Optimal and Provable Calibration in High-Dimensional Binary Classification: Angular Calibration and Platt Scaling

Yufan Li
Department of Statistics
Harvard University
Cambrdige, MA 02138
yufan_li@g.harvard.edu

Pragya Sur
Department of Statistics
Harvard University
Cambrdige, MA 02138
pragya@fas.harvard.edu

Abstract

We study the fundamental problem of calibrating a linear binary classifier of the form $\sigma(\hat{w}^\top x)$, where the feature vector x is Gaussian, σ is a link function, and $\hat{w}$ is an estimator of the true linear weight $w^\star$. By interpolating with a noninformative *chance classifier*, we construct a well-calibrated predictor whose interpolation weight depends on the angle $\angle(\hat{w}, w_\star)$ between the estimator $\hat{w}$ and the true linear weight $w_\star$. We establish that this angular calibration approach is provably well-calibrated in a high-dimensional regime where the number of samples and features both diverge, at a comparable rate. The angle $\angle(\hat{w}, w_\star)$ can be consistently estimated. Furthermore, the resulting predictor is uniquely *Bregman-optimal*, minimizing the Bregman divergence to the true label distribution within a suitable class of calibrated predictors. Our work is the first to provide a calibration strategy that satisfies both calibration and optimality properties provably in high dimensions. Additionally, we identify conditions under which a classical Platt-scaling predictor converges to our Bregman-optimal calibrated solution. Thus, Platt-scaling also inherits these desirable properties provably in high dimensions.

1 Introduction

Calibration of predictive models is a fundamental problem in statistics and machine learning, especially in applications that require reliable uncertainty quantification. A well-calibrated model ensures that its predicted probabilities align closely with true event probabilities—a property essential in fields such as medical decision-making [7, 41], meteorological forecasting [14, 28, 23, 65, 64], self-driving systems [60], and natural language processing [66, 29].

Numerous algorithms have been proposed for calibrating the outputs of a trained model, including classical methods such as Platt scaling [70, 13, 69, 31], histogram binning [89, 77], isotonic regression [90, 40, 12, 36, 44], and more recent approaches such as temperature scaling [29, 47], ensemble-based methods [50, 58, 86, 81], and Bayesian strategies [45, 21], among others.

While extensive prior work has studied calibration [48, 77, 30, 74, 43], this literature primarily focuses on traditional asymptotic theories or finite-sample learning theoretic arguments. These approaches often overlook the impact of problem dimensionality, which is particularly relevant for high-dimensional settings where the number of features may be substantial. Alternatively, a separate line of research has explored calibration within a high-dimensional proportional asymptotic regime, where the sample size n and the feature dimension d both diverge, at a comparable rate. This proportional scaling regime has gained significant traction in modern statistics and machine learning. In statistics, its popularity stems from the fact that theories derived under this regime

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-5069

capture high-dimensional phenomena observed in moderate to large sized datasets unusually well [42, 6, 25, 80, 91, 5, 78, 39, 63, 84, 52, 53]. Consequently, this has spurred the creation of innovative methods displaying remarkable practical performance [61, 27, 9, 54, 76, 57]. In machine learning, this regime has proven exceptionally valuable and effective in analyzing the behavior of modern neural networks and other interpolation learners under overparametrization [55, 34, 56, 59, 2, 75, 67]. For binary classification in this proportional regime, a substantial line of work [79, 78, 92] establishes that classical logistic regression yields seriously biased estimates; building upon these, [3] shows that logistic regression tends to be inherently overconfident, while [22] discusses the impact of regularization under the same model. Finally, [21] introduces expectation consistency and derives a limiting calibration error formula as a function of the signal prior and other problem parameters.

Despite these advancements, an approach that is provably calibrated in high dimensions, without knowledge of the true signal prior, is missing. Moreover, there is a lack of principled understanding regarding optimal calibration strategies from among the available options. Additionally, rigorous guarantees on the performance of classical calibration methods, such as Platt scaling, in modern high-dimensional scenarios is notably absent from the literature. In this paper, we address these gaps. We consider the challenge of calibrating a binary linear predictor in a frequentist setting under a Gaussian design. Our contributions are three-fold: (i) we introduce a data-driven predictor that can provably calibrate in a broad class of high-dimensional binary classification problems; (ii) we show that our calibrated predictor is *Bregman-optimal*, meaning it uniquely minimizes any Bregman divergence relative to the true label-generation probability; (iii) we establish conditions under which a classical Platt-scaled predictor converges to this Bregman-optimal calibrated solution, thereby formally showing that Platt scaling is both well-calibrated and Bregman optimal in our high-dimensional setting. Although we derive our theoretical results assuming Gaussian features, extensive recent universality results suggest that these should continue to hold for sufficiently light tailed distributions (see Section 8 for a discussion). We provide experiments that demonstrate this robustness to the Gaussian assumption (Section H.2 and 4).

We construct our calibrated predictor by interpolating with an uninformative ("chance") predictor, where the interpolation weight is determined by the angle $\angle(\hat{w}, w_\star)$ between the estimated linear weight $\hat{w}$ and the true weight $w_\star$. Our construction crucially leverages recent developments from the literature on observable estimation of unknown parameters in high dimensions. For instance, leveraging advances in [38, 8, 9, 10, 18, 54], we can show that the angle $\angle(\hat{w}, w_\star)$ is consistently estimable when n and d grow proportionally. To our knowledge, this is the first provable calibration method in a high-dimensional setting, and it uncovers a conceptual link between optimal calibration and $\angle(\hat{w}, w_\star)$: the poorer the alignment of $\widehat{w}$ with $w_\star$, the greater the noise needed to be injected to prediction logits to ensure calibration.

2 Setting

Suppose we observe i.i.d. data (y_i, x_i) satisfying

$$y_i \overset{\text{iid}}{\sim} \text{Bern}\left(\sigma\left(w_\star^\top x_i\right)\right), \quad i = 1, \ldots, n, \tag{1}$$

where $\sigma : \mathbb{R} \to [0, 1]$ denotes the link function and the covariates $x_i \in \mathbb{R}^d$ are drawn independently as $x_i \overset{\text{iid}}{\sim} N(0, \Sigma)$, with Σ assumed to be known (say from a separate unlabeled dataset as in [17, 18]). The true linear weight $w_\star \in \mathbb{R}^d$ is an arbitrary deterministic vector, and we assume without loss of generality that $w_\star^\top \Sigma w_\star = \|w_\star\|_\Sigma^2 = 1$. The training dataset is denoted as $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$ and $y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$.

To quantify the degree of miscalibration, we define the calibration error at level p for any predictor $\hat{f}$ as

$$\Delta_p^{\text{cal}}(\hat{f}) = p - \mathbb{E}_{x_{\text{new}}}\left[\sigma\left(w_\star^\top x_{\text{new}}\right) \mid \hat{f}\left(x_{\text{new}}\right) = p\right],$$

where $\mathbb{E}_{x_{\text{new}}}$ denotes the expectation over $x_{\text{new}} \sim N(0, \Sigma)$. A predictor is said to be *well-calibrated* if $\Delta_p^{\text{cal}}(\hat{f}) = 0$ for all p in the range of $\hat{f}$. Intuitively, this means that when the predictor assigns a probability p to label 1, the true probability of label 1 is indeed p.

We consider the regularized M-estimator

$$\widehat{w} = \arg\min_{w} \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}\left(w^\top x_i\right) + g(w),$$

where $g(\cdot)$ is a convex penalty and $\ell(\cdot)$ is a convex loss function. In this setting, we consider a sequence of problem instances $\{y(d), X(d), w_\star(d)\}_{d\geq 1}$ such that $X(d) \in \mathbb{R}^{n(d)\times d}$ and $y(d) \in \mathbb{R}^{n(d)}$ generated from (1). It is well-known that in the special case where $\ell(\cdot)$ equals the logistic loss and $g(\cdot)$ is zero, the corresponding predictor $\sigma(\hat{w}^\top x_{\rm new})$ is grossly mis-calibrated in the high-dimensional regime $\frac{n}{d} \to (0, +\infty)$ [79, 78, 3, 22], even where it is well-defined and unique [16].

In what follows, we present, for the first time, a predictor that is provably well-calibrated in this regime (for general convex losses and penalties beyond the special case mentioned above). We achieve this through an angular calibration idea, and furthermore, establish that this is optimal in the sense that it minimizes any Bregman divergence to the true label distribution. We conclude showing an interesting connection—Platt scaling converges to our angular predictor—and therefore is both provably well-calibrated and optimal in the aforementioned sense.

3 Introducing angular calibration

Most calibration strategies *adjust* a pre-trained predictor by learning a mapping $F\colon u \mapsto F(u)$ of the logits $\hat{w}^\top x_{\rm new}$. Platt scaling, for example, stipulates the parametric form $F(u) = \sigma(Au + B)$, where $A, B \in \mathbb{R}$ are fit on a holdout dataset. This raises a natural question:

Among all well-calibrated predictors of the form $F(\hat{w}^\top x_{\rm new})$, which one is "the best"?

In this section, we introduce a predictor with such an optimality property by interpolating between the prediction logits $\hat{w}^\top x$ and an uninformative chance predictor. Specifically, we show that if the interpolation weight is determined by the angle between the estimator $\widehat{w}$ and the true weight $w_\star$, given by

$$\theta_* = \arccos\left(\frac{\langle w_\star, \widehat{w}\rangle_\Sigma}{\|\widehat{w}\|_\Sigma \|w_\star\|_\Sigma}\right), \tag{2}$$

then the resulting interpolated predictor minimizes any Bregman divergence to the true label distribution among all predictors of the form $F(\hat{w}^\top x_{\rm new})$. To the best of our knowledge, a predictor that is both provably calibrated and optimal (in the aforementioned sense) has not been previously introduced for high-dimensional problems.

Notably, our predictor uses the angle defined in (2), thus to define a data-driven predictor, we require a consistent estimate of this angle. Fortunately, recent advances in the high-dimensional literature (c.f., [38, 8, 9, 10, 18, 54]) allow us to estimate the inner product $\langle \widehat{w}, w_\star\rangle_\Sigma$, and therefore θ_*, when n and d grow proportionally. We discuss the details of this estimation scheme later in Section 6. For now, we present our angular calibration idea assuming that we have access to a consistent estimator $\hat{\theta}$ for θ_*.

Definition 3.1. (Angular Predictor) Let

$$\hat{f}_{\rm ang}\left(\widehat{w}^\top x_{\rm new}; \hat{\theta}\right) = \mathbb{E}_Z\left(\sigma\left(\cos\left(\hat{\theta}\right) \cdot \left(\frac{\widehat{w}^\top x_{\rm new}}{\|\widehat{w}\|_\Sigma}\right) + \sin\left(\hat{\theta}\right) \cdot Z\right)\right) \tag{3}$$

where $\hat{\theta}$ is a consistent estimator of θ_* defined as in (13) and $\mathbb{E}_Z$ denotes expectation with respect to the *Gaussian noise* $Z \sim N(0,1)$. We will later refer to $\hat{f}_{\rm ang}$ as the angular predictor for simplicity.

