their behaviors.

We overcome this issue by performing a local analysis of the solutions at small κ. We prove that, for small enough κ, we have the local linear approximation

$$\begin{aligned} \alpha_\star &= \alpha_\star(\kappa) = 1 + \bar{\alpha}_0\kappa + O(\kappa^2), \\ \sigma_\star &= \sigma_\star^2(\kappa) = \bar{\sigma}_0^2 + O(\kappa), \\ \lambda_\star &= \lambda_\star(\kappa) = \bar{\lambda}_0\kappa + O(\kappa^2), \end{aligned}$$

with closed-form expressions for $(\bar{\alpha}_0, \bar{\sigma}_0, \bar{\lambda}_0)$. For example, we have $\bar{\sigma}_0^2 = \mathbb{E}[\rho'(Q_1)\rho'(-Q_1)]/(\mathbb{E}[\rho''(Q_1)])^2$ where $Q_1 \sim \mathsf{N}(0, \gamma^2)$. (See Lemma C.1 for the formal statement.) These approximations imply similar approximations for $R_\star, c_\star$, which allows us to analyze the behavior of the limiting calibration error (8) locally at small κ.

6.2. Determining sign of the limiting calibration error

Towards proving Theorem 3 & 1, it remains for us to derive the sufficient conditions for the sign of $C_{p,\kappa}(\sigma)$. Using the above local linear approximation for $(R_\star, c_\star)$ and performing first-order calculus, we obtain

$$\begin{aligned} \lim_{\kappa\to 0}\frac{C_{p,\kappa}(\sigma)}{\kappa} &= \sigma'(\sigma^{-1}(p)) \cdot \sigma^{-1}(p) \cdot \big(\bar{\alpha}_0 + \bar{\sigma}_0^2/\gamma^2\big) \\ &\quad - \frac{1}{2}\sigma''(\sigma^{-1}(p)) \cdot \bar{\sigma}_0^2. \end{aligned}$$

(Lemma D.1). We prove that $\bar{\alpha}_0 + \bar{\sigma}_0^2/\gamma^2 > 0$ always holds regardless of the activation,γ, and p. This implies that, as long as $\sigma''(\sigma^{-1}(p)) \le 0$, the right-hand side in the equation above is positive. This gives part (a) of Theorem 3. On the other hand, if $\sigma''(\sigma^{-1}(p)) > 2\sigma'(\sigma^{-1}(p))\sigma^{-1}(p)/\gamma^2$ and $\bar{\alpha}_0 < 0$, then the right-hand side above is negative. These are exactly the sufficient conditions we required in part (b) of Theorem 3.

7. Conclusion

This paper provides a precise theoretical study of the calibration error of logistic regression and a class of general binary classification problems. We show that logistic regression is inherently over-confident by $\Theta(d/n)$ when n/d is large, and establish sufficient conditions for the over- or under-confidence of unregularized ERM for general binary classification. Our results reveal that (1) Over-confidence is not just a result of over-parametrization; (2) Over-confidence is a common mode but not universal. We believe our work opens up a number of future questions, such as the interplay between calibration and model training (or regularization), or theoretical studies of calibration on nonlinear models.

Acknowledgment

We thank Yuxin Chen for the insightful discussions on the system of nonlinear equations. We thank all anonymous reviewers for the many helpful feedback on this paper.

References

Albert, A. and Anderson, J. A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011.

Begoli, E., Bhattacharya, T., and Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Candès, E. J., Sur, P., et al. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

Ding, Z., Han, X., Liu, P., and Niethammer, M. Local temperature scaling for probability calibration. *arXiv preprint arXiv:2008.05105*, 2020.

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Dusenberry, M., Jerfel, G., Wen, Y., Ma, Y., Snoek, J., Heller, K., Lakshminarayanan, B., and Tran, D. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pp. 2782–2792. PMLR, 2020.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110 (36):14557–14562, 2013.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gal, Y., Hron, J., and Kendall, A. Concrete dropout. *arXiv preprint arXiv:1705.07832*, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Gupta, C., Podkopaev, A., and Ramdas, A. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *arXiv preprint arXiv:2006.10564*, 2020.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Jung, C., Lee, C., Pai, M. M., Roth, A., and Vohra, R. Moment multicalibration for uncertainty estimation. *arXiv preprint arXiv:2008.08037*, 2020.

Kakade, S., Kalai, A. T., Kanade, V., and Shamir, O. Efficient learning of generalized linear and single index models with isotonic regression. *arXiv preprint arXiv:1104.2018*, 2011.

Karoui, N. E. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

Kull, M., Silva Filho, T., and Flach, P. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pp. 623–631. PMLR, 2017.

Kull, M., Perello-Nieto, M., Kängsepp, M., Song, H., Flach, P., et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *arXiv preprint arXiv:1910.12656*, 2019.

Kumar, A., Liang, P., and Ma, T. Verified uncertainty calibration. *arXiv preprint arXiv:1909.10155*, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Liu, J. Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv preprint arXiv:2006.10108*, 2020.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pp. 4051–4060. PMLR, 2019.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32:13153–13164, 2019.

Mai, X., Liao, Z., and Couillet, R. 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)*, pp. 3357–3361. IEEE, 2019.

Malinin, A., Mlodozeniec, B., and Gales, M. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

McCullagh, P. *Generalized linear models*. Routledge, 2018.

Michelmore, R., Kwiatkowska, M., and Gal, Y. Evaluating uncertainty quantification in end-to-end autonomous driving control. *arXiv preprint arXiv:1811.06817*, 2018.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H., and Dokania, P. K. Calibrating deep neural networks using focal loss. *arXiv preprint arXiv:2002.09437*, 2020.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pp. 13991–14002, 2019.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. 1999.

Rahimi, A., Shaban, A., Cheng, C.-A., Hartley, R., and Boots, B. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Shabat, E., Cohen, L., and Mansour, Y. Sample complexity of uniform convergence for multicalibration. *arXiv preprint arXiv:2005.01757*, 2020.

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

Stojnic, M. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

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

Taheri, H., Pedarsani, R., and Thrampoulidis, C. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. *arXiv preprint arXiv:2006.08917*, 2020.

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pp. 1683–1709. PMLR, 2015.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *arXiv preprint arXiv:1905.11001*, 2019.

Tran, L., Veeling, B. S., Roth, K., Swiatkowski, J., Dillon, J. V., Snoek, J., Mandt, S., Salimans, T., Nowozin, S., and Jenatton, R. Hydra: Preserving ensemble diversity for model distillation. *arXiv preprint arXiv:2001.04694*, 2020.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wen, Y., Tran, D., and Ba, J. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklf1yrYDr.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. Citeseer.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, J., Kailkhura, B., and Han, T. Y.-J. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, pp. 11117–11128. PMLR, 2020.

A. Technical tools

A.1. The prox operator

For any loss function ρ that satisfies Assumption A, for any $\lambda > 0$, we define its proximal mapping operator via

$$\mathsf{prox}_{\lambda\rho}(z) \equiv \arg\min_{t\in\mathbb{R}} \Big\{ \lambda\rho(t) + \frac{1}{2}(t-z)^2 \Big\}. \tag{14}$$

Lemma A.1 (Properties of the Proximal mapping operator)**.** *For any loss function ρ that satisfies Assumption A, its proximal mapping operator $\mathsf{prox}_{\lambda\rho}(z)$ is differentiable with respect to $z \in \mathbb{R}$ and $\lambda > 0$. Moreover, we have*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}z}\mathsf{prox}_{\lambda\rho}(z) &= \frac{1}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(z))}, \\ \frac{\mathrm{d}}{\mathrm{d}\lambda}\mathsf{prox}_{\lambda\rho}(z) &= -\frac{\rho'(\mathsf{prox}_{\lambda\rho}(z))}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(z))}. \end{aligned}$$

Proof of Lemma A.1. Denote $P(\lambda, z) = \mathsf{prox}_{\lambda\rho}(z)$. For any loss function ρ that satisfies Assumption A, the proximal mapping operator $P = P(\lambda, z)$ satisfies the following equation

$$\lambda\rho'(P(\lambda,z)) + P(\lambda,z) - z = 0.$$

Taking derivative with respect to z, we have

$$\lambda\rho''(P(\lambda,z))\partial_z P(\lambda,z) + \partial_z P(\lambda,z) - 1 = 0,$$

which gives

$$\partial_z P(\lambda,z) = \frac{1}{1+\lambda\rho''(P(\lambda,z))}.$$

Taking derivative with respect to λ, we have

$$\rho'(P(\lambda,z)) + \lambda\rho''(P(\lambda,z))\partial_\lambda P(\lambda,z) + \partial_\lambda P(\lambda,z) = 0,$$

which gives

$$\partial_\lambda P(\lambda,z) = -\frac{\rho'(P(\lambda,z))}{1+\lambda\rho''(P(\lambda,z))}.$$

This proves the lemma. □

A.2. Implicit function theorem

We state the standard implicit function theorem in the following.

