A Theoretical Analysis of Bayes-Optimal Calibration under the Logit-Noise Model

Anonymous authors
Paper under double-blind review

Abstract

We formulate the calibration of predicted probabilities as a problem in Bayesian decision theory. Under an additive Gaussian noise model in logit space, we show that the Bayes-optimal calibration function for the squared-error loss is uniquely characterized as the conditional posterior mean $\mathbb{E}[p^* \mid \hat{p}]$, and our main theorem gives it in closed form as a sigmoid integral over a Gaussian posterior under a Gaussian prior. We further show that an approximation to this optimal estimator takes the same functional form as temperature scaling, and we derive the correspondence between the temperature parameter and the noise variance. We also prove that the Bayes-optimal estimator under the Brier score reduces to the same posterior mean, providing a unified theoretical basis for squared-type losses. Finally, numerical experiments under two conditions synthetic logit data in which the assumption holds by construction, and data mediated by logistic regression in which the assumption holds only empirically corroborate the predicted theoretical behavior and quantify the range over which temperature-scaling-type correction is effective. This verification presupposes that the true logit is analytically known; extending noise-level estimation to real data, where it is not, remains an open problem.

1 Introduction

In classification problems in machine learning, the reliability of predicted probabilities has become as important a concern as classification accuracy itself. In safety-critical applications such as medical diagnosis and autonomous driving, predicted probabilities are required to reflect the true posterior probability accurately.

Calibration error is the standard measure for quantifying this requirement. A classifier is perfectly calibrated if, among the samples for which it outputs a predicted probability $\hat{p}$, the empirical rate of positive labels equals $\hat{p}$. However, recent deep learning models have been reported to trade calibration for accuracy, exhibiting a systematic tendency toward overconfidence (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017).

In the existing literature, temperature scaling has largely been treated as an empirical procedure that estimates the temperature parameter by minimizing the negative log-likelihood on a validation set (Guo et al., 2017). In contrast, the question of under what conditions a temperature-scaling-type correction corresponds to a Bayes-optimal estimator, once the generative process of the predicted probability is explicitly assumed, has not been fully addressed.

We formulate calibration as the decision problem of estimating the true posterior probability from an observed predicted probability, and we analyze it as a problem of conditional Bayes risk minimization under an additive noise generative model in logit space. We restrict our analysis to binary classification and present theoretical analysis and numerical verification in a low-dimensional setting where the logit-noise assumption holds, either by construction or empirically. The overconfidence of deep learning models mentioned above motivates this study, but a direct application to high-dimensional, multi-class deep learning models lies beyond the empirical scope of this paper and is left for future work.

The theoretical contributions of this paper consist of three parts. First, we integrate an additive Gaussian noise model in logit space with Bayesian decision theory: the Bayes-optimal calibration function is uniquely

characterized as $C(p) = \mathbb{E}[p^* \mid \hat{p} = p]$ (Proposition 7), and our main theorem gives it in closed form as a sigmoid integral over the Gaussian posterior (Theorem 11). Second, we show that the MacKay approximation to this optimal estimator takes the same functional form as temperature scaling, and we establish the correspondence $T = \sqrt{1 + \pi\tau^2/8}$ between the temperature parameter T and the noise variance τ^2 (Proposition 14; this isomorphism holds in the limit of a flat, noninformative prior in logit space). This provides a theoretical interpretation of temperature scaling from the standpoint of a logit-noise generative model. We further prove that the Bayes-optimal estimator reduces to the same posterior mean under both the squared-error loss and the Brier score (Proposition 19), providing a unified theoretical basis for squared-type losses.

In addition, we verify the above theoretical results numerically under two conditions, a synthetic-data condition and a learner-mediated condition (Section 7).

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 formulates the model and problem setting. Section 4 presents the main theorem, and Section 5 gives the extension to the Brier score. Section 6 presents an analysis under a concrete prior distribution together with an inverse-problem interpretation, and Section 7 verifies the theoretical behavior numerically. Section 8 discusses the results, and Section 9 concludes the paper.

2 Related Work

Calibration: definition and evaluation The calibration of predicted probabilities traces back to the framework of DeGroot and Fienberg (1983). In recent years, binning-based evaluation methods centered on the Expected Calibration Error (ECE) (Naeini et al., 2015) have been widely used. However, ECE has been shown to carry a bias that depends on the choice of binning (Kumar et al., 2019), which has motivated kernel-density-based methods (Zhang et al., 2020) and regression-based evaluation (Vaicenavicius et al., 2019). Toward mitigating this estimation bias, Roelofs et al. (2022) propose a debiased estimator, and Chidambaram et al. (2024) theoretically expose structural flaws of ECE through a logit-smoothing analysis. Futami and Fujisawa (2024) further develop an information-theoretic generalization analysis of ECE, providing a theoretical characterization of its finite-sample behavior.