Theorem 3.2 below shows that the angular predictor is well-calibrated. We defer the proof to Section A

Theorem 3.2. *Assume the link function σ is continuous. Then, the predictor $\hat{f}_{\rm ang}$ defined in* (3) *is well-calibrated as $d, n \to \infty, n/d \to (0, \infty)$. That is, for any p contained in the range of σ, we have that*

$$\Delta_p^{\rm cal}\left(\hat{f}_{\rm ang}\left(\cdot; \hat{\theta}\right)\right) = p - \mathbb{E}_{x_{\rm new}}\left[\sigma\left(w_\star^\top x_{\rm new}\right) \mid \hat{f}_{\rm ang}\left(\widehat{w}^\top x_{\rm new}; \hat{\theta}\right) = p\right] \to 0,$$

in probability where $\hat{\theta}$ is a consistent estimator for $\theta_\star$ (Cf. Proposition 6.2).

https://doi.org/10.52202/085713-5069

The above utilizes the result that when $\hat{\theta} = \theta_*$ exactly, $\hat{f}_{\text{ang}}(\cdot; \theta_*)$ is exactly well-calibrated (we state and prove this formally in Theorem A.1) and that $\hat{\theta}$ is consistent for $\theta_\star$. We will later show that $\hat{f}_{\text{ang}}\left(\cdot; \hat{\theta}\right) \approx \hat{f}_{\text{ang}}(\cdot; \theta_*)$ is in fact optimal in the sense that it minimizes any Bregman divergence to the true label distribution. The construction (3) admits an intuitive interpretation. By the basic trigonometric identity $\cos^2(\theta_*) + \sin^2(\theta_*) = 1$, we can see that the logits, that is, the argument of $\sigma(\cdot)$ in (3), is an interpolation between the informative component $\widehat{w}^\top x_{\text{new}}$ and the noninformative Gaussian noise Z. Notice that when $\widehat{w}$ is well aligned with $w_\star$ (i.e., $\cos^2(\theta_*) = 1$), the angular predictor $\hat{f}_{\text{ang}}$ lies closer to the informative predictor $\sigma(\widehat{w}^\top x_{\text{new}})$. Conversely, when $w_\star$ and $\widehat{w}$ are orthogonal (i.e., $\sin^2(\theta_*) = 1$), $\hat{f}_{\text{ang}}$ defaults to the non-informative chance predictor $\mathbb{E}[\sigma(Z)] = \mathbb{E}[\sigma(w_\star^\top x_{\text{new}})]$. In other words,

The poorer the alignment between $w_\star$ and $\widehat{w}$, the greater the magnitude of noise Z required to maintain calibration.

We will show in the next section that this angular interpolation idea leads to a uniquely Bregman optimal calibrated predictor. This provides the first calibration procedure that is calibrated and optimal in high dimensions, provably. Peusdocode for angular calibration, using angle estimator from Section 6, is included in Section G.

4 Main Result I: Calibrating Optimally using Angular Calibration

Before formally stating our results on optimality of angular calibration, we first define the following random probability vectors for label distribution,

$$q_\star := \begin{pmatrix} \sigma(w_\star^\top x_{\text{new}}) \\ 1 - \sigma(w_\star^\top x_{\text{new}}) \end{pmatrix}, \quad \hat{q}_F := \begin{pmatrix} F(\widehat{w}^\top x_{\text{new}}) \\ 1 - F(\widehat{w}^\top x_{\text{new}}) \end{pmatrix}, \quad \hat{q}_{\text{ang}}(\hat{\theta}) := \begin{pmatrix} \hat{f}_{\text{ang}}(\widehat{w}^\top x_{\text{new}}; \hat{\theta}) \\ 1 - \hat{f}_{\text{ang}}(\widehat{w}^\top x_{\text{new}}; \hat{\theta}) \end{pmatrix} \tag{4}$$

where $F : \mathbb{R} \to [0, 1]$ is any measurable function. Here, $q_\star$ corresponds to the ground-truth probability distribution of the new label, $\hat{q}_F$ the prediction probability distribution of an F-calibrated predictor, and q_{ang} the prediction probability distribution of our angular predictor.

Next, we define the Bregman loss function.

Definition 4.1 (Bregman Loss Functions)**.** Let $\phi : \mathbb{R}^2 \mapsto \mathbb{R}$ be a strictly convex differentiable function. Then, the Bregman loss function $D_\phi : \mathbb{R}^2 \times \mathbb{R}^2 \mapsto \mathbb{R}$ is defined as

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle.$$

The Bregman loss function class covers common losses such as the squared loss $D_\phi(x, y) := \|x - y\|_2^2$ and Kullback-Liebler (KL) divergence $D_\phi(x, y) = \sum_{j=1}^2 x_j \log(x_j/y_j)$ between two probability vectors x, y.

Theorem 4.2 states that the prediction probability $\hat{q}_{\text{ang}}$ generated by the angular predictor uniquely minimizes *any* Bregman loss against the ground-truth probability vector $\hat{q}_\star$ within the class of q_F for any F. We defer the proof of the Theorem below to Section B.

Theorem 4.2 (Optimality of angular predictor)**.** *Let $\phi : \mathbb{R}^2 \mapsto \mathbb{R}$ be any strictly convex differentiable function, and let D_ϕ be the corresponding Bregman loss function. Let $\mathbb{E}_{x_{\text{new}}}[\phi(q_\star)]$ be finite. Then, the expected Bregman loss $\mathbb{E}_{x_{\text{new}}}[D_\phi(q_\star, \hat{q}_F)]$ admits a unique minimizer (up to a.s. equivalence) among all $q_F, \forall F \in \mathcal{F} := \{f : \mathbb{R} \to [0, 1]\}$. Let this minimizer be $F_\star = \arg\min_{F \in \mathcal{F}} \mathbb{E}_{x_{\text{new}}}[D_\phi(q_\star, \hat{q}_F)]$. Further suppose that the link function σ is continuous. We then have that as $n, d \to \infty$, we have*

$$\left\| \hat{q}_{\text{ang}}(\hat{\theta}) - \hat{q}_{F_\star}(\widehat{w}^\top x_{\text{new}}) \right\|_2^2 \to 0$$

in probability. That is, the label prediction probability vector from angular calibration converges to the optimal label prediction probability vector given by $F_\star$.

We note that as $\hat{\theta} = \theta_*$, $\hat{q}_{\text{ang}}(\hat{\theta})$ precisely attains the optimal solution $F_*(\widehat{w}^\top x_{\text{new}})$. We defer the technical statement to Theorem B.2 in Section B.

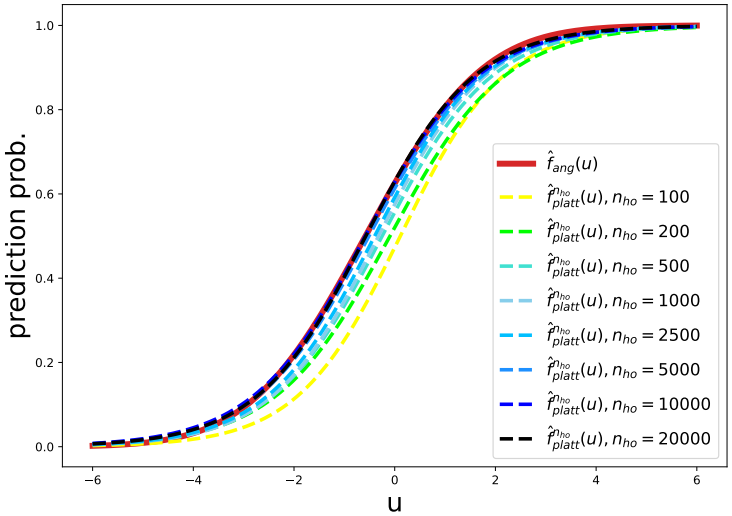


Figure 1: Platt scaling of a logistic ridge predictor converges to angular calibration predictor, as holdout set size increases. The plot is generated with Gaussian data with covariance $\Sigma = \frac{1}{d}\bar{\Sigma}$ where $\bar{\Sigma}_{kl} = 0.5^{|k-l|}, \forall k, l \in \{1, ..., d\}$, sigmoid link function in a data deficient setting where $n = 1000, p = 2000$. See more details in Section 7

5 Main Result II: Platt scaling is provably calibrated and Bregman-optimal

Platt scaling is arguably the most widely used calibration method in modern machine learning, yet its theoretical properties in high-dimensional settings remain unexplored. In this section, we identify conditions under which Platt scaling converges to our angular predictor, and is therefore well-calibrated and Bregman-optimal in high dimensions.

Platt scaling finds a mapping F of the prediction logits $\widehat{w}^\top x_{\text{new}}$ by minimizing the log-likelihood on a holdout dataset. In this section, we specifically consider the setting where we have a holdout dataset $(x_{\text{ho},i}, y_{\text{ho},i})_{i=1}^{n_{\text{ho}}}$ and the negative log-likelihood

$$\hat{\ell}_{n_{\text{ho}}}(F) := \sum_{i=1}^{n_{\text{ho}}} -y_{\text{ho},i} \log\big(F(\widehat{w}^\top x_{\text{ho},i})\big) - (1 - y_{\text{ho},i}) \log\big(1 - F(\widehat{w}^\top x_{\text{ho},i})\big). \tag{5}$$

The Platt calibration procedure then searches for a mapping F within some hypothesis class $\mathcal{F}_{\text{platt}}$ that minimizes the negative log-likelihood. Elementary asymptotic theory then shows that as $n_{\text{ho}} \to \infty$ (i.e. the holdout set is sufficiently large), $\hat{L}(\theta)$ in (5) converges to the population loss

$$\ell^\star(F) = \mathbb{E}_{x_{\text{new}}}\left[D_{\text{KL}}\left(\begin{pmatrix} \sigma(w_\star^\top x_{\text{new}}) \\ 1 - \sigma(w_\star^\top x_{\text{new}}) \end{pmatrix} \middle\| \begin{pmatrix} F(\widehat{w}^\top x_{\text{new}}) \\ 1 - F(\widehat{w}^\top x_{\text{new}}) \end{pmatrix}\right)\right] \tag{6}$$

almost surely up to a constant. This is exactly the argument of the Bregman loss $\mathbb{E}_{x_{\text{new}}}\left[D_\phi(q_\star, \hat{q}_F)\right]$ from Theorem 4.2 for ϕ specialized as the negative Shannon entropy. That is, such calibration procedures are essentially trying to optimizing the Bregman loss but within the restricted hypothesis class.

This naturally raises the question of whether calibration procedures such as Platt scaling can achieve the optimal Bregman loss, say in the limit of a sufficiently large holdout set. Theorem 5.1 shows that, if σ is a probit link function (or is closely approximated by one up to an affine transformation—for instance, $\text{sigmoid}(x) \approx \Phi(\sqrt{\pi/8}\, x)$), and if the negative log-likelihood (5) is minimized over the hypothesis class of the form $\sigma(Au + B)$, $A, B \in \mathbb{R}$ then the resulting Platt-scaled predictor converges to our angular predictor as $n_{\text{ho}} \to \infty$. Combining this connection with our results for the predictor $\hat{f}_{\text{ang}}(u; \theta_*)$ which is our angular predictor $\hat{f}_{\text{ang}}(u; \hat{\theta})$ with $\hat{\theta} = \theta_*$ exactly. As mentioned previously (see also Theorem A.1 and Theorem B.2 in Appendix), $\hat{f}_{\text{ang}}(u; \theta_*)$ is exactly calibrated and Bregman-optimal, which shows that Platt scaling is both provably calibrated and Bregman optimal. This offers

https://doi.org/10.52202/085713-5069

the first formal high-dimensional guarantees of this kind for the widely well-known Platt scaling procedure. We defer the proof of the Theorem below to Section C.