Lemma A.2 (Implicit function theorem)**.** *Let $\boldsymbol{F}(\boldsymbol{p},\kappa) : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a continuously differentiable vector-valued function on $\mathsf{B}(\boldsymbol{p}_0,\varepsilon) \times [0,\bar{\kappa}_0)$ for some $\bar{\kappa}_0 > 0$. Suppose $\boldsymbol{F}(\boldsymbol{p}_0, 0) = 0$ and*

$$\sigma_{\min}(\nabla_{\boldsymbol{p}} F(\boldsymbol{p}_0, 0)) > 0.$$

Then there exists a constant $\kappa_0 > 0$ and a continuous differentiable path $\boldsymbol{p}_\star(t) \in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)$, such that

$$\boldsymbol{F}(\boldsymbol{p}_\star(\kappa),\kappa) = 0, \quad \forall \kappa \in [0,\kappa_0).$$

A.3. Consistency of the convex ERM for general activation

Recall

$$R(\mathbf{w}) = \mathbb{E}\big[\rho(\mathbf{w}^\top\mathbf{x}) - y\mathbf{w}^\top\mathbf{x}\big]$$

and $\mathbb{E}[y|\mathbf{x}] = \sigma(\mathbf{w}_\star^\top\mathbf{x})$. If $\rho' = \sigma$, then we have

$$\nabla R(\mathbf{w}) = \mathbb{E}\big[\rho'(\mathbf{w}^\top\mathbf{x})\mathbf{x} - y\mathbf{x}\big] = \mathbb{E}\big[\big(\sigma(\mathbf{w}^\top\mathbf{x}) - \sigma(\mathbf{w}_\star^\top\mathbf{x})\big)\mathbf{x}\big].$$

This shows that $\nabla R(\mathbf{w}_\star) = 0$, which further by the convexity of ρ implies that $\mathbf{w}_\star \in \arg\min_{\mathbf{w}} R(\mathbf{w})$. □

B. Closed form expression for calibration error

Lemma B.1 (Closed-form expression for calibration error)**.** *Assume* $(\mathbf{X}, Y)$ *follows the binary linear model* (3) *with true coefficient* $\mathbf{w}_\star \in \mathbb{R}^d$*, and the activation function* σ *satisfies Assumption A. For any* $\mathbf{w} \in \mathbb{R}^d$*, and* $f_\mathbf{w}(\mathbf{x}) := \sigma(\mathbf{w}^\top \mathbf{x})$*, we have for any* $p \in (0,1)$ *that*

$$\Delta_p^{\mathsf{cal}}(f_\mathbf{w}) = p - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\left[\sigma\left(\frac{\|\mathbf{w}_\star\|}{\|\mathbf{w}\|}\cos\theta \cdot \sigma^{-1}(p) + \sin\theta\,\|\mathbf{w}_\star\|\,Z\right)\right], \tag{15}$$

where $\theta = \angle(\mathbf{w}, \mathbf{w}_\star)$ *is the angle between* $\mathbf{w}$ *and* $\mathbf{w}_\star$*.*

Proof. Recall by definition (2) that

$$\Delta_p^{\mathsf{cal}}(f_\mathbf{w}) = p - \mathbb{E}_{\mathbf{x},y}[y \mid f_\mathbf{w}(\mathbf{x}) = p] = p - \mathbb{E}_\mathbf{x}\big[\sigma(\mathbf{w}_\star^\top \mathbf{x}) \mid \sigma(\mathbf{w}^\top\mathbf{x}) = p\big].$$

As $\mathbf{x} \sim \mathsf{N}(0,1)$, the conditional distribution of $\mathbf{x}|\sigma(\mathbf{w}^\top\mathbf{x}) = p$ can be characterized precisely: Under this distribution we have $\mathbf{w}^\top\mathbf{x} = \sigma^{-1}(p)$ and $\mathbf{V}^\top\mathbf{x} \sim \mathsf{N}(0, \mathbf{I}_{d-1})$, where $\mathbf{V} \in \mathbb{R}^{d\times(d-1)}$ is the orthogonal complement subspace of $\mathbf{w}$. Therefore, conditioned on $\sigma(\mathbf{w}^\top\mathbf{x}) = p$, we have

$$\begin{aligned}
\mathbf{w}_\star^\top\mathbf{x} &= \mathbf{w}_\star^\top \frac{1}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top\mathbf{x} + \mathbf{w}_\star^\top\mathbf{V}^\top\mathbf{x} \\
&\overset{d}{=} \frac{\mathbf{w}_\star^\top\mathbf{w}}{\|\mathbf{w}\|^2}\sigma^{-1}(p) + \|\mathbf{w}_\star\|\sin\angle(\mathbf{w}_\star, \mathbf{w})\cdot Z = \frac{\|\mathbf{w}_\star\|}{\|\mathbf{w}\|}\cos\theta\cdot\sigma^{-1}(p) + \|\mathbf{w}\|_\star \sin\theta\cdot Z,
\end{aligned}$$

where $\overset{d}{=}$ denotes equal in distribution, and $Z \sim \mathsf{N}(0,1)$. This representation of the condition distribution yields the desired expression for $\Delta_p^{\mathsf{cal}}(f_\mathbf{w})$. □

C. Analysis of convex ERM in the high-dimensional limit

C.1. Concentration to a system of equations

We define the following system of equations in three variables $(\alpha, \sigma, \lambda) \in \mathbb{R}^3$, with two parameters $\kappa > 0$ and $\gamma > 0$:

$$\begin{cases}
\sigma^2 = \dfrac{1}{\kappa^2}\mathbb{E}\big[2\rho'(Q_1)\lambda^2\rho'(\mathsf{prox}_{\lambda\rho}(Q_2))^2\big], & \text{(16a)}\\
0 = \mathbb{E}\big[\rho'(Q_1)Q_1\lambda\rho'(\mathsf{prox}_{\lambda\rho}(Q_2))\big], & \text{(16b)}\\
1-\kappa = \mathbb{E}\left[\dfrac{2\rho'(Q_1)}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(Q_2))}\right], & \text{(16c)}
\end{cases}$$

where (Q_1, Q_2) follows a joint normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \gamma^2 & -\alpha\gamma^2 \\ -\alpha\gamma^2 & \alpha^2\gamma^2 + \kappa\sigma^2 \end{bmatrix}.$$

Let

$$(\alpha_\star, \sigma_\star, \lambda_\star) = (\alpha_\star(\kappa), \beta_\star(\kappa), \lambda_\star(\kappa)) \tag{17}$$

denote the solution to (16) whenever the solution exists and is unique (dropping dependence on γ for notational simplicity).

It is recently shown that the solutions $(\alpha_\star, \sigma_\star, \lambda_\star)$ is closely connected to the high-dimensional limit of the convex ERM (5). We state this in the following result.

Theorem C.1 (Concentration of the ERM (5); restatement of Theorem 2, (Sur & Candès, 2019))**.** *Let Assumption A hold. For any* $\gamma = \|\mathbf{w}_\star\|$*, there exists a* $\kappa_0 = \kappa_0(\gamma) > 0$ *such that solution* $(\alpha_\star, \sigma_\star, \lambda_\star)$ *defined in* (17) *uniquely exists in the positive region* $\Omega = \{(\alpha, \sigma, \lambda) : \alpha > 0, \sigma > 0, \lambda > 0\}$ *for any* $\kappa \in (0, \kappa_0]$*.*

Further, let $\widehat{\mathbf{w}}$ denote the solution to the convex ERM (5), and assume the data is generated from the binary linear model (3) with $\|\mathbf{w}_\star\| = \gamma$ fixed. As $d, n \to \infty$ with $d/n \to \kappa$ where $\kappa \in (0, \kappa_0]$ is a fixed constant, we have almost surely that $\widehat{\mathbf{w}}$ exists, and concentrates in the sense that

$$\begin{aligned} &\|\widehat{\mathbf{w}}\| \to R_\star = R_\star(\kappa,\gamma) := \sqrt{\alpha_\star^2\gamma^2 + \kappa\sigma_\star^2}, \\ &\cos\widehat{\theta} := \frac{\widehat{\mathbf{w}}^\top \mathbf{w}_\star}{\|\widehat{\mathbf{w}}\|\,\|\mathbf{w}_\star\|} \to c_\star = c_\star(\kappa,\gamma) := \frac{1}{\sqrt{1+\kappa\sigma_\star^2/\alpha_\star^2\gamma^2}}. \end{aligned} \tag{18}$$

Combining Lemma B.1 with our the concentration result in Theorem C.1 directly implies the following

Corollary C.1 (Concentration of the calibration error)**.** *In the same setting as Theorem C.1, the calibration error $\Delta_p^{\mathsf{cal}}(\widehat{f})$ (of the convex ERM) converges almost surely to the following limit as $d, n \to \infty$, $d/n \to \kappa$:*

$$\Delta_p^{\mathsf{cal}}(\widehat{f}) \to C_{p,\kappa}(\sigma) := p - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\left[\sigma\left(\frac{\|\mathbf{w}_\star\|}{R_\star}c_\star \cdot \sigma^{-1}(p) + \sqrt{1-c_\star^2}\,\|\mathbf{w}_\star\|\,Z\right)\right], \tag{19}$$

where $R_\star, c_\star$ are defined in (18).

Proof. This is a direct consequence of the concentration result of Theorem C.1. By Lemma B.1 we have for the ERM learned classifier $\widehat{f}(\mathbf{x}) = \sigma(\widehat{\mathbf{w}}^\top\mathbf{x})$ that

$$\Delta_p^{\mathsf{cal}}(\widehat{f}) = p - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\left[\sigma\left(\frac{\|\mathbf{w}_\star\|}{\|\widehat{\mathbf{w}}\|}\cos\widehat{\theta}\cdot\sigma^{-1}(p) + \sin\widehat{\theta}\,\|\mathbf{w}_\star\|\,Z\right)\right],$$

where $\cos\widehat{\theta} = \mathbf{w}_\star^\top\widehat{\mathbf{w}}/\|\mathbf{w}_\star\|\,\|\widehat{\mathbf{w}}\|$ is the cosine similarity between $\mathbf{w}_\star$ and $\widehat{\mathbf{w}}$. Now, by Theorem C.1, we have $\|\widehat{\mathbf{w}}\| \to R_\star$, $\cos\widehat{\theta} \to c_\star$, and $\sin\widehat{\theta} \to \sqrt{1-c_\star^2}$ with probability one as $n, d \to \infty$, $d/n \to \kappa$. (Note that the sign of $\sin\widehat{\theta}$ does not make a difference here as $Z \sim \mathsf{N}(0,1) \overset{d}{=} -Z$.) Therefore, the random variable inside the expectation in the above converges to the random variable

$$\sigma\left(\frac{\|\mathbf{w}_\star\|}{R_\star}c_\star\cdot\sigma^{-1}(p) + \sqrt{1-c_\star^2}\,\|\mathbf{w}_\star\|\,Z\right)$$

with probability one. Applying the bounded convergence theorem (since $|\sigma(\cdot)| \le 1$) yields that $\Delta_p^{\mathsf{cal}}(\widehat{f})$ converges to

$$C_{p,\kappa}(\sigma) := p - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\left[\sigma\left(\frac{\|\mathbf{w}_\star\|}{R_\star}c_\star\cdot\sigma^{-1}(p) + \sqrt{1-c_\star^2}\,\|\mathbf{w}_\star\|\,Z\right)\right],$$

as desired. □

C.2. Analysis of solution at small κ

We now analyze the solution $(\alpha_\star, \beta_\star, \kappa_\star)$ when κ is small. Throughout this section, we let $Q_2 = -\alpha Q_1 + \sqrt{\kappa}\sigma Z$, so that $Z \sim \mathsf{N}(0,1)$ and is independent of Q_1. Then we have $(Q_1, Q_2) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = [\gamma^2, -\alpha\gamma^2; -\alpha\gamma^2, \alpha^2\gamma^2 + \kappa\sigma^2]$.