Post-hoc calibration methods Platt scaling (Platt et al., 1999) applies logistic regression to a classifier's output, and temperature scaling (Guo et al., 2017) is positioned as a special case of it. Guo et al. (2017) show that deep learning models exhibit pronounced miscalibration and demonstrate the effectiveness of temperature scaling experimentally. Tang et al. (2024) propose Neural Clamping, which jointly learns an input perturbation together with temperature scaling and recovers temperature scaling as the special case of zero input perturbation. Other proposals include isotonic regression (Zadrozny & Elkan, 2002), histogram binning (Zadrozny & Elkan, 2001), and Bayesian binning (Naeini et al., 2014; 2015). Wenger et al. (2020) propose nonparametric multi-class calibration based on a latent Gaussian process. Most recently, Tao et al. (2024) propose Consistency Calibration, which exploits the consistency of outputs under perturbations of neighboring data.

Calibration during training Apart from post-hoc methods, approaches that suppress miscalibration by controlling the shape of the output distribution during training have also been studied extensively. The confidence penalty of Pereyra et al. (2017) indirectly mitigates overconfidence by penalizing low entropy in the output distribution, and it has been shown to be equivalent to label smoothing, as analyzed by Müller et al. (2019), up to the direction of the KL divergence. Mukhoti et al. (2020) show, both experimentally and analytically, that focal loss naturally yields better-calibrated predictions than ordinary cross-entropy, and Thulasidasan et al. (2019) report that the Mixup data-augmentation technique reduces calibration error. A theoretical analysis of the conditions and mechanism underlying the effect of Mixup is given by Zhang et al. (2022).

Theoretical foundations of calibration The relationship between calibration and proper scoring rules is discussed in detail by Bröcker (2009). The calibration–refinement decomposition of the Brier score traces back to Murphy (1973), who showed that calibration error separates out as one component of the Brier score.

Kull and Flach (2015) refine the decomposition of proper scoring rules into four terms, comprising adjustment loss, grouping loss, post-adjustment calibration loss, and irreducible loss, separating the deviation from the true posterior probability (grouping loss) from the bias of the average score (adjustment loss). However, to the best of our knowledge, a framework that derives the Bayes optimality of temperature scaling under an explicitly assumed generative model has not been fully worked out.

Positioning of this study The existing work reviewed above deals primarily with either (i) methods for evaluating calibration error, or (ii) post-hoc calibration methods based on empirical optimization on a validation set. This study differs in that it explicitly models the generative process of the predicted probability and derives an optimal calibration function grounded in Bayesian decision theory. Specifically, we show that, under the generative assumption of additive Gaussian noise in logit space, a temperature-scaling-type correction function arises naturally as an approximation to the Bayes-optimal estimator. This also distinguishes our analysis from Bayesian calibration methods that place a flexible, nonparametric prior over the calibration map itself, such as the latent-Gaussian-process model of Wenger et al. (2020): rather than fitting a general nonparametric corrector from data, we adopt a simple, low-dimensional generative assumption for the express purpose of obtaining an analytically tractable Bayes-optimal estimator and an interpretable closed-form correspondence with a widely used heuristic. The resulting characterization trades the flexibility of a nonparametric fit for an explicit account of when and why temperature scaling should be expected to work.

We note that Jang et al. (2021) add Gaussian or uniform noise to logits at inference time and propose a calibration method based on label-flip frequency; whereas they use noise as an internal operation of the correction algorithm, we treat noise as the generative model of the observed logit. As a contemporaneous line of work, Clarté et al. (2023) propose Expectation Consistency based on the Nishimori identity; whereas EC enforces an empirical consistency condition (equality between average confidence and average accuracy), our study differs in that it derives the correction function itself as a posterior mean under a generative model.

3 Problem Formulation

3.1 Basic Definitions

Definition 1 (Sigmoid and Logit Functions)**.** We define the sigmoid function $\sigma : \mathbb{R} \to (0, 1)$ and the logit function $\mathrm{logit} : (0, 1) \to \mathbb{R}$ by

$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}, \tag{1}$$

$$\mathrm{logit}(p) = \log \frac{p}{1 - p}. \tag{2}$$

These are mutual inverses: $\sigma(\mathrm{logit}(p)) = p$ and $\mathrm{logit}(\sigma(\theta)) = \theta$.

Definition 2 (Calibration Function)**.** For a classifier's output probability $\hat{p}$, we define the calibration function $C : [0, 1] \to [0, 1]$ by

$$C(p) = \mathbb{E}[p^* \mid \hat{p} = p], \tag{3}$$

where $p^* = \Pr(Y = 1 \mid X)$ is the true posterior probability. A classifier is said to be perfectly calibrated if $C(p) = p$ holds for all $p \in [0, 1]$.