Theorem 5.1. *Consider the predictor $\hat{f}^{n_{\rm ho}}_{\rm platt}(u)$ calibrated by the Platt scaling procedure, that is,*

$$\begin{aligned} &\hat{f}^{n_{\rm ho}}_{\rm platt}(\widehat{w}^\top x_{\rm new}) = \sigma(\hat{A}^{n_{\rm ho}} \cdot \widehat{w}^\top x_{\rm new} + \hat{B}^{n_{\rm ho}}), \\ &\textit{with } \hat{A}^{n_{\rm ho}}, \hat{B}^{n_{\rm ho}} = \operatorname*{argmin}_{(A,B)\in\mathcal{H}} \hat{\ell}_{n_{\rm ho}}(u \mapsto \sigma(Au+B)) \end{aligned} \tag{7}$$

for $\hat{\ell}_{n_{\rm ho}}(\cdot)$ defined in (5)*. If the link function σ satisfies $\sigma(x) = \Phi(a{\cdot}x{+}b)$ for some $a \in \mathbb{R}\backslash\{0\}, b \in \mathbb{R}$ and the point (A_*, B_*) defined in* (9) *is contained in a compact subset $\mathcal{H} \subset \mathbb{R}^2$, the angular predictor defined in* (3) *satisfies*

$$\hat{f}_{\rm ang}(u;\theta_*) = \sigma(A_* \cdot u + B_*) \in \mathcal{F}_{\rm platt}, \tag{8}$$

where

$$A_* = \frac{\cos(\theta_*)}{\|\hat{w}\|_\Sigma \sqrt{1+a^2\sin^2(\theta_*)}}, \qquad B_* = \frac{b}{a}\left(\frac{1}{\sqrt{1+a^2\sin^2(\theta_*)}} - 1\right) \tag{9}$$

and $\mathcal{F}_{\rm platt} = \{u \mapsto \sigma(Au+B) : A, B \in \mathbb{R}\}$. Moreover, as $n_{\rm ho} \to \infty$, we have that $\hat{A}^{n_{\rm ho}} \to A_*, \hat{B}^{n_{\rm ho}} \to B_*$ in probability and

$$\sup_{u\in\mathbb{R}} \left| \hat{f}^{n_{\rm ho}}_{\rm platt}(u) - \hat{f}_{\rm ang}(u;\theta_*) \right| \to 0 \tag{10}$$

in probability. Here, in-probability convergence is with respect to the randomness of $\{(x_{{\rm ho},i}, y_{{\rm ho},i})_{i=1}^{n_{\rm ho}}\}$.

To be clear, the above theorem considers the asymptotics in the holdout set $n_{\rm ho}$ for a fixed sample size and dimension n, d of the training dataset. We illustrate Theorem 5.1 in Figure 1 (see Section 7 for detailed settings) where the solid red line plots our angular predictor $u \mapsto \hat{f}_{\rm ang}(u)$ defined in (3) and the dashed lines plot the predictor $u \mapsto \hat{f}^{n_{\rm ho}}_{\rm platt}(u)$ calibrated by Platt scaling on increasingly large holdout sets $n_{\rm ho}$. We observe that the Platt scaling predictors indeed converge to our angular predictor as the holdout set sizes $n_{\rm ho}$ increase.

6 Consistent angle estimation

Observe that the angular predictor $\hat{f}_{\rm ang}$ depends on the unobserved quantity $\langle w_\star, \widehat{w}\rangle_\Sigma$. Using recent advancements from Equation (11) we are able to provide a consistent estimator for this quantity. For simplicity, we outline the estimation procedure here for a twice-differentiable loss function ℓ and strongly convex, twice-differentiable penalty g; analogous results for unregularized M-estimation and other losses/penalties found in [8]. We note that this result is part of a long line of development in observable estimation of unknown quantities in high dimensions [38, 9, 10, 18, 54].

A data driven estimator for $\langle w_\star, \widehat{w}\rangle_\Sigma^2$ proposed in [8] is:

$$\hat{a}_*^2 = \frac{\left(\frac{\hat{\upsilon}}{n}\|X\widehat{w} - \hat{\gamma}\hat{\psi}\|^2 + \frac{1}{n}\hat{\psi}^\top X\widehat{w} - \hat{\gamma}\hat{r}^2\right)^2}{\frac{1}{n^2}\left\|\Sigma^{-\frac{1}{2}}X^\top\hat{\psi}\right\|^2 + \frac{2\hat{\upsilon}}{n}\hat{\psi}^\top X\widehat{w} + \frac{\hat{\upsilon}^2}{n}\|X\widehat{w} - \hat{\gamma}\hat{\psi}\|^2 - \frac{d}{n}\hat{r}^2} \tag{11}$$

where $\hat{\psi} \in \mathbb{R}^n$ is the vector with components $\hat{\psi}_i = -\ell_i'\left(x_i^\top\widehat{w}\right)$, $\hat{\upsilon} = \frac{1}{n}\operatorname{Tr}\left(D - DX\hat{H}X^T D\right)$, $\hat{\gamma} = \operatorname{Tr}\left(X\hat{H}X^\top D\right)$ for $D = \operatorname{diag}\left(\ell_y''(X\widehat{w})\right)$ and $\hat{H} = \left(X^\top DX + n\nabla^2 g(\widehat{w})\right)^{-1}$ and $\hat{r} = (\frac{\|\hat{\psi}\|^2}{n})^{1/2}$. It can be shown that $\left|\hat{a}_*^2 - \langle w_\star, \widehat{w}\rangle_\Sigma^2\right| \to 0$ in suitable high-dimensional sense. We refer the technical statement to Theorem D.1 in Section D.

To estimate $\langle w_\star, \widehat{w}\rangle_\Sigma$, we also need to estimate its sign. We require reserving a constant fraction $n_{\rm ho} = \alpha \cdot n$ of the n training data for the sign estimator,

$$\widehat{\rm sgn} := \operatorname{sign}\left(\sum_{i=1}^{n_{\rm ho}} \widehat{w}^\top x_i^{\rm ho} \cdot y_i^{\rm ho}\right). \tag{12}$$

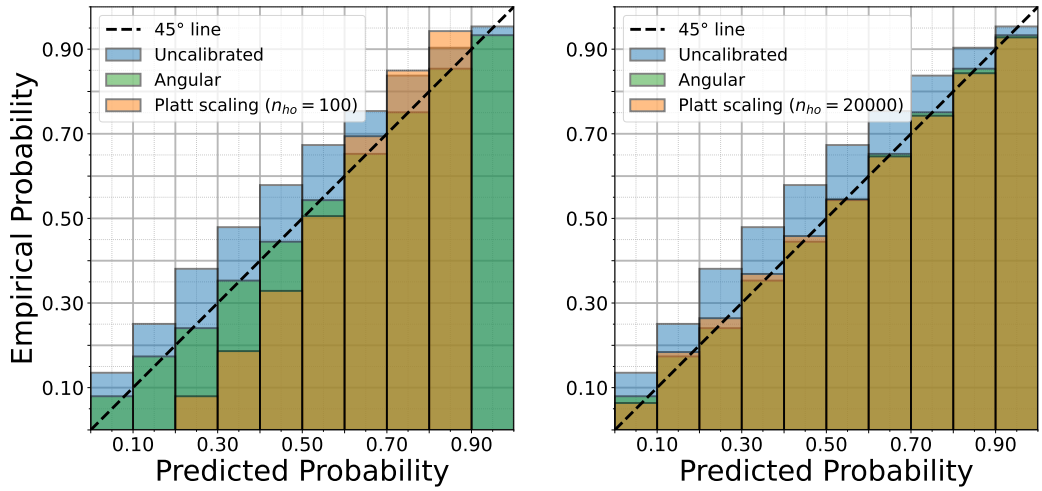


Figure 2: Reliability plots for angular calibration and Platt scaling of a logistic ridge predictor. Left panel uses a small holdout set for Platt scaling with $n_{\text{ho}} = 100$; Right panel uses a large holdout set with $n_{\text{ho}} = 2000$. The plot is generated with Gaussian data with covariance $\Sigma = \frac{1}{d}\bar{\Sigma}$ where $\bar{\Sigma}_{kl} = 0.5^{|k-l|}, \forall k, l \in \{1, ..., d\}$, sigmoid link function in a data deficient setting where $n = 1000, p = 2000$. See Section 7 for more details.

It can be shown that the probability of wrong sign identification using $\widehat{\text{sgn}}$ decreases exponentially with n_{ho}. We defer the proof to Section E.

Proposition 6.1. *Suppose that $\sigma'(x)$ is well-defined and non-negative almost everywhere and $\sigma'(x) > 0$ on a set with non-zero Lebesgue measure. We then have for some absolute constant $c > 0$,*

$$\mathbb{P}_{\text{ho}}\left(\widehat{\text{sgn}} = \text{sign}\left(\langle w_\star, \widehat{w}\rangle_\Sigma\right)\right) \geq 1 - 2\exp\left(-cn_{\text{ho}}\left(\cos\left(\theta_*\right) \cdot \mathbb{E}\sigma'(Z)\right)^2\right)$$

where $Z \sim N(0,1)$ and $\mathbb{P}_{\text{ho}}$ is with respect to the randomness in $\left(y_i^{\text{ho}}, x_i^{\text{ho}}\right)_{i=1}^{n_{\text{ho}}}$.

Plugging (11) and (12) into (2), we have the following estimator $\hat{\theta}$ for θ_* in (2)

$$\hat{\theta} := \arccos\left(\|\widehat{w}\|_\Sigma^{-1}\widehat{\text{sgn}} \cdot \sqrt{\hat{a}_*^2}\right). \tag{13}$$

Theorem 6.2 below shows that $\hat{\theta}$ is consistent.

Corollary 6.2. *Under the assumption of Proposition 6.1 and Theorem D.1, as $n, d \to \infty$, we have that $|\hat{\theta} - \theta_*| \to 0$ in probability where $n_{\text{ho}} = \alpha \cdot n$ for a fixed constant $\alpha > 0$.*

7 Numerical experiments

7.1 Simulations

This section presents a simple simulation to demonstrate results in Section 6. We generate i.i.d. samples $x_i \overset{\text{iid}}{\sim} N(0, \Sigma), i = 1, ..., n$ where $\Sigma = \frac{1}{d}\bar{\Sigma}$ and $\bar{\Sigma}_{kl} = 0.5^{|k-l|}, \forall k, l \in \{1, ..., d\}$; we also generate labels from (1) with $\sigma(u) = \text{sigmoid}(3u + 1) = 1/(1 + \exp(-(3u + 1)))$ and $w_\star \sim N(0, I_d)$ (normalized to $\|w_\star\|_\Sigma = 1$). We consider the case of ridge logistic regression with $\ell_{y_i}(w^\top x) = -y_i \log(\hat{p}_w(x_i)) - (1 - y_i)\log(1 - \hat{p}_w(x_i))$ with $\hat{p}_w(x_i) = \frac{1}{1+\exp(-x_i w)}$ and $g(w) = \frac{\lambda}{2d}\|w\|_2^2$ with $\lambda = 0.5$. We assume that we are in a data deficient setting where $n = 1000, p = 2000$.

The realizability plots in Figure 2 are generated from a test set of size $n_{\text{test}} = 20000$. To produce these plots, we bin the predicted probabilities for label 1 (on the x-axis) and then compute the average of the observed label within each bin (on the y-axis). Perfect calibration would align the binned

https://doi.org/10.52202/085713-5069

Table 1: **ECE on pretrained feature extractors.** "Uncal." = uncalibrated; "Angular" = our method; "Platt/Iso" use $n_{\mathrm{ho}} \in \{100, 500\}$ labeled hold-out points.

Model–Dataset	Uncal.	Angular	Platt 100	Iso 100	Platt 500	Iso 500
ResNet-34–CIFAR-10	0.1236	0.0199	0.0561	0.0484	0.0259	0.0298
MiniLM–20 Newsgroups	0.1392	0.0249	0.0931	0.1107	0.0679	0.0813
ChemBERTa–Tox21	0.1389	0.0132	0.0236	0.0497	0.0175	0.0293

points with the 45° line. In the left and right panels, Platt scaling is derived using holdout sets of $n_{\mathrm{ho}} = 100$ and $n_{\mathrm{ho}} = 20000$, respectively, whereas both the uncalibrated predictor and the angular predictor remain unchanged across the two panels.

From the reliability plots, we see that the uncalibrated predictor (blue) is poorly calibrated, while, as expected, the angular predictor (green) shows good calibration. Here, we have used the angle estimator (11) and the sign estimator (12) to estimate the value of $\langle w_\star, \widehat{w} \rangle_\Sigma$. The estimated value for $\langle w_\star, \widehat{w} \rangle_\Sigma$ is 0.4356 while the true value is 0.4526. We also ran 5000 Monte Carlo trials where we found the probability of incorrect sign estimation to be 0.89% with a holdout set of size $n_{\mathrm{ho}} = 100$.