For any $(\bar{\alpha}, \bar{\sigma}, \bar{\lambda}, \kappa) \in \overline{\Omega} \times (0, \kappa_0)$ where $\overline{\Omega} = \{(\bar{\alpha}, \bar{\sigma}, \bar{\lambda}) : \bar{\alpha} > -1/\kappa, \bar{\sigma} > 0, \bar{\lambda} > 0\}$, we consider the following equivalent version of the system of equations (16), defined via

$$\begin{aligned} F_1(\bar{\alpha},\bar{\sigma},\bar{\lambda},\kappa) &= \bar{\sigma}^2 - \bar{\lambda}^2\mathbb{E}[2\rho'(Q_1)\rho'(\mathsf{prox}_{\kappa\bar{\lambda}\rho}(-(1+\kappa\bar{\alpha})Q_1 + \sqrt{\kappa}\bar{\sigma}Z))^2], \\ F_2(\bar{\alpha},\bar{\sigma},\bar{\lambda},\kappa) &= \kappa^{-1}\mathbb{E}[\rho'(Q_1)Q_1\rho'(\mathsf{prox}_{\kappa\bar{\lambda}\rho}(-(1+\kappa\bar{\alpha})Q_1 + \sqrt{\kappa}\bar{\sigma}Z))], \\ F_3(\bar{\alpha},\bar{\sigma},\bar{\lambda},\kappa) &= \kappa^{-1}\{1 - \kappa - \mathbb{E}[2\rho'(Q_1)/[1 + \kappa\bar{\lambda}\rho''(\mathsf{prox}_{\kappa\bar{\lambda}\rho}(-(1+\kappa\bar{\alpha})Q_1 + \sqrt{\kappa}\bar{\sigma}Z))]]\}. \end{aligned} \tag{20}$$

We further define $\boldsymbol{p} = (\bar{\alpha}, \bar{\sigma}, \bar{\lambda})$ as a shorthand for the three variables. For $(\boldsymbol{p}, \kappa) \in \overline{\Omega} \times (0, \kappa_0)$, we let

$$\boldsymbol{F}(\boldsymbol{p},\kappa) = (F_1(\boldsymbol{p},\kappa), F_2(\boldsymbol{p},\kappa), F_3(\boldsymbol{p},\kappa)). \tag{21}$$

Equation (21) and the system (16) are equivalent up to a change of variables: For any fixed κ, any solution $(\alpha_\star, \sigma_\star, \lambda_\star)$ of Eq. (16) yields a solution $((\alpha_\star - 1)/\kappa, \sigma_\star, \kappa\lambda_\star, \kappa)$ of $\boldsymbol{F}(\boldsymbol{p},\kappa) = \mathbf{0}$ where $\boldsymbol{F}$ is defined as in Eq. (21), and vice versa. As a consequence, Theorem C.1 implies that, for any $\kappa \in (0, \kappa_0)$, there exists a unique solution $\boldsymbol{p}(\kappa)$ of $\boldsymbol{F}(\boldsymbol{p}, \kappa) = 0$ in the corresponding region $\overline{\Omega} = \{(\bar{\alpha}, \bar{\sigma}, \bar{\lambda}) : \bar{\alpha} > -1/\kappa, \bar{\sigma} > 0, \bar{\lambda} > 0\}$.

Now we define

$$\bar{\alpha}_0 := -\frac{\mathbb{E}[Q_1 \rho'''(Q_1)] \cdot \mathbb{E}[\rho'(Q_1)\rho'(-Q_1)]}{2\mathbb{E}[Q_1^2 \rho''(Q_1)] \cdot \mathbb{E}[\rho''(Q_1)]^2}, \tag{22a}$$

$$\bar{\sigma}_0^2 := \frac{\mathbb{E}[\rho'(Q_1)\rho'(-Q_1)]}{(\mathbb{E}[\rho''(Q_1)])^2} > 0, \tag{22b}$$

$$\bar{\lambda}_0 := \frac{1}{\mathbb{E}[\rho''(Q_1)]} > 0, \tag{22c}$$

where the expectation is taken with respect to $Q_1 \sim \mathsf{N}(0, \gamma^2)$.

Lemma C.1 (Local linear analysis of solutions at small κ)**.** *In the same setting as Theorem C.1, the solutions $(\alpha_\star, \sigma_\star, \lambda_\star)$ defined in (17) satisfy*

$$\alpha_\star(\kappa) = 1 + \bar{\alpha}_0 \kappa + o(\kappa), \tag{23a}$$

$$\sigma_\star(\kappa) = \bar{\sigma}_0 + o(1), \tag{23b}$$

$$\lambda_\star(\kappa) = \bar{\lambda}_0 \kappa + o(\kappa), \tag{23c}$$

where $(\bar{\alpha}_0, \bar{\sigma}_0, \bar{\lambda}_0)$ are defined in (22), and $o(\cdot)$ is the standard small-o notation ($o(\kappa)/\kappa \to 0$ as $\kappa \to 0$).

Lemma C.1 formalizes the arguments we sketched in Section 6.1. (We note that Section 6.1 stated the error terms in terms of the slightly stronger $O(\kappa^2)$ instead of $o(\kappa)$, yet $o(\kappa)$ as stated above is sufficient for our subsequent results.) The rest of this section is devoted to proving Lemma C.1.

C.3. Proof of Lemma C.1

We take $\boldsymbol{p}_0 = (\bar{\alpha}_0, \bar{\sigma}_0, \bar{\lambda}_0)$, where $\bar{\alpha}_0, \bar{\sigma}_0, \bar{\lambda}_0$ are defined in (22). Observe that, in order to prove (23), it suffices to prove that under the change of variables

$$(\bar{\alpha}(\kappa), \bar{\sigma}(\kappa), \bar{\lambda}(\kappa)) := \left(\frac{\alpha_\star(\kappa) - 1}{\kappa}, \sigma_\star(\kappa), \frac{\lambda_\star(\kappa)}{\kappa}\right),$$

the new set of solutions obey

$$\begin{aligned} \bar{\alpha}(\kappa) &= \bar{\alpha}_0 + o(1), \\ \bar{\sigma}(\kappa) &= \bar{\sigma}_0 + o(1), \\ \bar{\lambda}(\kappa) &= \bar{\lambda}_0 + o(1). \end{aligned}$$

In other words, letting $\boldsymbol{p}(\kappa) = (\bar{\alpha}(\kappa), \bar{\sigma}(\kappa), \bar{\lambda}(\kappa))$ denote the solution to the transformed equation $\boldsymbol{F}(\boldsymbol{p}, \kappa) = 0$, we wish to prove that $\boldsymbol{p}(\kappa) \to \boldsymbol{p}_0$. We achieve this through a continuity analysis of the function $\boldsymbol{F}$ and the implicit function theorem. The key intermediate results are the following two auxiliary lemmas, whose proofs are deferred to Section C.3.1 and C.3.2 respectively.

Lemma C.2. *Let Assumption A hold. Let $\boldsymbol{F}$ be as defined in Eq. (21). Then for any ε such that $\mathsf{B}(\boldsymbol{p}_0, 2\varepsilon) \subseteq \overline{\Omega} = \{(\bar{\alpha}, \bar{\sigma}, \bar{\lambda}) : \bar{\alpha} > -1/\kappa, \bar{\sigma} > 0, \bar{\lambda} > 0\}$, there exists a continuous matrix function $\boldsymbol{J} : \mathsf{B}(\boldsymbol{p}_0, \varepsilon) \to \mathbb{R}^{3\times 3}$ with*

$$\sigma_{\min}(\boldsymbol{J}(\boldsymbol{p}_0)) > 0, \tag{24}$$

and

$$\lim_{\kappa \to 0} \sup_{\boldsymbol{p} \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)} \left\| \nabla_{\boldsymbol{p}} \boldsymbol{F}(\boldsymbol{p}, \kappa) - \boldsymbol{J}(\boldsymbol{p}) \right\|_{\mathrm{op}} = 0.$$

Lemma C.3. *Let Assumption A hold. Let $\boldsymbol{F}$ be as defined in Eq. (21). Then for any ε such that $\mathsf{B}(\boldsymbol{p}_0, 2\varepsilon) \subseteq \overline{\Omega} = \{(\bar{\alpha}, \bar{\sigma}, \bar{\lambda}) : \bar{\alpha} > -1/\kappa, \bar{\sigma} > 0, \bar{\lambda} > 0\}$, there exists two continuous vector functions $\boldsymbol{F}_0, \boldsymbol{g} : \mathsf{B}(\boldsymbol{p}_0, \varepsilon) \to \mathbb{R}^3$ such that*

$$\lim_{\kappa \to 0} \sup_{\boldsymbol{p} \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)} \left\| \boldsymbol{F}(\boldsymbol{p}, \kappa) - \boldsymbol{F}_0(\boldsymbol{p}) \right\|_2 = 0,$$

$$\lim_{\kappa \to 0} \sup_{\boldsymbol{p} \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)} \left\| \partial_\kappa \boldsymbol{F}(\boldsymbol{p}, \kappa) - \boldsymbol{g}(\boldsymbol{p}) \right\|_2 = 0.$$

Moreover, we have

$$\lim_{\kappa\to 0+} \boldsymbol{F}(\boldsymbol{p}_0, \kappa) = \boldsymbol{F}_0(\boldsymbol{p}_0) = 0.$$

By Lemma C.2 and C.3, we can continuously extend the function $\boldsymbol{F}$ to the region $\mathsf{B}(\boldsymbol{p}_0, \varepsilon) \times [0, \kappa_0)$ for some small κ_0, such that $\boldsymbol{F}(\boldsymbol{p}, \kappa)$ is continuously differentiable in the same region. Moreover, by Lemma C.3, we have $\boldsymbol{F}(\boldsymbol{p}_0, 0) = \lim_{\kappa\to 0} \boldsymbol{F}(\boldsymbol{p}_0, \kappa) = 0$. Finally, by Lemma C.2, we have $\sigma_{\min}(\nabla_{\boldsymbol{p}} F(\boldsymbol{p}_0, 0)) > 0$.