Remark 3 (Population Function versus Estimator)**.** The function C in Definition 2 is a population-level function determined by the joint distribution of the random variables $(p^*, \hat{p})$. In this paper, we treat calibration as the problem of estimating p^* from the observed predicted probability $\hat{p}$, and we design the calibrated output as a correction function of the form $\tilde{p} = \psi(\hat{p})$. As we show later, under the squared-error loss the optimal correction function ψ^* satisfies $\psi^*(p) = C(p)$ (Proposition 7).

3.2 The Logit-Noise Model

Assumption 1 (Logit-Noise Model)**.** The observed predicted probability $\hat{p}$ follows a model in which Gaussian noise is added in logit space to the true posterior probability p^*:

$$\theta^* = \mathrm{logit}(p^*), \tag{4}$$

$$\hat{\theta} = \theta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \tau^2), \tag{5}$$

$$\hat{p} = \sigma(\hat{\theta}), \tag{6}$$

where $\tau > 0$ is the noise standard deviation, $p^* = \Pr(Y = 1 \mid X)$ is the true posterior probability depending on the input X, and $\theta^* = \mathrm{logit}(p^*)$ is the corresponding latent logit.

Remark 4 (Empirical Basis for Assumption 1)**.** As theoretical background for this assumption, a related study (Kumoi et al., 2023) builds an analysis framework for the performance of noisy binary classifiers (classifiers whose estimated posterior probability is corrupted by error), and its treatment of an estimated posterior probability with noise is consistent with the logit-noise model of this paper. Moreover, Ferianc et al. (2023) systematically confirm that noise injected at the training and inference stages contributes to calibration, and the logit-noise assumption of this paper can be regarded as an explicit generative-model formulation of this empirical regularity. Building on this observation, we focus on clarifying the theoretical properties of the optimal calibration function based on conditional Bayes risk minimization. The residual diagnostics in our numerical verification (Section 7, Figure 4) confirm that the zero-mean, variance-τ^2 structure of the assumption holds even under a learner-mediated condition. On the other hand, the same figure shows that normality and homoscedasticity with respect to θ^* hold only partially, which is a factor that constrains the effective range of temperature-scaling-type correction.

Assumption 2 (Prior Distribution)**.** The true logit $\theta^* = \mathrm{logit}(p^*)$ follows a prior distribution $g(\theta)$. Its correspondence with the prior distribution $\pi(p^*)$ in probability space is given, via the Jacobian of the change of variables, by

$$g(\theta) = \pi(\sigma(\theta)) \cdot \sigma(\theta)(1 - \sigma(\theta)). \tag{7}$$

3.3 Loss Functions and Bayes Optimality

We write the calibrated output probability as $\tilde{p} = \psi(\hat{p})$. We consider two loss functions in this study.

Definition 5 (Loss Functions)**.** We define the squared-error loss and the Brier score loss for an estimator $\tilde{p}$ by

$$L_{\mathrm{SE}}(\tilde{p}, p^*) = (\tilde{p} - p^*)^2, \tag{8}$$

$$L_{\mathrm{BS}}(\tilde{p}, Y) = (\tilde{p} - Y)^2, \tag{9}$$

where p^* is the true posterior probability and $Y \in \{0, 1\}$ is the true label.

Definition 6 (Conditional Bayes Risk and Bayes-Optimal Estimator)**.** Given the observation $\hat{\theta} = \mathrm{logit}(\hat{p})$, we define the conditional Bayes risk for the squared-error loss by

$$R_{\mathrm{SE}}(\tilde{p} \mid \hat{\theta}) = \mathbb{E}\left[(\tilde{p} - p^*)^2 \mid \hat{\theta}\right]. \tag{10}$$

We call the value of $\tilde{p}$ that minimizes this quantity for each $\hat{\theta}$ the Bayes-optimal estimator.

The following proposition shows that the optimal correction function coincides with the calibration function. This result requires no generative-model assumption and holds in general.

Proposition 7 (Coincidence of the Optimal Correction Function and the Calibration Function)**.** Fix an arbitrary joint distribution of the random variables $(p^*, \hat{p})$. Under the squared-error loss, for each $p \in [0, 1]$ the value $\psi^*(p)$ of the correction function $\psi^* : [0, 1] \to [0, 1]$ that minimizes the conditional Bayes risk $\mathbb{E}[(\psi(p) - p^*)^2 \mid \hat{p} = p]$ is uniquely given by

$$\psi^*(p) = \mathbb{E}[p^* \mid \hat{p} = p] = C(p). \tag{11}$$

Proof. Fix $\hat{p} = p$. Completing the square gives

$$\mathbb{E}[(\tilde{p} - p^*)^2 \mid \hat{p} = p] = \mathrm{Var}(p^* \mid \hat{p} = p) + \big(\tilde{p} - \mathbb{E}[p^* \mid \hat{p} = p]\big)^2.$$