In contrast, the left panel of Figure 2 shows that Platt scaling (orange) with a holdout set size of $n_{\mathrm{ho}} = 100$ fails to properly calibrate. However, when the holdout set size is increased to $n_{\mathrm{ho}} = 20000$, Platt scaling also calibrates correctly. When $n_{\mathrm{ho}} = 20000$, the predictor calibrated from Platt scaling is found to be (using scikit-learn package's CalibratedClassifierCV routine [68])

$$\hat{f}^{n_{\mathrm{ho}}}_{\mathrm{platt}}(\widehat{w}^\top x_{\mathrm{new}}) = \sigma(\hat{A}^{n_{\mathrm{ho}}} \cdot \widehat{w}^\top x_{\mathrm{new}} + \hat{B}^{n_{\mathrm{ho}}}), \text{ with } \hat{A}^{n_{\mathrm{ho}}} = 0.3396, \hat{B}^{n_{\mathrm{ho}}} = -0.1521.$$

We now check if $\hat{A}^{n_{\mathrm{ho}}}$ and $\hat{B}^{n_{\mathrm{ho}}}$ are indeed close to A_*, B_* as claimed in Theorem 5.1. Note that even though the link function is not strictly a probit function as required by Theorem 5.1, by approximating $\mathrm{sigmoid}(u) \approx \Phi(\sqrt{\pi/8} \cdot u)$, we have $\sigma(u) \approx \Phi(\sqrt{\pi/8}(3u + 1))$. We then obtain $A_* = 0.2991, B_* = -0.1597$ from (9) setting $a = 3\sqrt{\pi/8}, b = \sqrt{\pi/8}$, which are indeed quite close to $\hat{A}^{n_{\mathrm{ho}}} = 0.3396, \hat{B}^{n_{\mathrm{ho}}} = -0.1521$.

7.2 Semi-real experiments

We assess angular calibration on semi-real tasks that keep real-data covariates but simulate labels from the known generative model. We maintain settings in Section 7.1 but replace data covariates with: (i) final-layer logits of pretrained deep networks; and (ii) classic UCI benchmarks.

We found that (11) plug-in estimator for $\langle w^\star, \hat{w} \rangle$ is unstable on these real datasets. This is a known issue for estimators like (11) that are based on Wigner-type random-matrix-theoretic assumptions. Modifying these estimator are an active research area [54, 57]. To isolate calibration effects, we simulate $w^\star$ and labels, using the true angle.

Each dataset is split into training set n, a large unlabeled pool $n_{\mathrm{cov}} \gg n_{\mathrm{train}}$ for covariance estimation, and n_{test}. We report Expected Calibration Error (ECE; lower is better) [29]. Post-hoc baselines use a labeled hold-out of size $n_{\mathrm{ho}} \in \{100, 500\}$ ("Platt 100/500" and "Iso 100/500").

Pretrained representations. We fit a linear head on frozen embeddings and calibrate the resulting logits: ResNet-34 (ImageNet-1K pretrain) on CIFAR-10 [35, 24, 46], MiniLM sentence embeddings on 20 Newsgroups [85, 72, 51], and ChemBERTa on Tox21 from MoleculeNet [20, 87]. We have $n \times d = (300 \times 512)/(800 \times 384)/(800 \times 768)$, $n_{\mathrm{cov}} = 30{,}000/3{,}000/3{,}000$ and $n_{\mathrm{test}} = 10{,}000/1{,}000/500$ for CIFAR-10 / 20NG / Tox21. The results are reported in Table 1; the reliability plots are deferred to Section H.2.

UCI benchmarks. On Communities & Crime, Splice-junction Gene Sequences, and Madelon [71, 1, 32], we train linear predictors on raw covariates with $n \times d = (200 \times 100)/(100 \times 180)/(300 \times 500)$, $n_{\mathrm{cov}} = 700/2000/900$, and $n_{\mathrm{test}} = 593/900/900$ for Communities & Crime / Splice-junction / Madelon. The results are reported in Table 2; the reliability plots are deferred to Section H.2.

Table 2: **ECE on UCI datasets.** Conventions as in Table 1.

Dataset	**Uncal.**	**Angular**	**Platt 100**	**Iso 100**	**Platt 500**	**Iso 500**
Communities & Crime	0.1700	0.0296	0.0620	0.0661	0.0262	0.0436
Splice-junction Gene Seq	0.1262	0.0208	0.0568	0.1041	0.0556	0.0879
Madelon	0.1473	0.0267	0.0721	0.0719	0.0299	0.0624

8 Extensions and future directions

We derived our theoretical results assuming that the covariates are Gaussian—although at first pass this might appear stylistic, recent universality results [37, 56, 33, 26, 49, 62] demonstrate that these results should continue to hold as long as the covariates have sufficiently light tails. We demonstrate this with further experiments. In Section H.2 and Figure 4, we reproduce Figure 1 and 2 with non-Gaussian design matrices (iid Rademacher and uniform entries respectively) where we observe that our results continue to be accurate. Establishing such universality formally should be an interesting avenue for future work—we include an informal discussion here. Consider a general setting where $x_i \overset{\mathrm{d}}{=} \Sigma^{1/2} z_i$, where z_i has iid entries with zero mean, unit variance and finite moments, and $\Sigma = p^{-1}\bar{\Sigma}$ where $\bar{\Sigma}$ has bounded condition number. Denote $\widetilde{w} = \bar{\Sigma}^{1/2}\widehat{w}$, $\widetilde{w}_\star = \bar{\Sigma}^{1/2} w^\star$ and $x_{\text{new}}, z_{\text{new}}$ to be observations at test time with the same distribution as x_i, z_i respectively. If we could apply the multivariate CLT [11], we would obtain

$$\left(p^{-1/2} \sum_{i=1}^{p} \widetilde{w}_{\star,i} z_{\text{new},i}, p^{-1/2} \sum_{i=1}^{p} \widetilde{w}_i z_{\text{new},i} \right) \Rightarrow (Z_1, Z_2), \tag{14}$$

where $(Z_1, Z_2) \sim N(0, L)$ for some positive definite covariance matrix $L \in \mathbb{R}^{2\times 2}$. To apply the multivariate CLT, we require to check the following moment condition (c.f. [11])

$$\begin{aligned} & p^{-3/2} \textstyle\sum_{i=1}^{p} \mathbb{E}\left[z_{\text{new},i}^{\frac{3}{2}} \right] \left[(\widetilde{w}_{\star,i}, \widetilde{w}_i) \begin{pmatrix} \|\widehat{w}\|_\Sigma^2 & \langle \widehat{w}, w_\star \rangle_\Sigma \\ \langle \widehat{w}, w_\star \rangle_\Sigma & \|w_\star\|_\Sigma^2 \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{w}_{\star,i} \\ \widetilde{w}_i \end{pmatrix} \right]^{\frac{3}{2}} \\ & \leq p^{-1/2} \mathbb{E}\left[z_{\text{new},i}^{\frac{3}{2}} \right] \sigma_{\min}^{-1}(L) \left(\tfrac{1}{p} \textstyle\sum_{i=1}^{p} |\widetilde{w}_{\star,i}|^3 + \tfrac{1}{p} \sum_{i=1}^{p} |\widetilde{w}_i|^3 \right) = o\left(\tfrac{1}{\sqrt{p}} \right). \end{aligned}$$

We claim that the above holds almost surely for sufficiently large p, if we have constants $W_1, W_2 > 0$ for which the following holds

$$\begin{pmatrix} \|\widehat{w}\|_\Sigma^2 & \langle \widehat{w}, w_\star \rangle_\Sigma \\ \langle \widehat{w}, w_\star \rangle_\Sigma & \|w_\star\|_\Sigma^2 \end{pmatrix} \to L, \quad \left(\frac{1}{p} \sum_{i=1}^{p} |\widetilde{w}_{\star,i}|^3, \frac{1}{p} \sum_{i=1}^{p} |\widetilde{w}_i|^3 \right) \to (W_1, W_2).$$

Recent universality results for either approximate message passing algorithms [19] or convex gaussian minmax theorems (CGMT) [33] allow one to prove this beyond Gaussian designs. Numerous works have already applied such arguments in the context of other high-dimensional problems [37, 62, 56, 49]. Using (14), the conditional distribution (15) in the proof of Theorem 3.2 can be extended to non-Gaussian, Wigner-type features preconditioned by some known $\Sigma^{1/2}$, thus leading to a proof of Theorem 2 beyond Gaussian designs. In the interest of space, we defer formalizing this to future work.

Finally, we consider binary classification in this work; it would be interesting to extend our results to multi-index models, which includes multi-class classification, additive and interaction models, and two-layer neural networks [82, 88, 15]; see details in Section F. Multi-index model can be defined as follows: for $K \geq 2$ and unobserved indices $W_\star = [w_{\star 1}, \ldots, w_{\star K}] \in \mathbb{R}^{d\times K}$, the true logits and model outputs are $G := W_\star^\top x_{\text{new}} \in \mathbb{R}^K$, $\pi(x_{\text{new}}) = g(G)$ where g is a generalized link (vector- or scalar-valued). We show in Section F that, given an estimator $\widehat{W} = [\widehat{w}_1, \ldots, \widehat{w}_K]$ of $W_\star$, angular predictor in Definition 3.1 may be extended for the multi-index model as follows,

$$\widehat{f}_{\text{ang}}\left(\widehat{W}^\top x_{\text{new}} \right) := \mathbb{E}_Z \left[g\left(M_\star S + L_\star Z \right) \right]$$

where the matrix quantity $M_\star$ and $L_\star$ depends on cross-index angles $\langle w_{\star k}, \widehat{w}_\ell \rangle_\Sigma, \ell, k \in [K]$. Though no estimators are given for these cross-index angles in literature as far as we know, recent theory for multi-index models [82] suggests that analogues of the single-index angle estimators [8] are feasible. We leave this to future works.

 https://doi.org/10.52202/085713-5069

Acknowledgments and Disclosure of Funding

P.S. was funded partially by NSF DMS Award No. 2113426 and NSF CAREER Award No. 2440824.

References

[1] Molecular biology (splice-junction gene sequences) [dataset]. UCI Machine Learning Repository, 1991.

[2] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.

[3] Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. In *International conference on machine learning*, pages 566–576. PMLR, 2021.

[4] Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005.

[5] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[6] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.

[7] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

[8] Pierre C Bellec. Observable adjustments in single-index models for regularized m-estimators. *arXiv preprint arXiv:2204.06990*, 2022.

[9] Pierre C Bellec and Cun-Hui Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28(2):713–743, 2022.

[10] Pierre C Bellec and Cun-Hui Zhang. Debiasing convex regularized estimators and interval estimation in linear models. *The Annals of Statistics*, 51(2):391–436, 2023.

[11] Vidmantas Bentkus. A lyapunov-type bound in rd. *Theory of Probability & Its Applications*, 49(2):311–323, 2005.

[12] Eugene Berta, Francis Bach, and Michael Jordan. Classifier calibration with roc-regularized isotonic regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR, 2024.

[13] Björn Böken. On the appropriateness of platt scaling in classifier calibration. *Information Systems*, 95:101641, 2021.

[14] Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009.

[15] Joan Bruna and Daniel Hsu. Survey on algorithms for multi-index models. *arXiv preprint arXiv:2504.05426*, 2025.

[16] Emmanuel J Candès and Pragya Sur. 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.

[17] Michael Celentano and Andrea Montanari. Correlation adjusted debiased lasso: debiasing the lasso with inaccurate covariate model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae039, 2024.

[18] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220, 2023.

[19] Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electron. J. Probab.*, 26:1–44, 2021.

[20] Sowmya Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[21] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Expectation consistency for calibration of neural networks. In *Uncertainty in Artificial Intelligence*, pages 443–453. PMLR, 2023.

[22] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Theoretical characterization of uncertainty in high-dimensional linear classification. *Machine Learning: Science and Technology*, 4(2):025029, 2023.

[23] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[25] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[26] Rishabh Dudeja, Subhabrata Sen, and Yue M Lu. Spectral universality in regularized linear regression with nearly deterministic sensing matrices. *IEEE Transactions on Information Theory*, 2024.