Therefore, the conditions in the Implicit Function Theorem (Lemma A.2) are satisfied, from which we can conclude that there exists a continuously differentiable path $\{\boldsymbol{p}(\kappa); \kappa \in [0, \kappa_0)\} \subset \mathsf{B}(\boldsymbol{p}_0, \varepsilon)$, such that $\boldsymbol{F}(\boldsymbol{p}(\kappa), \kappa) = 0$ for any $\kappa \in [0, \kappa_0)$. By Theorem C.1 , for any $\kappa \in [0, \kappa_0)$, the equation $\boldsymbol{F}(\boldsymbol{p}, \kappa) = 0$ has at most one solution in Ω, which should be $\boldsymbol{p}(\kappa)$. By the differentiability of the path $\boldsymbol{p}(\kappa)$ at $\kappa = 0$ (as a conclusion of Lemma A.2), we get $\boldsymbol{p}(\kappa) \to \boldsymbol{p}_0$ as $\kappa \to 0$. This proves Lemma C.1. □

C.3.1. Proof of Lemma C.2

Throughout this proof, we let $Q_2 = -(1 + \kappa\bar{\alpha})Q_1 + \sqrt{\kappa}\bar{\sigma}Z$, so that $Z \sim \mathsf{N}(0, 1)$ and is independent of Q_1. We define $P = \mathsf{prox}_{\kappa\bar{\lambda}\rho}(Q_2)$. Then by Lemma A.1, we can define

$$\begin{aligned}
P_z &= \frac{\mathrm{d}}{\mathrm{d}z}\mathsf{prox}_{\kappa\bar{\lambda}\rho}(z)|_{z=Q_2} = 1/(1 + \kappa\bar{\lambda}\rho''(P)),\\
P_\lambda &= \kappa^{-1}\frac{\mathrm{d}}{\mathrm{d}\bar{\lambda}}\mathsf{prox}_{\kappa\bar{\lambda}\rho}(z)|_{z=Q_2} = -\rho'(P)/(1 + \kappa\bar{\lambda}\rho''(P)).
\end{aligned}$$

The derivatives of $\boldsymbol{F}$ gives

$$\begin{aligned}
\partial_{\bar{\alpha}} F_1 &= 4\bar{\lambda}^2\mathbb{E}[\rho'(Q_1)\rho'(P)\rho''(P)P_z Q_1]\kappa,\\
\partial_{\bar{\sigma}} F_1 &= 2\bar{\sigma} - 4\bar{\lambda}^2\mathbb{E}[\rho'(Q_1)\rho'(P)\rho''(P)P_z Z]\sqrt{\kappa}\\
&= 2\bar{\sigma} - 4\bar{\lambda}^2\mathbb{E}\Big[\rho'(Q_1)P_z\frac{\rho''(P)^2 + \kappa\bar{\lambda}\rho''(P)^3 + \rho'(P)\rho'''(P)}{(1 + \kappa\bar{\lambda}\rho''(P))^2}\Big]\bar{\sigma}\kappa\\
&= 2\bar{\sigma} - 4\bar{\lambda}^2\mathbb{E}\Big[\rho'(Q_1)P_z^3\Big(\rho''(P)^2 + \kappa\bar{\lambda}\rho''(P)^3 + \rho'(P)\rho'''(P)\Big)\Big]\bar{\sigma}\kappa,\\
\partial_{\bar{\lambda}} F_1 &= -4\bar{\lambda}\mathbb{E}[\rho'(Q_1)\rho'(P)^2] - 4\bar{\lambda}\mathbb{E}[\rho'(Q_1)\rho'(P)\rho''(P)P_\lambda]\kappa,\\
\partial_{\bar{\alpha}} F_2 &= -\mathbb{E}[\rho'(Q_1)\rho''(P)P_z Q_1^2],\\
\partial_{\bar{\sigma}} F_2 &= \mathbb{E}[\rho'(Q_1)Q_1\rho''(P)P_z Z]\kappa^{-1/2} = \mathbb{E}[\rho'(Q_1)Q_1\rho'''(P)P_z^3]\bar{\sigma},\\
\partial_{\bar{\lambda}} F_2 &= \mathbb{E}[\rho'(Q_1)Q_1\rho''(P)P_\lambda],\\
\partial_{\bar{\alpha}} F_3 &= -2\mathbb{E}[\rho'(Q_1)P_z^2\rho'''(P)P_z Q_1]\kappa\bar{\lambda},\\
\partial_{\bar{\sigma}} F_3 &= 2\mathbb{E}[\rho'(Q_1)P_z^2\kappa\bar{\lambda}\rho'''(P)P_z Z]\sqrt{\kappa}\kappa^{-1}\\
&= 2\mathbb{E}[\rho'(Q_1)(\rho''''(P) + \kappa\bar{\lambda}(\rho''(P)\rho''''(P) - 3\rho'''(P)^2))P_z^5]\bar{\sigma}\bar{\lambda}\kappa,\\
\partial_{\bar{\lambda}} F_3 &= 2\mathbb{E}[\rho'(Q_1)P_z^2(\rho''(P) + \kappa\bar{\lambda}\rho'''(P)P_\lambda)]\\
&= 2\mathbb{E}[\rho'(Q_1)P_z^2\rho''(P)] + 2\kappa\bar{\lambda}\mathbb{E}[\rho'(Q_1)P_z^2\rho'''(P)P_\lambda].
\end{aligned}$$

We next show that

$$\lim_{\kappa\to 0}\sup_{\boldsymbol{p}\in\mathsf{B}(\boldsymbol{p}_0,\varepsilon)} \|\nabla_{\boldsymbol{p}}\boldsymbol{F}(\boldsymbol{p}, \kappa) - \boldsymbol{J}(\boldsymbol{p})\|_{\mathrm{op}} = 0, \tag{25}$$

where

$$\begin{aligned}
\boldsymbol{J}(\boldsymbol{p}) &= \begin{bmatrix} 0 & 2\bar{\sigma} & -4\bar{\lambda}\mathbb{E}[\rho'(Q_1)\rho'(-Q_1)^2] \\ -\mathbb{E}[\rho'(Q_1)\rho''(Q_1)Q_1^2] & \mathbb{E}[\rho'(Q_1)Q_1\rho'''(-Q_1)]\bar{\sigma} & -\mathbb{E}[\rho'(Q_1)\rho'(-Q_1)Q_1\rho''(-Q_1)] \\ 0 & 0 & 2\mathbb{E}[\rho'(Q_1)\rho''(-Q_1)] \end{bmatrix}\\
&= \begin{bmatrix} 0 & 2\bar{\sigma} & -4\bar{\lambda}\mathbb{E}[\rho'(Q_1)\rho'(-Q_1)^2] \\ -(1/2)\mathbb{E}[\rho''(Q_1)Q_1^2] & \mathbb{E}[\rho'(Q_1)Q_1\rho'''(-Q_1)]\bar{\sigma} & -\mathbb{E}[\rho'(Q_1)\rho'(-Q_1)Q_1\rho''(Q_1)] \\ 0 & 0 & \mathbb{E}[\rho''(Q_1)] \end{bmatrix},
\end{aligned} \tag{26}$$

where the last equality is by the fact that $\rho'(z) - 1/2$ is an odd function.

The above result follows from (uniform) continuity arguments using the fact that ρ is 4-th continuously differentiable: $\sup_z |\rho^{(k)}(z)| < \infty$ for $k = 1, 2, 3, 4$. In the following, we show the result for the $(3, 3)$-th entry: $\lim_{\kappa\to 0}\sup_{\boldsymbol{p}\in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)} |\nabla_{\bar\lambda} F_3(\boldsymbol{p}, \kappa) - J_{33}(\boldsymbol{p})| = 0$ as a demonstration of the proof of Eq. (25). Note we have

$$\begin{aligned}
&\sup_{\boldsymbol{p}\in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)} |\nabla_{\bar\lambda} F_3(\boldsymbol{p}, \kappa) - J_{33}(\boldsymbol{p})| \\
\le &\sup_{\boldsymbol{p}\in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)} \left|2\mathbb{E}[\rho'(Q_1)P_z^2\rho''(P)] + 2\kappa\bar\lambda\mathbb{E}[\rho'(Q_1)P_z^2\rho'''(P)P_\lambda] - 2\mathbb{E}[\rho'(Q_1)\rho''(-Q_1)]\right| \\
\le &\sup_{\boldsymbol{p}\in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)} \left|2\mathbb{E}[\rho'(Q_1)(P_z^2\rho''(P) - \rho''(-Q_1))]\right| + \kappa \sup_{\boldsymbol{p}\in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)} \left|2\bar\lambda\mathbb{E}[\rho'(Q_1)P_z^2\rho'''(P)P_\lambda]\right| \\
\le &K\cdot \mathbb{E}\Big[\sup_{\boldsymbol{p}\in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)} |P_z^2\rho''(P) - \rho''(-Q_1)|\Big] + K\cdot\kappa
\end{aligned}$$

for some constant K that does not depend on κ and $\boldsymbol{p}$. Note we have $\sup_{\boldsymbol{p}\in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)} |P_z^2\rho''(P) - \rho''(-Q_1)|$ bounded and goes to 0 as $\kappa \to 0+$, for any fixed (Q_1, Q_2). Then by bounded convergence theorem, we have

$$\lim_{\kappa\to 0+}\sup_{\boldsymbol{p}\in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)} |\nabla_{\bar\lambda} F_3(\boldsymbol{p}, \kappa) - J_{33}(\boldsymbol{p})| = 0.$$

The proof of convergence of other entries within the matrix $\nabla_{\boldsymbol{p}}\boldsymbol{F}(\boldsymbol{p}, \kappa) - \boldsymbol{J}(\boldsymbol{p})$ follow from the same proof strategy using fact that $\sup_z |\rho^{(i)}(z)| < \infty$ for $i = 1, 2, 3, 4$.