[27] Oliver Y Feng, Ramji Venkataramanan, Cynthia Rush, Richard J Samworth, et al. A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 15(4):335–536, 2022.

[28] Tilmann Gneiting and Adrian E Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

[29] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[30] Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.

[31] Chirag Gupta and Aaditya Ramdas. Online platt scaling with calibeating. In *International Conference on Machine Learning*, pages 12182–12204. PMLR, 2023.

[32] Isabelle Guyon. Madelon [dataset]. UCI Machine Learning Repository, 2003.

[33] Qiyang Han and Yandi Shen. Universality of regularized regression estimators in high dimensions. *The Annals of Statistics*, 51(4):1799–1823, 2023.

[34] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

 https://doi.org/10.52202/085713-5069

[36] Alexander Henzi, Johanna F Ziegel, and Tilmann Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):963–993, 2021.

[37] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

[38] Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593 – 2622, 2018.

[39] Kuanhao Jiang, Rajarshi Mukherjee, Subhabrata Sen, and Pragya Sur. A new central limit theorem for the augmented ipw estimator: Variance inflation, cross-fit covariance and beyond. *arXiv preprint arXiv:2205.10198*, 2022.

[40] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Smooth isotonic regression: a new method to calibrate predictive models. *AMIA Summits on Translational Science Proceedings*, 2011:16, 2011.

[41] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

[42] Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data, 2009.

[43] Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

[44] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, volume 1, page 9, 2009.

[45] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR, 2020.

[46] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[47] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

[48] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] Samriddha Lahiry and Pragya Sur. Universality in block dependent linear models with applications to nonlinear regression. *IEEE Transactions on Information Theory*, 2024.

[50] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[51] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 331–339, 1995.

[52] Yufan Li, Zhou Fan, Subhabrata Sen, and Yihong Wu. Random linear estimation with rotationally-invariant designs: Asymptotics at high temperature. *IEEE Transactions on Information Theory*, 70(3):2118–2153, 2023.

[53] Yufan Li, Subhabrata Sen, and Ben Adlam. Understanding optimal feature transfer via a fine-grained bias-variance analysis. *arXiv preprint arXiv:2404.12481*, 2024.

[54] Yufan Li and Pragya Sur. Spectrum-aware adjustment: A new debiasing framework with applications to principal components regression. *arXiv preprint arXiv:2309.07810*, 2023.

[55] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. 2020.

[56] Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and minimum-l1-norm interpolated classifiers. *The Annals of Statistics*, 50(3):1669–1695, 2022.

[57] Kevin Luo, Yufan Li, and Pragya Sur. Roti-gcv: Generalized cross-validation for right-rotationally invariant data. *arXiv preprint arXiv:2406.11666*, 2024.

[58] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

[59] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[60] Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. Evaluating uncertainty quantification in end-to-end autonomous driving control. *arXiv preprint arXiv:1811.06817*, 2018.

[61] Marco Mondelli and Ramji Venkataramanan. Approximate message passing with spectral initialization for generalized linear models. In *International Conference on Artificial Intelligence and Statistics*, pages 397–405. PMLR, 2021.

[62] Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

[63] Andrea Montanari, Subhabrata Sen, et al. A friendly tutorial on mean-field spin glass techniques for non-physicists. *Foundations and Trends® in Machine Learning*, 17(1):1–173, 2024.

[64] Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

[65] Allan H Murphy and Robert L Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1):41–47, 1977.

[66] Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015.

[67] Pratik Patil, Jin-Hong Du, and Ryan J Tibshirani. Optimal ridge regularization for out-of-distribution prediction. *arXiv preprint arXiv:2404.01233*, 2024.

[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[69] Nathan Phelps, Daniel J Lizotte, and Douglas G Woolford. Using platt's scaling for calibration after undersampling–limitations and how to address them. *arXiv preprint arXiv:2410.18144*, 2024.

[70] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[71] Michael Redmond. Communities and crime [dataset]. UCI Machine Learning Repository, 2002.

[72] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992, 2019.

[73] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[74] Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. *Advances in Neural Information Processing Systems*, 33:13331–13340, 2020.

 https://doi.org/10.52202/085713-5069

[75] Yanke Song, Sohom Bhattacharya, and Pragya Sur. Generalization error of min-norm interpolators in transfer learning. *arXiv preprint arXiv:2406.13944*, 2024.

[76] Yanke Song, Xihong Lin, and Pragya Sur. Hede: Heritability estimation in high dimensions by ensembling debiased estimators. *arXiv preprint arXiv:2406.11184*, 2024.

[77] Zeyu Sun, Dogyoon Song, and Alfred Hero. Minimum-risk recalibration of classifiers. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[79] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175:487–558, 2019.

[80] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

[81] Linh Tran, Bastiaan S Veeling, Kevin Roth, Jakub Swiatkowski, Joshua V Dillon, Jasper Snoek, Stephan Mandt, Tim Salimans, Sebastian Nowozin, and Rodolphe Jenatton. Hydra: Preserving ensemble diversity for model distillation. *arXiv preprint arXiv:2001.04694*, 2020.

[82] Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Fundamental computational limits of weak learnability in high-dimensional multi-index models. *arXiv preprint arXiv:2405.15480*, 2024.

[83] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[84] Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791 – 2823, 2020.

[85] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*, 2020.

[86] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.

[87] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

[88] Zhuoran Yang, Krishnakumar Balasubramanian, Zhaoran Wang, and Han Liu. Estimating high-dimensional non-gaussian multiple index models via stein's lemma. *Advances in Neural Information Processing Systems*, 30, 2017.

[89] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616, 2001.

[90] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

[91] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[92] Qian Zhao, Pragya Sur, and Emmanuel J Candes. The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861, 2022.

A Proof of Theorem 3.2

Before proving Theorem 3.2, we first show that $\hat{f}_{\text{ang}}(\cdot;\theta_*)$ is exactly calibrated.

Theorem A.1. *The predictor $\hat{f}_{\text{ang}}$ defined in* (3) *is well-calibrated at all $p \in [0,1]$ when . That is, for any $p \in [0,1]$ and any $d, n \in \mathbb{N}_+$*

$$\Delta_p^{\text{cal}}\left(\hat{f}_{\text{ang}}(\cdot;\theta_*)\right) = p - \mathbb{E}_{x_{\text{new}}}\left[\sigma\left(w_\star^\top x_{\text{new}}\right) \mid \hat{f}_{\text{ang}}\left(\widehat{w}^\top x_{\text{new}};\theta_*\right) = p\right] = 0.$$

Proof of Theorem A.1. Let us define the following event

$$\mathcal{A} := \left\{\hat{f}_{\text{ang}}\left(w_\star^\top x_{\text{new}};\theta_*\right) = p\right\}.$$

We have

$$\begin{aligned}
&\mathbb{E}_{x_{\text{new}}}\left[\sigma\left(w_\star^\top x_{\text{new}}\right) \mid \mathcal{A}\right] \\
&\overset{(i)}{=} \mathbb{E}_{x_{\text{new}}}\left[\mathbb{E}_{x_{\text{new}}}\left[\sigma\left(w_\star^\top x_{\text{new}}\right) \mid x_{\text{new}}^\top \widehat{w}\right] \mid \mathcal{A}\right] \\
&\overset{(ii)}{=} \mathbb{E}_{x_{\text{new}}}\left[\mathbb{E}_Z\left[\sigma\left(\frac{1}{\|\widehat{w}\|_\Sigma}\cdot\cos(\theta_*)\cdot x_{\text{new}}^\top\widehat{w} + \sin(\theta_*)\cdot Z\right)\right] \mid \mathcal{A}\right] \\
&= \mathbb{E}_{x_{\text{new}}}\left[\hat{f}_{\text{ang}}\left(w_\star^\top x_{\text{new}};\theta_*\right) \mid \mathcal{A}\right] \\
&= p.
\end{aligned}$$

where (i) follows from tower property of expectation and the fact that $\hat{f}_{\text{ang}}(x_{\text{new}})$ depends on x_{new} only through $x_{\text{new}}^\top\widehat{w}$, (ii) follows from conditional expectation of multivariate Gaussian distribution

$$\left[w_\star^\top x_{\text{new}} \mid x_{\text{new}}^\top\widehat{w}\right] \overset{L}{=} \frac{1}{\|\widehat{w}\|_\Sigma}\cdot\cos(\theta_*)\cdot x_{\text{new}}^\top\widehat{w} + \sin(\theta_*)\cdot Z \tag{15}$$

for some $Z \sim N(0,1)$. □

Proof of Theorem 3.2. Using result from Theorem A.1, it suffices to show that as $|\hat{\theta} - \theta_*| \to 0$ in probability, we have that

$$\left|\Delta_p^{\text{cal}}\left(\hat{f}_{\text{ang}}(\cdot;\theta_*)\right) - \Delta_p^{\text{cal}}\left(\hat{f}_{\text{ang}}\left(\cdot;\hat{\theta}\right)\right)\right| \to 0. \tag{16}$$

Let us introduce the following notation for the ease of presentation:

$$X := \sigma(w_\star^\top x_{\text{new}}), \quad \hat{Y} = \hat{f}_{\text{ang}}\left(\widehat{w}^\top x_{\text{new}};\hat{\theta}\right), \quad Y_* = \hat{f}_{\text{ang}}\left(\widehat{w}^\top x_{\text{new}};\theta_*\right).$$

Then, we can write LHS of (16) as

$$\left|\mathbb{E}[X \mid \hat{Y} = p] - \mathbb{E}\left[X \mid Y_\star = p\right]\right| = \left|\frac{1}{f_{\widehat{Y}}(p)}\int_0^1 x f_{X,\hat{Y}}(x,p)\mathrm{d}x - \frac{1}{f_{Y_\star}(p)}\int_0^1 x f_{X,Y_\star}(x,p)\mathrm{d}x\right|$$

where $f_{\hat{Y}}, f_{Y_*}, f_{X,Y_\star}, f_{X,\hat{Y}}$ are the distribution density functions of $\hat{Y}, Y_*$ and joint density functions of $(X, Y_\star)$ and $(X, \hat{Y})$. We now show that the RHS of the above converges to 0. Firstly, $|\frac{1}{f_{Y_\star}(p)} - \frac{1}{f_{\hat{Y}}(p)}| \to 0$ because $|\hat{Y} - Y_\star| \to 0$ in probability (and thus in distribution) by continuous mapping theorem. Secondly,

$$\left|\int_0^1 x f_{X,\hat{Y}}(x,p)\mathrm{d}x - \int_0^1 x f_{X,Y_\star}(x,p)\mathrm{d}x\right| \to 0$$

by bounded convergence theorem and the fact that $(X, \hat{Y})$ converges to (X, Y_*) jointly. We conclude the proof. □

https://doi.org/10.52202/085713-5069

B Proof of Theorem 4.2

We first state a result from [4] for general random variables.

Proposition B.1 (Theorem 1, [4]). *Let $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ be a strictly convex differentiable function, and let D_ϕ be the corresponding Bregman loss function. Let X be an arbitrary random variable taking values in $\mathbb{R}^d$ for which both $\mathbb{E}[X]$ and $\mathbb{E}[\phi(X)]$ are finite. Then, among all functions of Z, the conditional expectation is the unique minimizer (up to a.s. equivalence) of the expected Bregman loss, i.e.,*

$$\arg\min_{Y \in \sigma(Z)} \mathbb{E}\left[D_\phi(X, Y)\right] = \mathbb{E}[X \mid Z].$$

Using the above results, we show that angular calibration with $\hat{\theta} = \theta_*$ minimizes Bregman divergence to true label distribution among predictors of the form $F(\widehat{w}^\top x_{\text{new}})$.

Theorem B.2 (Optimality of angular predictor). *Let $\phi : \mathbb{R}^2 \mapsto \mathbb{R}$ be any strictly convex differentiable function, and let D_ϕ be the corresponding Bregman loss function. Let $\mathbb{E}_{x_{\text{new}}}[\phi(q_\star)]$ be finite. Then, the expected Bregman loss $\mathbb{E}_{x_{\text{new}}}\left[D_\phi(q_\star, \hat{q}_F)\right]$ admits a unique minimizer (up to a.s. equivalence) among all $q_F, \forall F \in \mathcal{F} := \{f : \mathbb{R} \to [0,1]\}$. Let this minimizer be $F_\star = \arg\min_{F \in \mathcal{F}} \mathbb{E}_{x_{\text{new}}}\left[D_\phi(q_\star, \hat{q}_F)\right]$. We then have that almost surely*

$$\hat{q}_{\text{ang}}(\theta_*) = F_*(\widehat{w}^\top x_{\text{new}})$$

where $\hat{q}_{\text{ang}}(\theta_*)$ is the label prediction probability vector by angular calibration given in (4) *with $\hat{\theta}$ replaced by θ_*.*

Proof of Theorem B.2. Firstly, we set X, Y, Z in Theorem B.1 as

$$X \leftarrow \begin{pmatrix} \sigma(w_\star^\top x_{\text{new}}) \\ 1 - \sigma(w_\star^\top x_{\text{new}}) \end{pmatrix}, \quad Y \leftarrow \begin{pmatrix} F(\widehat{w}^\top x_{\text{new}}) \\ 1 - F(\widehat{w}^\top x_{\text{new}}) \end{pmatrix}, \quad Z \leftarrow \widehat{w}^\top x_{\text{new}}.$$

The result then follows from Theorem B.1 and the following

$$\mathbb{E}[X \mid Z] = \mathbb{E}\left[\begin{pmatrix} \sigma(w_\star^\top x_{\text{new}}) \\ 1 - \sigma(w_\star^\top x_{\text{new}}) \end{pmatrix} \mid x_{\text{new}}^\top \widehat{w}\right] = \begin{pmatrix} \hat{f}_{\text{ang}}(x_{\text{new}}) \\ 1 - \hat{f}_{\text{ang}}(x_{\text{new}}) \end{pmatrix}$$

where we used (15) and (3) for the last equality. □

Now we are ready to state proof of Theorem 4.2.

Proof. Using result from Theorem B.2, it suffices to show that

$$\left\| \hat{q}_{\text{ang}}(\hat{\theta}) - \hat{q}_{\text{ang}}(\theta_*) \right\|_2 \to 0$$

in probability. This is an immediate consequence of the continuous mapping theorem under the assumption σ is continuous. □

C Proof of Theorem 5.1

Before proving Theorem 5.1, we first state two classic analysis results that we will later use.

Proposition C.1 (Theorem 5.7, [83]). *Let ℓ_n be random functions on $\mathcal{H}$, $\ell^\star$ be a fixed function on $\mathcal{H}$, and $\theta^\star \in \mathcal{H}$ such that (i) uniform convergence of $n^{-1}\ell_n$ to $\ell^\star$ holds:*

$$\sup_{\theta \in \mathcal{H}} \left| \frac{1}{n}\ell_n(\theta) - \ell^\star(\theta) \right| \xrightarrow[n \to \infty]{\text{inprob.}} 0,$$

(ii) the mode of $\ell^\star$ is well-separated, i.e for all $\varepsilon > 0$,

$$\sup_{\theta \in \mathcal{H} : d(\theta, \theta^\star) \geq \varepsilon} \ell^\star(\theta) < \ell^\star(\theta^\star)$$

Then any sequence $\hat{\theta}_n$ maximizing ℓ_n converges in probability to $\theta^\star$.

https://doi.org/10.52202/085713-5069

Proposition C.2 (Theorem 10.8, [73]). *Let C be a relatively open convex set, and let $f_1, f_2, \ldots,$ be a sequence of finite convex functions on C. Suppose that the sequence converges pointwise on a dense subset of C, i.e. that there exists a subset C' of C such that its closure satisfies $\mathrm{cl}C' \supset C$ and, for each $x \in C'$, the limit of $f_1(x), f_2(x), \ldots,$ exists and is finite. The limit then exists for every $x \in C$, and the function f, where*

$$f(x) = \lim_{i\to\infty} f_i(x)$$

is finite and convex on C. Moreover the sequence $f_1, f_2, \ldots,$ converges to f uniformly on each closed bounded subset of C.

Now we are ready to prove Theorem 5.1.

Proof of Theorem 5.1. (8) is obtained from applying the well-known identity below for probit function $\Phi(\cdot)$

$$\mathbb{E}\Phi(\mu + \sigma \cdot Z) = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right), \qquad Z \sim N(0,1)$$

to (3).

To prove that $\hat{A}^{n_{\mathrm{ho}}} \to A_*, \hat{B}^{n_{\mathrm{ho}}} \to B_*$ as $n_{\mathrm{ho}} \to \infty$, we would like to apply (C.1) by setting $n \leftarrow n_{\mathrm{ho}}, \theta \leftarrow \{A, B\}, \theta^\star \leftarrow \{A_*, B_*\}$,

$$\ell_n(\theta) \leftarrow \ell_{n_{\mathrm{ho}}}(A,B) := \sum_{i=1}^{n_{\mathrm{ho}}} -y_{\mathrm{ho},i} \log\big(F_{A,B}(\widehat{w}^\top x_{\mathrm{ho},i})\big) - (1-y_{\mathrm{ho},i}) \log\big(1 - F_{A,B}(\widehat{w}^\top x_{\mathrm{ho},i})\big), \tag{17}$$

and

$$\begin{aligned}\ell^\star(\theta) \leftarrow \ell^\star(A,B) :=& \mathbb{E}_{x_{\text{new}}}\Big[-\sigma\left(w_\star^\top x_{\text{new}}\right) \log\left(F_{A,B}\left(\widehat{w}^\top x_{\text{new}}\right)\right) \\ & - \left(1 - \sigma\left(w_\star^\top x_{\text{new}}\right)\right) \log\left(1 - F_{A,B}\left(\widehat{w}^\top x_{\text{new}}\right)\right)\Big].\end{aligned} \tag{18}$$

where we used the notation

$$F_{A,B}(u) := \sigma(Au + B) = \Phi(a(Au+B)+b).$$

Here, RHS of (18) is up to an affine transform of the KL divergence (6) and, therefore, it follows from (8) and (**??**) that its minimizer is indeed A_*, B_*.

To verify condition (i) of Theorem C.1, we first note that (17) converges to (18) point-wise in probability following from law of large number theorem. Furthermore, we note that the functions

$$f_1(u) := -\log(\Phi(u)), \qquad f_2(u) := -\log(1 - \Phi(u))$$

are both strictly convex functions. To see this, one can show second derivatives are positive for all $u \in \mathbb{R}$ by utilizing the two elementary inequalities: $u\Phi(u) + \Phi'(u) > 0$, $u\Phi(u) + \Phi'(u) > u$. It then follows that $\ell_{n_{\mathrm{ho}}}(A,B)$ is a convex function of (A,B) on the domain $\mathbb{R}^2$. Uniform convergence on $\mathcal{H}$ is then an application of Theorem C.2.

To verify condition (ii) of Theorem C.1, we only need to show that $\ell^\star(A,B)$ is a strictly convex function in (A,B). Let us write the inside of the expectation of (18) as

$$f(A,B) = \sigma(H) f_1(aA \cdot H + aB + b) + (1 - \sigma(H)) f_2(aA \cdot H + aB + b)$$

where $H := w_\star^\top x_{\mathrm{new}} \sim N(0,1)$. Then, we have that

$$\nabla^2 f(A,B) = \big(\sigma(H) \cdot f_1''(aA \cdot H + aB + b) \cdot + (1 - \sigma(H)) f_2''(aA \cdot H + aB + b)\big) \cdot \begin{pmatrix} a^2H^2 & a^2H \\ a^2H & a^2 \end{pmatrix}$$

which is positive-definite almost surely when $a \neq 0$. Hence, we have that almost surely

$$f\left(t(A_1,B_1) + (1-t)(A_1,B_1)\right) < tf\left((A_1,B_1)\right) + (1-t) f\left((A_1,B_1)\right)$$

which implies that

$$\mathbb{E}_{x_{\text{new}}} f\left(t(A_1,B_1) + (1-t)(A_1,B_1)\right) < t\mathbb{E}_{x_{\text{new}}} f\left((A_1,B_1)\right) + (1-t)\mathbb{E}_{x_{\text{new}}} f\left((A_1,B_1)\right).$$

 https://doi.org/10.52202/085713-5069

The claim that $\ell^\star(A, B)$ is strictly convex follows.

It then follows from Theorem C.1 that $\hat{A}^{n_{\rm ho}} \to A_*, \hat{B}^{n_{\rm ho}} \to B_*$ in probability as $n_{\rm ho} \to \infty$. The uniform convergence $\hat{f}^{n_{\rm ho}}_{\rm platt}(u) \to \hat{f}_{\rm ang}(u)$ follows immediately.

□

D Inner product estimation

We restate the following results from Theorem 4.4, [8]. We note that the quantity $\frac{\hat{r}^4}{\hat{v}^2\hat{t}^2}$ in the error bound is observable and is typically of constant order in the proportional regime. See [8] for details.

Theorem D.1. *Suppose ℓ is continuously differentiable and g is strongly convex and twice differential penalty function. Assume also that $\frac{1}{2\delta} \le \frac{d}{n} \le \frac{1}{\delta}$ for some $\delta > 0$, for arbitrarily large probability $1-\delta$, the following holds*

$$\mathbb{E}\left|\hat{a}_*^2 - \langle w_\star, \widehat{w}\rangle_\Sigma^2\right| \le \frac{C\hat{r}^4}{\hat{v}^2\hat{t}^2} \cdot n^{-1/2}$$

where C is a constant depending only on g and δ.

E Sign estimation

Proof of Theorem 6.1. With respect to randomness in validation dataset (that is, $\widehat{w}$ is treated as deterministic), we have that $\widehat{w}^\top x_i^{\rm ho} \cdot y_i^{\rm ho}$ are iid across i and satisfies that

$$\begin{aligned}\widehat{w}^\top x_i^{\rm ho} \cdot y_i^{\rm ho} &\overset{L}{=} \widehat{w}^\top x_i^{\rm ho} \cdot \mathrm{Bern}_i\left(\sigma\left(\frac{\langle w_\star, \widehat{w}\rangle_\Sigma}{\widehat{w}^\top\Sigma\widehat{w}}\widehat{w}^\top x_i^{\rm ho} + \sin(\theta_*)\cdot Z_i\right)\right)\\ &\overset{L}{=} \|\widehat{w}\|_\Sigma U_i \cdot \mathrm{Bern}_i\left(\sigma\left(\cos(\theta_*)U_i + \sin(\theta_*)\cdot Z_i\right)\right) = H_i\end{aligned}$$

where $Z_i \overset{iid}{\sim} N(0,1)$ and we have used $w^\top x_i^{\rm ho} \overset{L}{=} \|\widehat{w}\|_\Sigma \cdot U_i$ for $U_i \overset{\rm iid}{\sim} N(0,1)$. It follows from Gaussian integration by parts that

$$\mathbb{E}H_i = \langle w_\star, \widehat{w}\rangle_\Sigma \cdot \mathbb{E}\sigma'\left(\cos(\theta_*)U_i + \sin(\theta_*)\cdot Z_i\right) = \langle w_\star, \widehat{w}\rangle_\Sigma \mathbb{E}\sigma'(Z).$$

Meanwhile, H_i is subGaussian with subGaussian norm

$$\|H_i\|_{\psi_2}^2 \le \|\widehat{w}\|_\Sigma^2 \|U_i\|_{\psi_2}^2 \le 3\|\widehat{w}\|_\Sigma^2.$$

By theorem assumption, $\mathbb{E}\sigma'(Z) > 0$ and

$$\mathrm{sign}\left(\langle w_\star, \widehat{w}\rangle_\Sigma \mathbb{E}\sigma'(Z)\right) = \mathrm{sign}\left(\langle w_\star, \widehat{w}\rangle_\Sigma\right)$$