Finally, we show Eq. (24). By Assumption A we have $\mathbb{E}[\rho''(Q_1)] > 0$, $\mathbb{E}[\rho''(Q_1)Q_1^2] > 0$, and $\bar\sigma_0 > 0$, which yields $\sigma_{\min}(\boldsymbol{J}(\boldsymbol{p}_0)) > 0$. This proves Eq. (24).

□

C.3.2. Proof of Lemma C.3

In this proof, we follow the same notations with the proof of Lemma C.2 as in Section C.3.2. Especially, the quantities Z, P, P_z, and P_λ are defined accordingly.

For any $(\bar\alpha, \bar\sigma, \bar\lambda, \kappa) \in \overline{\Omega} \times (0, \kappa_0)$ where $\overline{\Omega} = \{(\bar\alpha, \bar\sigma, \bar\lambda) : \bar\alpha > -1/\kappa, \bar\sigma > 0, \bar\lambda > 0\}$, we define

$$\begin{aligned}
f_1(\boldsymbol{p}, \kappa) &= \mathbb{E}[2\rho'(Q_1)\rho'(\mathsf{prox}_{\kappa\bar\lambda\rho}(-(1 + \kappa\bar\alpha)Q_1 + \sqrt{\kappa}\bar\sigma Z))^2], \\
f_2(\boldsymbol{p}, \kappa) &= \mathbb{E}[\rho'(Q_1)Q_1\rho'(\mathsf{prox}_{\kappa\bar\lambda\rho}(-(1 + \kappa\bar\alpha)Q_1 + \sqrt{\kappa}\bar\sigma Z))], \\
f_3(\boldsymbol{p}, \kappa) &= \mathbb{E}[2\rho'(Q_1)/[1 + \kappa\bar\lambda\rho''(\mathsf{prox}_{\kappa\bar\lambda\rho}(-(1 + \kappa\bar\alpha)Q_1 + \sqrt{\kappa}\bar\sigma Z))]].
\end{aligned} \tag{27}$$

By the definition of F_1, F_2, F_3 as in Eq. (20), we have

$$\begin{aligned}
F_1(\boldsymbol{p}, \kappa) &= \bar\sigma^2 - \bar\lambda^2 f_1(\boldsymbol{p}, \kappa), \\
F_2(\boldsymbol{p}, \kappa) &= \kappa^{-1} f_2(\boldsymbol{p}, \kappa), \\
F_3(\boldsymbol{p}, \kappa) &= \kappa^{-1}\{1 - \kappa - f_3(\boldsymbol{p}, \kappa)\}.
\end{aligned} \tag{28}$$

Then, Lemma C.3 holds as long as we show that there exists continuous functions $\boldsymbol{F}_0(\boldsymbol{p}) = (F_{0,1}(\boldsymbol{p}), F_{0,2}(\boldsymbol{p}), F_{0,3}(\boldsymbol{p}))$ and $\boldsymbol{g}(\boldsymbol{p}) = (g_1(\boldsymbol{p}), g_2(\boldsymbol{p}), g_3(\boldsymbol{p}))$ such that

$$f_1(\boldsymbol{p}, \kappa) = F_{0,1}(\boldsymbol{p}) + o(1), \tag{29}$$

$$\partial_\kappa f_1(\boldsymbol{p}, \kappa) = -(\bar\sigma^2/\bar\lambda^2)g_1(\boldsymbol{p}) + o(1), \tag{30}$$

$$f_2(\boldsymbol{p}, \kappa) = o(1), \tag{31}$$

$$\partial_\kappa f_2(\boldsymbol{p}, \kappa) = F_{0,2}(\boldsymbol{p}) + o(1), \tag{32}$$

$$\partial_\kappa^2 f_2(\boldsymbol{p}, \kappa) = g_2(\boldsymbol{p}) + o(1), \tag{33}$$

$$f_3(\boldsymbol{p}, \kappa) = 1 + o(1), \tag{34}$$

$$\partial_\kappa f_3(\boldsymbol{p}, \kappa) = -F_{0,3}(\boldsymbol{p}) + o(1), \tag{35}$$

$$\partial_\kappa^2 f_3(\boldsymbol{p}, \kappa) = -g_3(\boldsymbol{p}) + o(1), \tag{36}$$

where the $o(1)$ terms convergence to 0 uniformly over $\boldsymbol{p} \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)$ as $\kappa \to 0+$. Moreover, we need

$$F_{0,1}(\boldsymbol{p}_0) = \bar{\sigma}_0^2/\bar{\lambda}_0^2, \tag{37}$$

$$F_{0,2}(\boldsymbol{p}_0) = 0, \tag{38}$$

$$F_{0,3}(\boldsymbol{p}_0) = 1. \tag{39}$$

We first prove Eq. (29), (30) and (37). First, we have

$$\lim_{\kappa \to 0+} f_1(\boldsymbol{p}, \kappa) = \mathbb{E}[2\rho'(Q_1)\rho'(-Q_1)^2] = \bar{\sigma}_0^2/\bar{\lambda}_0^2,$$

where the last equality is by the definition in Eq. (22). Further, by smoothness of ρ and the prox operator, and the fact that the neighborhood $\mathsf{B}(\boldsymbol{p}_0, \varepsilon)$ is bounded, this convergence is uniform over $\boldsymbol{p} = (\bar{\alpha}, \bar{\sigma}, \bar{\lambda}) \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)$. This proves Eq. (29) and (37).

Second, we have

$$\begin{aligned} \partial_\kappa f_1(\boldsymbol{p}, \kappa) &= 4\mathbb{E}\Big[\rho'(Q_1)\rho'(P)\rho''(P)\Big(P_\lambda \bar{\lambda} - P_z \bar{\alpha} Q_1 + P_z \bar{\sigma} Z/(2\sqrt{\kappa})\Big)\Big] \\ &= 4\mathbb{E}\Big[\rho'(Q_1)\rho'(P)\rho''(P)\Big(P_\lambda \bar{\lambda} - P_z \bar{\alpha} Q_1\Big)\Big] \\ &\quad + 2\mathbb{E}\Big[\rho'(Q_1)[\rho''(P)^2 + \kappa\bar{\lambda}\rho''(P)^3 + \rho'(P)\rho'''(P)]P_z^3\Big]\bar{\sigma}^2, \end{aligned}$$

where the last inequality is by Stein's identity for $Z \sim \mathcal{N}(0, 1)$. So this gives

$$\begin{aligned} \lim_{\kappa \to 0+} \partial_\kappa f_1(\boldsymbol{p}, \kappa) &= 4\mathbb{E}\Big[\rho'(Q_1)\rho'(-Q_1)\rho''(-Q_1)\Big(- \rho'(-Q_1)\bar{\lambda} - \bar{\alpha} Q_1\Big)\Big] \\ &\quad + 2\mathbb{E}\Big[\rho'(Q_1)[\rho''(-Q_1)^2 + \rho'(-Q_1)\rho'''(-Q_1)]\Big]\bar{\sigma}^2. \end{aligned}$$

Again, by the smoothness of ρ and the prox operator, and the fact that the neighborhood $\mathsf{B}(\boldsymbol{p}_0, \varepsilon)$ is bounded, this convergence is uniform over $\boldsymbol{p} = (\bar{\alpha}, \bar{\sigma}, \bar{\lambda}) \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)$. This proves Eq. (30). The proof of other equations within (29) to (39) follow from similar continuity arguments. This proves Lemma C.3. □

D. Proofs of main theorems

This section provides the proofs of our main theorems, building on the developments in Section C. We first prove Theorem 3 in Section D.1 and prove Theorem 1 in Section D.2 as a direct corollary of Theorem 3. We prove Corollary 2 in Section D.3 and Corollary 4 in Section D.4.

D.1. Proof of Theorem 3

Recall from Corollary C.1 that, under Assumption A, in the limit of $d, n \to \infty$, $d/n \to \kappa$ where $\kappa < 0$, the calibration error $\Delta_p^{\mathsf{cal}}(\widehat{f})$ converges to the limit

$$C_{p,\kappa}(\sigma) := p - \mathbb{E}_{Z \sim \mathsf{N}(0,1)}\left[\sigma\left(\frac{\gamma}{R_\star} c_\star \cdot \sigma^{-1}(p) + \sqrt{1 - c_\star^2}\gamma \cdot Z\right)\right], \tag{40}$$

where $\gamma = \|\mathbf{w}_\star\|$. This is the first part of the claim in Theorem 3.