So the sign identification of $\widehat{\mathrm{sgn}}$ is correct if the following event holds

$$\left|\frac{1}{n_{\rm ho}}\sum_{i=1}^{n_{\rm ho}} w^\top x_i \cdot y_i - \langle w_\star, \widehat{w}\rangle_\Sigma \mathbb{E}\sigma'(Z)\right| < \left|\langle w_\star, \widehat{w}\rangle_\Sigma \mathbb{E}\sigma'(Z)\right|.$$

By Hoeffding's inequality,

$$\begin{aligned}&\mathbb{P}_{\rm ho}\left(\left|\frac{1}{n_{\rm ho}}\sum_{i=1}^{n_{\rm ho}} w^\top x_i \cdot y_i - \langle w_\star, \widehat{w}\rangle_\Sigma \mathbb{E}\sigma'(Z)\right| > \left|\langle w_\star, \widehat{w}\rangle_\Sigma \mathbb{E}\sigma'(Z)\right|\right)\\ &\le 2\exp\left(-\frac{cn_{\rm ho}\left(\mathbb{E}\sigma'(Z)\right)^2\langle w_\star, \widehat{w}\rangle_\Sigma^2}{\|\widehat{w}\|_\Sigma^2}\right) = 2\exp\left(-cn_{\rm ho}\left(\cos(\theta_*)\cdot\mathbb{E}\sigma'(Z)\right)^2\right).\end{aligned}$$

The theorem statement follows. □

F Angular calibration for multi-index models

Let $x_{\text{new}} \sim \mathcal{N}(0, \Sigma) \subset \mathbb{R}^d$. Fix $K \geq 2$ and let $W_\star = [w_{\star 1}, \ldots, w_{\star K}] \in \mathbb{R}^{d \times K}$. Define

$$G := W_\star^\top x_{\text{new}} \in \mathbb{R}^K, \quad \pi(x_{\text{new}}) = g(G),$$

where g is a generalized link (vector- or scalar-valued). This setup covers:

- Two-layer nets (frozen outer layer): $g(u) = \sum_k a_k \, \sigma(u_k)$
- Multi-class softmax: $g(u) = \mathrm{softmax}(u)$
- Additive index model: $g(u) = \sum_k f_k(u_k)$
- Interaction index model: $g(u) = \sum_k f_k(u_k) + \sum_{k<\ell} h_{k\ell}(u_k, u_\ell)$

The following extends angular calibration to mutli-index models. Note that to apply the angular predictor (19), we must estimate cross-index angles $\langle w_{\star,k}, \widehat{w}_\ell \rangle_\Sigma$, similarly to the single-index case. Though no estimator is given in literature as far as we know, recent theory for multi-index models [82] suggests that analogues of the single-index angle estimators [8] are feasible. We leave derivation of these estimators to future works.

Theorem F.1 (Angular calibration for multi-index models)**.** *Let $\widehat{W} = [\widehat{w}_1, \ldots, \widehat{w}_K] \in \mathbb{R}^{d \times K}$ be any estimator. Set*

$$D := \mathrm{diag}(\|\widehat{w}_1\|_\Sigma, \ldots, \|\widehat{w}_K\|_\Sigma), \quad S := D^{-1} \widehat{W}^\top x_{\text{new}} \in \mathbb{R}^K.$$

Then we have $K \times K$ covariance blocks

$$\mathrm{Cov}(G) = W_\star^\top \Sigma W_\star, \quad R := \mathrm{Cov}(S) = D^{-1} \widehat{W}^\top \Sigma \widehat{W} D^{-1}, \quad C := \mathrm{Cov}(G, S) = W_\star^\top \Sigma \widehat{W} D^{-1}$$

where

$$R_{k\ell} = \frac{\langle \widehat{w}_k, \widehat{w}_\ell \rangle_\Sigma}{\|\widehat{w}_\ell\|_\Sigma \|\widehat{w}_k\|_\Sigma}, \qquad C_{k\ell} = \frac{\langle w_{\star k}, \widehat{w}_\ell \rangle_\Sigma}{\|\widehat{w}_\ell\|_\Sigma \|w_{\star k}\|_\Sigma}$$

Then, assuming that R is invertible, we may define

$$M_\star := C R^{-1}, \qquad \Sigma_\star := \mathrm{Cov}(G) - C R^{-1} C^\top$$

and we have that $G \mid S \sim \mathcal{N}(M_\star S, \Sigma_\star)$. For any factor $L_\star$ with $L_\star L_\star^\top = \Sigma_\star$ and any $Z \sim \mathcal{N}(0, I_K)$ independent of everything, define the multi-index angular predictor

$$\widehat{f}_{\text{ang}}\big(\widehat{W}^\top x_{\text{new}}\big) := \mathbb{E}_Z \left[g\big(M_\star S + L_\star Z\big) \right]. \tag{19}$$

Then for any $p \in \Delta^{K-1}$ and any $d, n \in \mathbb{N}_+$,

$$p - \mathbb{E}\left[\pi(x_{\text{new}}) \;\middle|\; \widehat{f}_{\text{ang}}(\widehat{W}^\top x_{\text{new}}) = p \right] = 0.$$

Proof. Let $x_{\text{new}} \sim \mathcal{N}(0, \Sigma)$ and define

$$G := W_\star^\top x_{\text{new}} \in \mathbb{R}^K, \qquad S := D^{-1} \widehat{W}^\top x_{\text{new}} \in \mathbb{R}^K,$$

with $D = \mathrm{diag}(\|\widehat{w}_1\|_\Sigma, \ldots, \|\widehat{w}_K\|_\Sigma)$ and $\|u\|_\Sigma := \sqrt{u^\top \Sigma u}$, $\langle u, v \rangle_\Sigma := u^\top \Sigma v$. Set the $2K \times d$ linear map

$$T := \begin{bmatrix} W_\star^\top \\ D^{-1} \widehat{W}^\top \end{bmatrix}, \qquad Y := \begin{bmatrix} G \\ S \end{bmatrix} = T \, x_{\text{new}}.$$

Since x_{new} is Gaussian and Y is a linear transform, Y is jointly Gaussian with mean 0 and covariance

$$\mathrm{Cov}(Y) = T \Sigma T^\top = \begin{bmatrix} W_\star^\top \Sigma W_\star & W_\star^\top \Sigma \widehat{W} D^{-1} \\ D^{-1} \widehat{W}^\top \Sigma W_\star & D^{-1} \widehat{W}^\top \Sigma \widehat{W} D^{-1} \end{bmatrix}.$$

 https://doi.org/10.52202/085713-5069

Thus,

$$\mathrm{Cov}(G) = W_\star^\top \Sigma W_\star, \quad R := \mathrm{Cov}(S) = D^{-1}\widehat{W}^\top \Sigma \widehat{W} D^{-1}, \quad C := \mathrm{Cov}(G,S) = W_\star^\top \Sigma \widehat{W} D^{-1}.$$

In particular, for $k, \ell \in [K]$,

$$R_{k\ell} = \frac{\langle \widehat{w}_k, \widehat{w}_\ell \rangle_\Sigma}{\|\widehat{w}_k\|_\Sigma \, \|\widehat{w}_\ell\|_\Sigma}, \qquad C_{k\ell} = \frac{\langle w_{\star k}, \widehat{w}_\ell \rangle_\Sigma}{\|w_{\star k}\|_\Sigma \, \|\widehat{w}_\ell\|_\Sigma}.$$

Define the $K \times K$ matrices

$$M_\star := CR^{-1}, \qquad \Sigma_\star := \mathrm{Cov}(G) - CR^{-1}C^\top.$$

Consider the linear residual

$$U := G - M_\star S = G - CR^{-1}S.$$

Because Y is Gaussian, U is Gaussian; further,

$$\mathrm{Cov}(U,S) = \mathrm{Cov}(G,S) - CR^{-1}\,\mathrm{Cov}(S,S) = C - CR^{-1}R = 0,$$

so U and S are independent (uncorrelated jointly Gaussian vectors are independent). Moreover,

$$\mathrm{Cov}(U) = \mathrm{Cov}(G) - CR^{-1}C^\top = \Sigma_\star.$$

Hence we have the orthogonal decomposition

$$G = M_\star S + U, \qquad U \perp\!\!\!\perp S, \qquad U \sim \mathcal{N}(0, \Sigma_\star).$$

Equivalently, the conditional distribution is

$$G \,\big|\, S \sim \mathcal{N}\big(M_\star S,\ \Sigma_\star\big),$$

which is the standard multivariate normal conditioning formula via the Schur complement.

Finally, let $L_\star$ be any matrix satisfying $L_\star L_\star^\top = \Sigma_\star$ and let $Z \sim \mathcal{N}(0, I_K)$ independent of (G,S). Then $U \overset{d}{=} L_\star Z$ and

$$G \,\big|\, S \overset{d}{=} M_\star S + L_\star Z.$$

Therefore, for any measurable g (vector- or scalar-valued) with the requisite integrability,

$$\mathbb{E}[g(G) \mid S] = \mathbb{E}_Z\big[g\big(M_\star S + L_\star Z\big)\big],$$

which yields the stated multi-index angular predictor

$$\widehat{f}_{\mathrm{ang}}\big(\widehat{W}^\top x_{\mathrm{new}}\big) := \mathbb{E}_Z\big[g\big(M_\star S + L_\star Z\big)\big].$$

To conclude, set $X := \pi(x_{\mathrm{new}}) = g(G) \in \Delta^{K-1}$ and

$$Y \ := \ \widehat{f}_{\mathrm{ang}}\big(\widehat{W}^\top x_{\mathrm{new}}\big) = \mathbb{E}[g(G) \mid S].$$

Then Y is $\sigma(S)$-measurable and $Y = \mathbb{E}[X \mid S]$ almost surely. We claim that

$$\mathbb{E}[X \mid Y] = Y \qquad \text{a.s.}$$

Indeed, for any bounded measurable $\varphi : \Delta^{K-1} \to \mathbb{R}$,

$$\mathbb{E}\big[\varphi(Y)\,(X - Y)\big] = \mathbb{E}\Big[\ \mathbb{E}\big[\varphi(Y)\,(X-Y) \mid S\big]\ \Big] = \mathbb{E}\big[\varphi(Y)\,\big(\mathbb{E}[X \mid S] - Y\big)\big] = 0,$$

so $\mathbb{E}[X \mid Y] = Y$ (coordinate-wise) by the defining property of conditional expectation.

By the existence of regular conditional expectations, this implies that for $\mathbb{P}_Y$-almost every $p \in \Delta^{K-1}$,

$$\mathbb{E}\Big[\pi(x_{\mathrm{new}}) \ \Big|\ \widehat{f}_{\mathrm{ang}}(\widehat{W}^\top x_{\mathrm{new}}) = p\Big] = \mathbb{E}[X \mid Y = p] = p,$$

i.e.,

$$p - \mathbb{E}\Big[\pi(x_{\mathrm{new}}) \ \Big|\ \widehat{f}_{\mathrm{ang}}(\widehat{W}^\top x_{\mathrm{new}}) = p\Big] = 0.$$

This establishes exact calibration of the multi-index angular predictor. □

G Pseudocode

Algorithm 1: Angular Calibration

Input : Training data $\{(x_i, y_i)\}_{i=1}^n$; link σ; convex loss ℓ and penalty g;
covariance Σ (known or estimated from unlabeled data);
holdout set $\{(x_i^{\mathrm{ho}}, y_i^{\mathrm{ho}})\}_{i=1}^{n_{\mathrm{ho}}}$ for sign estimation;

Output : Calibrated predictor $\hat{f}_{\mathrm{ang}}(x)$ that returns $\hat{p} \in [0, 1]$ for any new x.