We now analyze the value $C_{p,\kappa}(\sigma)$ by plugging in the formulas for $R_\star, c_\star$ from Theorem C.1:

$$R_\star = \sqrt{\alpha_\star^2\gamma^2 + \kappa\sigma_\star^2} \quad \text{and} \quad c_\star = \frac{1}{\sqrt{1 + \kappa\sigma_\star^2/\alpha_\star^2\gamma^2}},$$

which yields

$$\frac{\gamma c_\star}{R_\star} = \gamma \cdot \frac{1}{\sqrt{1 + \kappa\sigma_\star^2/\alpha_\star^2\gamma^2}} \cdot \frac{1}{\sqrt{\alpha_\star^2\gamma^2 + \kappa\sigma_\star^2}} = \frac{1}{\alpha_\star + \kappa\sigma_\star^2/\alpha_\star\gamma^2},$$

and

$$\sqrt{1-c_\star^2}\cdot\gamma = \sqrt{1-\frac{1}{1+\kappa\sigma_\star^2/\alpha_\star^2\gamma^2}}\cdot\gamma = \sqrt{\frac{\kappa\sigma_\star^2}{\alpha_\star^2+\kappa\sigma_\star^2/\gamma^2}}.$$

Now, Lemma C.1 states that the following first order approximation holds for small κ:

$$\begin{aligned}\alpha_\star &= 1+\bar{\alpha}_0\kappa+o(\kappa),\\ \sigma_\star^2 &= \bar{\sigma}_0^2+o(1).\end{aligned}$$

where $o(1)$ denotes an error term that converges to 0 as $\kappa\to 0$ (and similarly, $o(\kappa)/\kappa\to 0$). Plugging these into the preceding displays, we get

$$\begin{aligned}\frac{\gamma c_\star}{R_\star} &= \frac{1}{1+\bar{\alpha}_0\kappa+o(\kappa)+\kappa(\bar{\sigma}_0^2+o(1))/(1+o(1))\gamma^2}\\ &= \frac{1}{1+(\bar{\alpha}_0+\bar{\sigma}_0^2/\gamma^2)\kappa+o(\kappa)}\\ &= \underbrace{1-(\bar{\alpha}_0+\bar{\sigma}_0^2/\gamma^2)\kappa+o(\kappa)}_{:=a(\kappa)},\end{aligned} \tag{41}$$

and

$$\sqrt{1-c_\star^2}\gamma = \sqrt{\frac{\kappa\bar{\sigma}_0^2+o(\kappa)}{1+\bar{\alpha}_0\kappa+o(\kappa)+\kappa(\bar{\sigma}_0^2+o(1))/\gamma^2}} = \underbrace{\sqrt{\kappa\bar{\sigma}_0^2+o(\kappa)}}_{:=b(\kappa)}. \tag{42}$$

We now use the following Lemma for the limiting calibration error $C_{p,\kappa}(\sigma)$, whose proof is deferred to Section D.1.1:

Lemma D.1. *Let $C_{p,\kappa}(\sigma)$ be defined as in (40), then we have*

$$\begin{aligned}\lim_{\kappa\to 0}\frac{C_{p,\kappa}(\sigma)}{\kappa} &= \lim_{\kappa\to 0}\frac{p-\mathbb{E}_{Z\sim\mathsf{N}(0,1)}\big[\sigma\big(a(\kappa)\cdot\sigma^{-1}(p)+b(\kappa)\cdot Z\big)\big]}{\kappa}\\ &= \underbrace{\sigma'(\sigma^{-1}(p))\cdot\sigma^{-1}(p)\cdot\big(\bar{\alpha}_0+\bar{\sigma}_0^2/\gamma^2\big)-\frac{1}{2}\sigma''(\sigma^{-1}(p))\cdot\bar{\sigma}_0^2}_{:=C(p,\bar{\alpha}_0,\bar{\sigma}_0^2,\sigma)}.\end{aligned}$$

By Lemma D.1, the limiting analysis of $C_{p,\kappa}(\sigma)$ reduces to the analysis of the constant $C(p,\bar{\alpha}_0,\bar{\sigma}_0^2,\sigma)$. First, observe that for $p\in(0.5,1)$, by Assumption A, we have $\sigma^{-1}(p)>0$, and (by the strict monotonicity) we have $\sigma'(\sigma^{-1}(p))>0$.

Proof of part (a) Suppose $\sigma''(\sigma^{-1}(p))\le 0$, then we have

$$C(p,\bar{\alpha}_0,\bar{\sigma}_0^2,\sigma)\ge\sigma'(\sigma^{-1}(p))\cdot\sigma^{-1}(p)\big(\bar{\alpha}_0+\bar{\sigma}_0^2/\gamma^2\big).$$

We now prove that $\bar{\alpha}_0+\bar{\sigma}_0^2/\gamma^2>0$ always holds under Assumption A, which implies that $C(p,\bar{\alpha}_0,\bar{\sigma}_0^2,\sigma)\ge 0$ and thus $C_{p,\kappa}(\sigma)>0$ for sufficiently small κ. Indeed, applying the expression (22a) and (22b) for $\bar{\alpha}_0$, $\bar{\sigma}_0$, we have

$$\bar{\alpha}_0+\bar{\sigma}_0^2/\gamma^2 = \underbrace{\frac{\mathbb{E}[\rho'(Q_1)\rho'(-Q_1)]}{(\mathbb{E}[\rho''(Q_1)])^2}}_{>0}\cdot\left\{-\frac{\mathbb{E}[Q_1\rho'''(Q_1)]}{2\mathbb{E}[Q_1^2\rho''(Q_1)]}+\frac{1}{\gamma^2}\right\}$$

(where $Q_1\sim\mathsf{N}(0,\gamma^2)$). Therefore, it suffices to show that the quantity inside the $\{\cdot\}$ is positive, which (by $\mathbb{E}[Q_1^2\rho''(Q_1)]>0$, as $\rho''(\cdot)=\sigma'(\cdot)>0$) is equivalent to

$$\gamma^2\mathbb{E}[Q_1\rho'''(Q_1)]<2\mathbb{E}\big[Q_1^2\rho''(Q_1)\big].$$

Let $Z \sim \mathsf{N}(0,1)$. The above is equivalent to

$$\gamma \mathbb{E}[Z\rho'''(\gamma Z)] < 2\mathbb{E}\big[Z^2\rho''(\gamma Z)\big]. \tag{43}$$

Applying Stein's identity $\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$ on $f(z) = z\rho''(\gamma z)$, we get $\mathbb{E}\big[Z^2\rho''(\gamma Z)\big] = \mathbb{E}[\rho''(\gamma Z) + \gamma Z\rho'''(\gamma Z)]$, which implies that $\gamma\mathbb{E}[Z\rho'''(\gamma Z)] = \mathbb{E}\big[(Z^2-1)\rho''(\gamma Z)\big]$. Therefore, (43) is further equivalent to

$$\mathbb{E}\big[(Z^2-1)\rho''(\gamma Z)\big] < \mathbb{E}\big[2Z^2\rho''(\gamma Z)\big],$$

which holds as we assumed $\rho''(\cdot) = \sigma'(\cdot) > 0$.

Proof of part (b) It is straightforward to see taht the following is a group a sufficient conditions for $C(p, \bar\alpha_0, \bar\sigma_0^2, \sigma) > 0$:

$$\bar\alpha_0 \le 0 \quad \text{and} \quad \frac{1}{2}\sigma''(\sigma^{-1}(p)) > \sigma'(\sigma^{-1}(p)) \cdot \sigma^{-1}(p)/\gamma^2.$$

Further, recall the expression (22a) for $\bar\alpha_0$ (where $Q_1 \sim \mathsf{N}(0,\gamma^2)$:

$$\bar\alpha_0 = -\frac{\mathbb{E}[Q_1\rho'''(Q_1)] \cdot \mathbb{E}[\rho'(Q_1)\rho'(-Q_1)]}{2\mathbb{E}[Q_1^2\rho''(Q_1)] \cdot \mathbb{E}[\rho''(Q_1)]^2}.$$

Because $\rho' = \sigma > 0$, $\rho'' = \sigma' > 0$, all expectations in the above expression are positive except for $\mathbb{E}[Q_1\rho'''(Q_1)]$. Therefore, $\bar\alpha_0 \le 0$ is equivalent to $\mathbb{E}[Q_1\rho'''(Q_1)] \ge 0$. This proves $\mathbb{E}[Q_1\rho'''(Q_1)] \ge 0$ and $\frac{1}{2}\sigma''(\sigma^{-1}(p)) > \sigma'(\sigma^{-1}(p)) \cdot \sigma^{-1}(p)/\gamma^2$ is a sufficient condition for $C(p, \bar\alpha_0, \bar\sigma_0^2, \sigma) < 0$, which by Lemma D.1 implies that $C_{p,\kappa}(\sigma) < 0$ for all small κ, yielding part (b). □

D.1.1. Proof of Lemma D.1

Let $a(\kappa) = 1 - (\bar\alpha_0 + \bar\sigma_0^2/\gamma^2)\kappa + e(\kappa)$ and $b(\kappa) = \sqrt{\kappa\bar\sigma_0^2 + f(\kappa)}$, where $e(\kappa), f(\kappa) = o(\kappa)$, i.e. they are bounded and satisfy $e(\kappa)/\kappa \to 0$ and $f(\kappa)/\kappa \to 0$ for small enough κ.

By a second-order Taylor expansion of σ at $\sigma^{-1}(p)$ (which exists as Assumption A assumed ρ is four-times continuously differentiable, i.e. σ is three-times continuously differentiable), we have

$$\begin{aligned}
&\frac{C_{p,\kappa}(\sigma)}{\kappa} = \frac{p - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\big[\sigma\big(a(\kappa)\cdot\sigma^{-1}(p) + b(\kappa)\cdot Z\big)\big]}{\kappa} \\
&= \frac{p - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\Big[\sigma\Big(\sigma^{-1}(p) - (\bar\alpha_0 + \bar\sigma_0^2/\gamma^2)\sigma^{-1}(p)\cdot\kappa + e(\kappa)\sigma^{-1}(p) + \sqrt{\kappa\bar\sigma_0^2 + f(\kappa)}\cdot Z\Big)\Big]}{\kappa} \\
&= \mathrm{IV}_\kappa + \mathrm{I}_\kappa + \mathrm{II}_\kappa + \mathrm{III}_\kappa,
\end{aligned}$$

where terms I, II, III, IV are

$$\begin{aligned}
\mathrm{IV}_\kappa &= \frac{p - \sigma(\sigma^{-1}(p))}{\kappa} = 0, \\
\mathrm{I}_\kappa &= -\frac{\sigma'(\sigma^{-1}(p))\cdot\mathbb{E}\Big[-(\bar\alpha_0 + \bar\sigma_0^2/\gamma^2)\sigma^{-1}(p)\cdot\kappa + e(\kappa)\sigma^{-1}(p) + \sqrt{\kappa\bar\sigma_0^2 + f(\kappa)}\cdot Z\Big]}{\kappa} \\
&= \sigma'(\sigma^{-1}(p))\sigma^{-1}(p)\cdot(\bar\alpha_0 + \bar\sigma_0^2/\gamma^2) + \sigma'(\sigma^{-1}(p))\sigma^{-1}(p)\cdot e(\kappa)/\kappa, \\
\mathrm{II}_\kappa &= -\frac{1}{2}\sigma''(\sigma^{-1}(p))\cdot\frac{\mathbb{E}\Big[\Big(-(\bar\alpha_0 + \bar\sigma_0^2/\gamma^2)\sigma^{-1}(p)\cdot\kappa + e(\kappa)\sigma^{-1}(p) + \sqrt{\kappa\bar\sigma_0^2 + f(\kappa)}\cdot Z\Big)^2\Big]}{\kappa}, \\
|\mathrm{III}_\kappa| &\le \mathbb{E}\left[\frac{1}{6}|\sigma'''(\xi)|\cdot\left|-(\bar\alpha_0 + \bar\sigma_0^2/\gamma^2)\sigma^{-1}(p)\cdot\kappa + e(\kappa)\sigma^{-1}(p) + \sqrt{\kappa\bar\sigma_0^2 + f(\kappa)}\cdot Z\right|^3\right]\Big/\kappa.
\end{aligned}$$

By these expressions, at the limit $\kappa \to 0$, we have

$$|\mathrm{III}_\kappa| \le \mathbb{E}\Big[C_1\big|C_2\kappa + C_3\sqrt{\kappa}Z\big|^3\Big]/\kappa \le C_4\kappa^{3/2}/\kappa = O(\sqrt{\kappa}) \to 0,$$

$$
\begin{aligned}
\mathrm{II}_\kappa &= -\frac{1}{2}\sigma''(\sigma^{-1}(p)) \cdot \mathbb{E}\Big[\big(-(\bar{\alpha}_0 + \bar{\sigma}_0^2/\gamma^2)\sigma^{-1}(p)\cdot\kappa + e(\kappa)\sigma^{-1}(p)\big)^2 + \big(\kappa\bar{\sigma}_0^2 + f(\kappa)\big)\cdot Z^2\Big] \Big/ \kappa \\
&\quad - \frac{1}{2}\sigma''(\sigma^{-1}(p)) \cdot \big[O(\kappa^2)/\kappa + (\bar{\sigma}_0^2 + f(\kappa)/\kappa)\big] \to -\frac{1}{2}\sigma''(\sigma^{-1}(p))\bar{\sigma}_0^2, \\
\mathrm{I}_\kappa &\to \sigma'(\sigma^{-1}(p))\sigma^{-1}(p)\cdot(\bar{\alpha}_0 + \bar{\sigma}_0^2/\gamma^2).
\end{aligned}
$$

Therefore we have

$$
\begin{aligned}
&\lim_{\kappa\to 0}\frac{C_{p,\kappa}(\sigma)}{\kappa} = \lim_{\kappa\to 0}\mathrm{I}_\kappa + \lim_{\kappa\to 0}\mathrm{II}_\kappa \\
&= \sigma'(\sigma^{-1}(p))\sigma^{-1}(p)\cdot(\bar{\alpha}_0 + \bar{\sigma}_0^2/\gamma^2) - \frac{1}{2}\sigma''(\sigma^{-1}(p))\bar{\sigma}_0^2.
\end{aligned}
$$

This is the desired result. □

D.2. Proof of Theorem 1

Theorem 1 is a special case of Theorem 3(a). Indeed, it is straightforward to check that the logistic activation $\sigma(t) = \frac{1}{1+e^{-t}}$ along with the logistic loss $\rho(t) = \log(1+e^t)$ satisfies Assumption A. Further, for any $t > 0$, we have

$$
\sigma''(t) = -\frac{e^t(e^t-1)}{(1+e^t)^3} < 0.
$$

Therefore the sufficient conditions in Theorem 3(a) is satisfied, from which we conclude that $\lim_{\kappa\to 0} C_{p,\kappa}/\kappa = C_p > 0$ for all $p \in (0.5, 1)$. □

D.3. Proof of Corollary 2

Observe that $\widehat{f}(\mathbf{x}) = \sigma(\widehat{\mathbf{w}}^\top \mathbf{x}) \overset{d}{=} \sigma(\|\widehat{\mathbf{w}}\|\, x_1)$ where $x_1 \sim \mathsf{N}(0,1)$. By definition of the CE and the closed-form expression for $\Delta_p^{\mathsf{cal}}(\widehat{f})$ in Lemma B.1, we have

$$
\begin{aligned}
&\mathrm{CE}(\widehat{f}) = \mathbb{E}_{\mathbf{x}}\left[\left|\Delta_p^{\mathsf{cal}}(\widehat{f})\Big|_{p=\widehat{f}(\mathbf{x})}\right|\right] = \int_{-\infty}^{\infty}\left|\Delta_p^{\mathsf{cal}}(\widehat{f})\Big|_{p=\sigma(\|\widehat{\mathbf{w}}\|x_1)}\right|\varphi(x_1)dx_1 \\
&= \int_{-\infty}^{\infty}\left|\sigma(\|\widehat{\mathbf{w}}\|\, x_1) - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\left[\sigma\left(\frac{\|\mathbf{w}_\star\|}{\|\widehat{\mathbf{w}}\|}\cos\widehat{\theta}\cdot\sigma^{-1}(\sigma(\|\widehat{\mathbf{w}}\|\, x_1)) + \sin\widehat{\theta}\,\|\mathbf{w}_\star\|\, Z\right)\right]\right|\varphi(x_1)dx_1 \\
&= \int_{-\infty}^{\infty}\left|\underbrace{\sigma(\|\widehat{\mathbf{w}}\|\, x_1) - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\left[\sigma\left(\|\mathbf{w}_\star\|\cos\widehat{\theta}\cdot x_1 + \sin\widehat{\theta}\,\|\mathbf{w}_\star\|\, Z\right)\right]}_{:=g(\|\widehat{\mathbf{w}}\|,\cos\widehat{\theta},x_1)}\right|\varphi(x_1)dx_1 \\
&= \int_{-\infty}^{\infty}\left|g(\|\widehat{\mathbf{w}}\|, \cos\widehat{\theta}, x_1)\right|\varphi(x_1)dx_1,
\end{aligned}
$$

where $\varphi(t) = \exp(-t^2/2)/\sqrt{2\pi}$ is the $\mathsf{N}(0,1)$ density.

We next show that

$$
g(R, c, x_1) = \sigma(Rx_1) - \mathbb{E}_{Z\sim\mathsf{N}(0,1)}\left[\sigma\left(\gamma c\cdot x_1 + \sqrt{1-c^2}\gamma\cdot Z\right)\right]
$$

is locally Lipschitz with respect to (R, c) around any (R_0, c_0) such that $c_0 < 1$. ($\gamma = \|\mathbf{w}_\star\|$ for notational simplicity.) We have $|g'_R| = |\sigma'(Rx_1)\cdot x_1| \le C_1|x_1|$, and

$$
|g'_C| = \left|\mathbb{E}_Z\left[\sigma'(\gamma c x_1 + \sqrt{1-c^2}\gamma Z)\cdot(\gamma x_1 - \frac{c}{\sqrt{1-c^2}}\gamma Z)\right]\right| \le C_2\left(|x_1| + \frac{1}{\sqrt{1-c_0^2}}\right).
$$

Therefore we indeed have, for (R, c) in a neighborhood of (R_0, c_0),

$$
|g(R_1, c_1, x_1) - g(R_2, c_2, x_1)| \le C_3\left(|x_1| + \frac{1}{\sqrt{1-c_0^2}}\right)\cdot(|R_1 - R_2| + |c_1 - c_2|).
$$

Above, $C_3 > 0$ is an absolute constant.

Now, Theorem C.1 shows that as $d, n \to \infty$, $d/n \to \kappa$, with probability one we have $(\|\widehat{\mathbf{w}}\|, \cos\widehat{\theta}) \to (R_\star, c_\star)$ where $c_\star < 1$. Therefore, we have

$$\begin{aligned} &\left|\mathrm{CE}(\widehat{f}) - \int_{-\infty}^{\infty} |g(R_\star, c_\star, x_1)|\varphi(x_1)dx_1\right| \\ &\le \int_{-\infty}^{\infty} C_3\left(|x_1| + \frac{1}{\sqrt{1-c_\star^2}}\right) \cdot (|R - R_\star| + |c - c_\star|)\varphi(x_1)dx_1 \to 0. \end{aligned}$$

This shows that, with probability one,

$$\mathrm{CE}(\widehat{f}) \to \int_{-\infty}^{\infty} |g(R_\star, c_\star, x_1)|\varphi(x_1)dx_1 = \int_{-\infty}^{\infty} \left|C_{\sigma(|R_\star x_1|),\kappa}\right|\varphi(x_1)\cdot dx_1 =: C_\kappa > 0.$$

where the last step used the limiting calibration error (8), and the fact that for $p < 0.5$ we have $|\Delta_p^{\mathsf{cal}}(\widehat{f})| = |\Delta_{1-p}^{\mathsf{cal}}(\widehat{f})|$ by symmetry of the activation function. This shows the first part of the corollary.