(A) Fit base linear model

$\widehat{w} \leftarrow \arg\min_w \ \frac{1}{n} \sum_{i=1}^n \ell_{y_i}(w^\top x_i) + g(w)$

$\|\widehat{w}\|_\Sigma \leftarrow (\widehat{w}^\top \Sigma\, \widehat{w})^{1/2}$

(B) Observable magnitude of $\langle w_\star, \widehat{w}\rangle_\Sigma$

$\hat{\psi}_i \leftarrow -\,\ell'_{y_i}(x_i^\top \widehat{w}); \quad \hat{\psi} \leftarrow (\hat{\psi}_1, \ldots, \hat{\psi}_n)^\top$

$D \leftarrow \mathrm{diag}\big(\ell''_y(X\widehat{w})\big); \quad \hat{H} \leftarrow \big(X^\top D X + n\,\nabla^2 g(\widehat{w})\big)^{-1}$

$\hat{v} \leftarrow \frac{1}{n}\,\mathrm{Tr}(D - DX\hat{H}X^\top D); \quad \hat{\gamma} \leftarrow \mathrm{Tr}(X\hat{H}X^\top D); \quad \hat{r}^2 \leftarrow \|\hat{\psi}\|^2/n$

$$\hat{a}_*^2 \leftarrow \frac{\left(\frac{\hat{v}}{n}\|X\widehat{w} - \hat{\gamma}\hat{\psi}\|^2 + \frac{1}{n}\hat{\psi}^\top X\widehat{w} - \hat{\gamma}\hat{r}^2\right)^2}{\frac{1}{n^2}\left\|\Sigma^{-\frac{1}{2}} X^\top \hat{\psi}\right\|^2 + \frac{2\hat{v}}{n}\hat{\psi}^\top X\widehat{w} + \frac{\hat{v}^2}{n}\|X\widehat{w} - \hat{\gamma}\hat{\psi}\|^2 - \frac{d}{n}\hat{r}^2} \quad \texttt{// Est. of } \langle w_\star, \widehat{w}\rangle_\Sigma^2$$

(C) Sign via holdout correlation

$\widehat{\mathrm{sgn}} \leftarrow \mathrm{sign}\big(\sum_{i=1}^{n_{\mathrm{ho}}} (\widehat{w}^\top x_i^{\mathrm{ho}})\, y_i^{\mathrm{ho}}\big)$

(D) Angle & interpolation weights

$c \leftarrow \dfrac{\widehat{\mathrm{sgn}}\,\sqrt{\hat{a}_*^2}}{\|\widehat{w}\|_\Sigma}; \quad c \leftarrow \min\{1, \max\{-1,\, c\}\}$ `// numerical clip`

$\hat{\theta} \leftarrow \arccos(c); \quad \alpha \leftarrow \cos\hat{\theta} = c; \quad \beta \leftarrow \sin\hat{\theta} = \sqrt{1 - \alpha^2}$

(E) Define calibrated predictor $\hat{f}_{\mathrm{ang}}$

For any new x:

$u \leftarrow \widehat{w}^\top x; \quad s \leftarrow u/\|\widehat{w}\|_\Sigma$

draw $z_1, \ldots, z_M \overset{iid}{\sim} \mathcal{N}(0, 1)$ and set $\hat{f}_{\mathrm{ang}}(x) \approx \frac{1}{M}\sum_{j=1}^M \sigma(\alpha s + \beta z_j)$

return $\hat{f}_{\mathrm{ang}}(x)$

https://doi.org/10.52202/085713-5069

H Additional plots

H.1 Universality

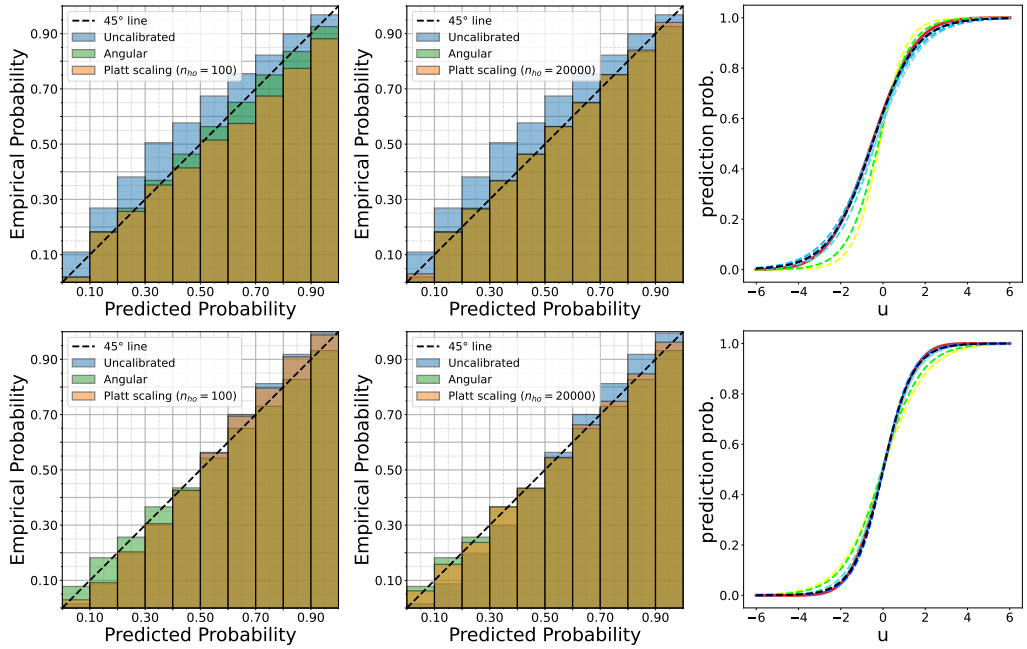


Figure 3: Reproduce Figure 1 (in the third column) and 2 (in first two columns) for Rademacher entries. Upper Row: rerun simulations in Section 7 but with subGaussian designs $W\Sigma^{1/2}$ where W_{ij} are sampled iid from Rademacher distribution, taking values $+1, -1$ with equal probability. Bottom Row: we replace the sigmoid link function in Section 7 with a clipped relu link function $\sigma(x) = \text{clip}(3x + 0.5)$ where $\text{clip}(x) = x, \forall x \in [0, 1]$, $\text{clip}(x) = 0, \forall x < 0$ and $\text{clip}(x) = 1, \forall x > 1$.

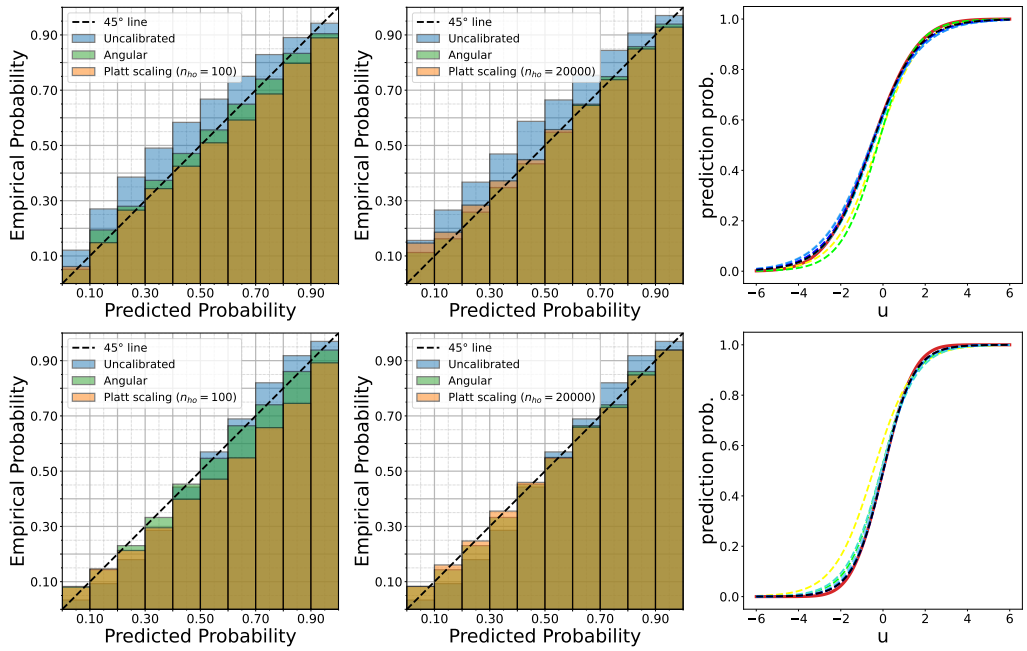
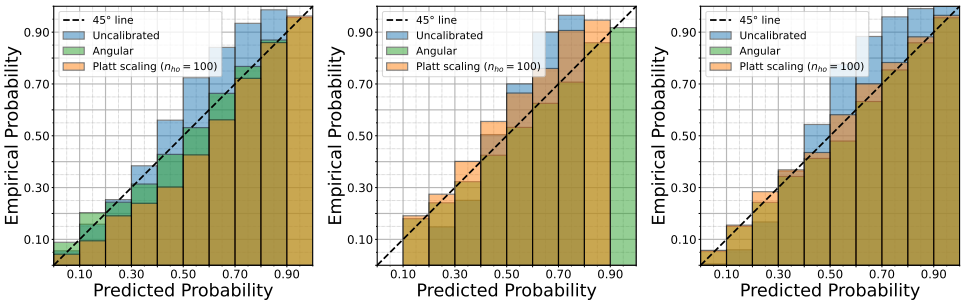


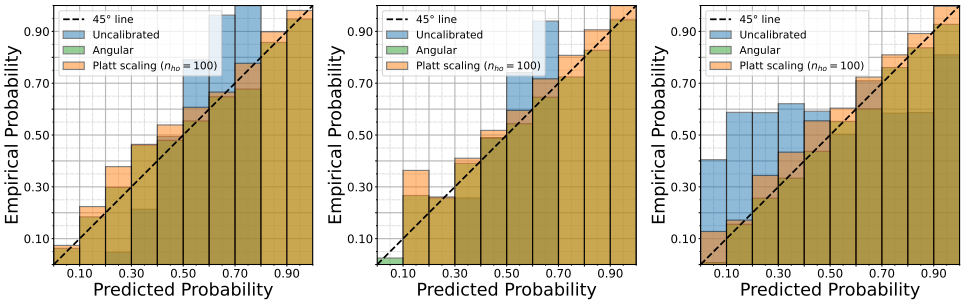
Figure 4: Reproduce Figure 1 (the third column) and 2 (first two columns) for uniform entries. Upper Row: rerun simulations in Section 7 but with non-Gaussian designs $W\Sigma^{1/2}$ where W_{ij} are sampled iid from uniform distribution, taking values in interval $[-\sqrt{12}/2, \sqrt{12}/2]$ uniformly at random. Bottom Row: we replace the sigmoid link function in Section 7 with a clipped relu link function $\sigma(x) = \text{clip}(3x + 0.5)$ where $\text{clip}(x) = x, \forall x \in [0, 1]$, $\text{clip}(x) = 0, \forall x < 0$ and $\text{clip}(x) = 1, \forall x > 1$.

https://doi.org/10.52202/085713-5069

H.2 UCI benchmarks and pretrained embeddings



(a) CIFAR-10 (b) 20 Newsgroups (c) Tox21



(d) Communities & Crime (e) DNA (f) Madelon

I Simulation code and reproduction

The simulation code Angular_calibration.ipynb, Semi_experiments.ipynb included as supplementary material. The experiment can be run on modern personal computers without needing special computing hardware. Detailed instructions is provided in the simulation code.

https://doi.org/10.52202/085713-5069

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clearly state the claims made, including the contributions made in the paper and important assumptions and limitations, in abstract and introduction

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Limitations and future directions section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are included either in section 2 or in each theorem statement. Full proofs are included in appendices.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have fully disclose all the information needed to reproduce the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

 https://doi.org/10.52202/085713-5069

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided source code for our simulations in supplementary file, along with detailed instructions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All simulation details are disclosed in Section 7 Simulations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix F

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work is purely theoretical, aimed at advancing the mathematical understanding of calibration in an idealistic setting. It does not involve any personal, sensitive, or protected information. As such, there is no plausible way for our methods to be used for discrimination, profiling, surveillance, or any other harmful purpose.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/085713-5069

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/085713-5069

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM is only used for minor grammar editing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.