Finally, Theorem 1 asserts that $C_{p,\kappa} = C_p\kappa + o(\kappa)$ for sufficiently small κ. Further, this $o(\kappa)$ by the proof of Lemma D.1 is uniform over $p \in (0.5, 1)$, provided that $\sup_z\{|\sigma'(z)z|, |\sigma''(z)z^2|, |\sigma'''(z)\cdot\mathbb{E}_{G\sim\mathsf{N}(0,1)}[|z + aG|^3]|\} < \infty$. For the logistic activation $\sigma(z) = 1/(1+e^{-z})$ these are all satisfied due to the exponential decay of $\sigma', \sigma'', \sigma'''$. This means we can plug in $C_{p,\kappa} = C_p\kappa$ into the above and obtain

$$C_\kappa = \underbrace{\int_{-\infty}^{\infty} C_{\sigma(|R_\star x_1|)}\varphi(x_1)dx_1}_{:=C} \cdot\kappa + o(\kappa),$$

where $C > 0$. This shows the second part of the corollary. □

D.4. Proof of Corollary 4

We first check that $\sigma_{\mathrm{underconf}}$ defined in (13) satisfy the sufficient conditions of Theorem 3(b) at $\|\mathbf{w}_\star\| = 1$. Recall

$$\sigma_{\mathrm{underconf}}(z) = \begin{cases} 0, & z < -2\pi, \\ \dfrac{1}{2} + \dfrac{1}{4\pi}(z - \sin z), & |z| \le 2\pi, \\ 1, & z > 2\pi. \end{cases} \tag{44}$$

First, through numerical calculations, we get

$$\mathbb{E}_{Q_1\sim\mathsf{N}(0,1)}[Q_1\sigma''(Q_1)] = \mathbb{E}_{Q_1\sim\mathsf{N}(0,1)}\left[Q_1 \cdot \frac{1}{4\pi}\sin(Q_1)\right] = 0.0483 > 0.$$

This verifies condition (11). Second, for $z = \sigma^{-1}(p)$ and $|z| \le 2\pi$ we have

$$\sigma''_{\mathrm{underconf}}(z) - 2\sigma'_{\mathrm{underconf}}(z)\cdot z = \frac{1}{4\pi}(\sin z + 2z\cos z - 2z).$$

Numerical calculation shows that the above is strictly positive for (at least) $z \in (0, 0.96]$. This shows that the under-confidence provably happens at $p \in (0.5, \sigma(0.96)] = (0.5, 0.5112]$ in the limit of $\kappa \to 0$. This verified condition (12) for this range of p (and sufficiently small κ). We remark in passing that the upper range $p \approx 0.5112$ of under-confidence agrees well with the simulations at $d/n = 0.01$ in Figure 2(c).

Finally, note that $\sigma_{\mathrm{underconf}}$ satisfies all symmetry and low-order smoothness assumptions in Assumption A, but does not satisfy the high-order smoothness ($\sigma'' = \rho'''$ is still continuous, but $\sigma''' = \rho''''$ does not exist) as well as the strict positivity $\sigma'(z) > 0$. However, since $\sigma_{\mathrm{underconf}}$ satisfy the above two sufficient conditions with a positive margin, we can slightly perturb and smoothify it to some $\widetilde{\sigma}_{\mathrm{underconf}}$ which does satisfy higher-order smoothness and strict positivity, and preserves the above two sufficient conditions. This proves Corollary 4. □

E. Additional experimental details

E.1. Simulations

We choose $d = 100$ and $d/n \in \{0.01, 0.05, 0.1, 0.25\}$ which corresponds to $n \in \{10000, 2000, 1000, 400\}$. For all settings we generalize 5 problem instances. We solve the convex ERM (with either $\sigma_{\rm logistic}$ or $\sigma_{\rm underconf}$) with gradient descent. We run gradient descent with step size 0.01 on the full loss $\widehat{R}_n(\mathbf{w})$, until the gradient norm $\left\|\nabla \widehat{R}_n(\mathbf{w})\right\| < 10^{-5}$ (as $\widehat{R}_n$ is convex, gradient norm is a proper measure of global optimality). Each problem instance yields a $\widehat{\mathbf{w}}$ for which we can compute $\|\widehat{\mathbf{w}}\|$ and $\cos\widehat{\theta} = \widehat{\mathbf{w}}^\top \mathbf{w}_\star / \|\widehat{\mathbf{w}}\| \|\mathbf{w}_\star\|$, and then compute the calibration curve from the closed-form expression (6), as shown in Figure 2.

E.2. CIFAR10 experiment

We train a multi-class logistic (softmax) classifier $\widehat{\mathbf{W}} \in \mathbb{R}^{3072\times 5}$ for the (5-class subset of) CIFAR10 data. Note that the classifier does not involve an intercept (which does not restrict the capacity of the model as the data is class-balanced). We train on the training set (with $n = 25000$ examples) with the momentum optimizer with learning rate $\{10^{-3}, 10^{-4}, 10^{-5}\}$ for 30 epochs each (totally 90 epochs), and momentum 0.9. We use minibatch gradients with batch-size 128. The training and test images are pre-processed by subtracting the mean and normalizing by the standard deviation, and **without** any further data augmentation.

The trained $\widehat{\mathbf{W}}$ is then used as-is to generate (random) pseudo-labels $y_i^{\sf pseudo}$ as described in Section 5.2, for both the training and test sets. We then use the same training setup on the pseudo training set $(\mathbf{x}_i, y_i^{\sf pseudo})$ to learn a logistic classifier $\widetilde{\mathbf{W}} \in \mathbb{R}^{3072\times 5}$, and evaluate its accuracy and confidence on the test (pseudo-labeled) dataset.

The training and test losses for both the true labels and the pseudo labels are plotted in Figure 4. As indicated by the final training loss, the problem is under-parametrized (training loss is well above 0) for both tasks. Further, observe that both losses have stabilized at the end of training.

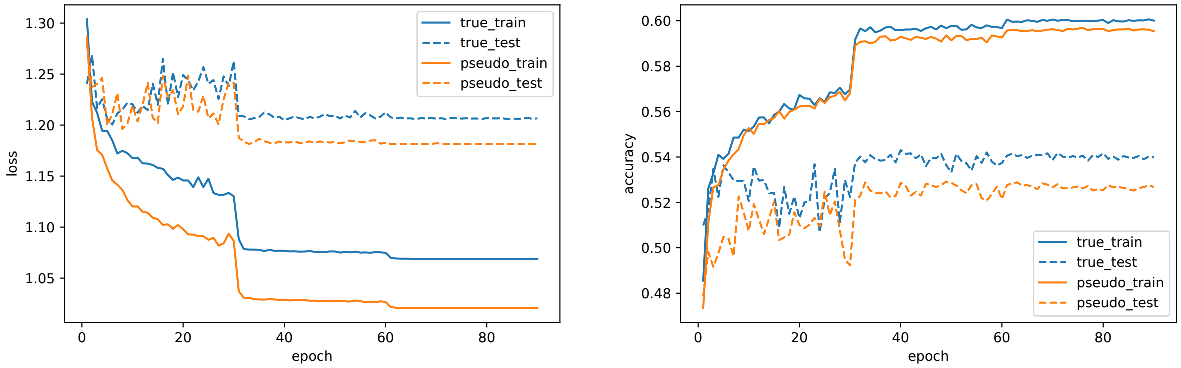


Figure 4. Training and test losses (left) and accuracies (right) for the CIFAR10 logistic regression experiment in Section 5.2.

For evaluating the confidence and accuracies on both the true labels and pseudo labels, we partition $[0.2, 10]$ to 10 equally spaced bins $B_1, \dots, B_{10}$, and compute its average confidence and accuracy within B_j on the test set as

$$\overline{\rm conf}_j = \frac{1}{|B_j|} \sum_{i:{\rm conf}_i \in B_j} {\rm conf}_i.$$

$$\overline{\rm acc}_j = \frac{1}{|B_j|} \sum_{i:{\rm conf}_i \in B_j} \mathbf{1}\left\{ \arg\max_k \left[\widetilde{\mathbf{W}}^\top \mathbf{x}_i\right]_k = y_i^{\sf pseudo} \right\},$$

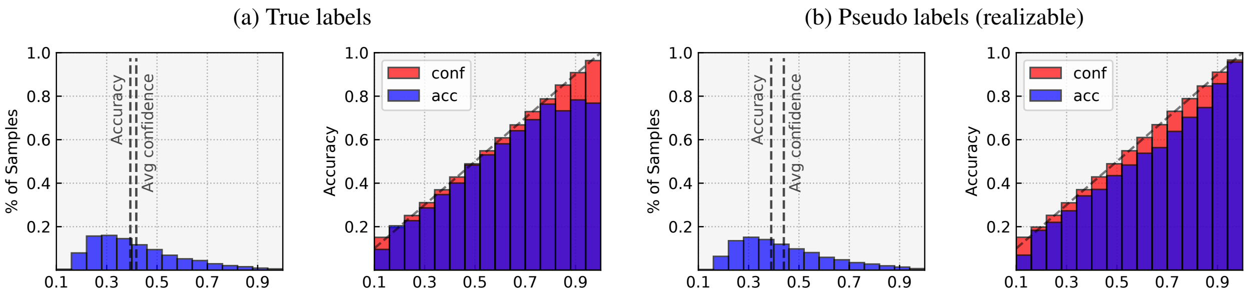
where

$${\rm conf}_i = \max_k \frac{\exp\left(\left[\widetilde{\mathbf{W}}^\top \mathbf{x}_i\right]_k\right)}{\sum_{k'=1}^5 \exp\left(\left[\widetilde{\mathbf{W}}^\top \mathbf{x}_i\right]_{k'}\right)} \in [0.2, 1]$$

is the confidence on the i-th test example for $i \in \{1, \dots, 5000\}$. The $\overline{\mathrm{conf}}_j$ and $\overline{\mathrm{acc}}_j$ are plotted in the calibration diagrams in Figure 3.

E.3. An ablation

Figure 5 reports the results of re-running the CIFAR experiment in Section 5.2, but with the full 10 classes of data (instead of 5), and 15 confidence bins (instead of 10). The test accuracy of logistic regression is now 40%. We find that logistic regression is over-confident on both true labels and pseudo-labels, similar as in Section 5.2.


(a) True labels

(b) Pseudo labels (realizable)

Figure 5. Calibration of multi-class logistic regression on (all 10 classes of) CIFAR10. Confidences are partitioned into $B = 15$ bins. The x-axes denote the confidences (predicted top probabilities) of the models.

E.4. Details of Figure 1

The deep neural net presented in Figure 1 is a WideResNet-50-2 (Zagoruyko & Komodakis, 2016) trained on the ImageNet image classification dataset. We train for 100 epochs with the momentum optimizer with Nesterov momentum 0.9, initial learning rate 0.1, a learning rate decay factor of $10\times$ at the $\{30, 60, 90\}$ epochs, and weight decay 5×10^{-4}. We use a batchsize of 256 distributed to 8 GPUs. The final test accuracy of this model is 76.27%. The realizable logistic regression uses simulated data with $n = 2000, d = 100$, and the same training protocol as described in Section E.1.