The first term on the right-hand side does not depend on $\tilde{p}$, so the minimizer is uniquely $\tilde{p} = \mathbb{E}[p^* \mid \hat{p} = p]$. □

3.4 Components of the Post-Correction Residual

To streamline the subsequent discussion and the numerical verification in Section 7, we discuss the residual $\tilde{p} - p^*$, i.e., the difference between the output of a temperature-scaling-type correction $\tilde{p} = \sigma(\mathrm{logit}(\hat{p})/T)$ (the temperature parameter T is formally introduced in Proposition 14) and the true posterior probability p^*, in terms of the following three conceptual components, classified according to their source.

Definition 8 (Components of the Post-Correction Residual)**.** We distinguish the sources of the post-correction residual under a temperature-scaling-type correction as follows.

(I-a) Approximation error. The structural error incurred when the Bayes-optimal estimator $\mathbb{E}_{Z\sim\mathcal{N}(\mu,v^2)}[\sigma(Z)]$ is replaced by $\sigma(\mu/\sqrt{1+\pi v^2/8})$. This originates from Lemma 13 and is intrinsic to the theoretical framework itself.

(I-b) Parameter estimation error. The finite-sample error incurred when the noise level τ is estimated from calibration data as $\hat{\tau}$. It vanishes as the number of calibration samples increases.

(I-c) Model misspecification error. Error arising from deviations from Assumption 1 (additive Gaussian noise, homoscedasticity, normality). This includes non-normality of the residuals, systematic bias depending on θ^*, and heteroscedasticity.

Remark 9 (Meaning and Limitations of the Decomposition)**.** Definition 8 does not strictly decompose the squared post-correction residual into a sum of three components; it is a set of labels for conceptually classifying the sources of error. Under conditions that satisfy Assumption 1 by construction, such as synthetic logit data, (I-c) vanishes by definition and the residual is explained solely by (I-a) and (I-b). With data mediated by a learner, (I-c) is generally nonzero and is observed as a component that temperature scaling cannot fully correct. In this paper we use this vocabulary as a diagnostic tool for delineating the range of validity of the theory.

We note that this (I-a)/(I-b)/(I-c) classification separates the post-correction residual according to its source (approximation, estimation, or model misspecification), which serves a different purpose from the calibration–refinement decomposition of Murphy (1973), which decomposes the score itself based on its statistical properties, or the four-term decomposition of Kull and Flach (2015). Whereas these classical decompositions split the expectation of a fixed score into components, the classification in this paper is distinctive in that it delineates the origin of the residual relative to a specific generative model (Assumption 1). In particular, (I-c) (model misspecification error) in this paper conceptually overlaps with the grouping loss, which measures deviation from the true posterior probability, but the two are not identical, since (I-c) refers specifically to a deviation from the generative structure of Assumption 1 itself.

4 Main Results

We first give an integral representation of the calibration function for a general prior (Section 4.1), then present the closed-form main theorem under a Gaussian prior (Section 4.2), and finally establish the isomorphism with temperature scaling (Section 4.3).

4.1 Integral Representation of the Calibration Function

Proposition 10 (Integral Representation of the Calibration Function)**.** Under Assumptions 1 and 2, the calibration function admits the integral representation

$$C(p) = \frac{\displaystyle\int_{-\infty}^{\infty} \sigma(\theta) \cdot \varphi\left(\frac{\mathrm{logit}(p) - \theta}{\tau}\right) \cdot g(\theta)\, d\theta}{\displaystyle\int_{-\infty}^{\infty} \varphi\left(\frac{\mathrm{logit}(p) - \theta}{\tau}\right) \cdot g(\theta)\, d\theta}, \tag{12}$$

where $\varphi(\cdot)$ is the standard normal density, and the factor $(1/\tau)$ common to the numerator and denominator has been canceled.

Proof. By Assumption 1, the density of $\hat{\theta}$ conditional on θ^* is $f_{\hat{\theta}|\theta^*}(\hat{\theta} \mid \theta^*) = (1/\tau)\varphi((\hat{\theta} - \theta^*)/\tau)$. Using the Jacobian $|d\mathrm{logit}(p)/dp| = 1/(p(1-p))$ of the change of variables $p = \sigma(\hat{\theta})$, the likelihood of p conditional on p^* is given by

$$f(p \mid p^*) = \frac{1}{\tau \cdot p(1-p)} \cdot \varphi\left(\frac{\mathrm{logit}(p) - \mathrm{logit}(p^*)}{\tau}\right). \tag{13}$$

By Bayes' theorem and Proposition 7, $C(p) = \mathbb{E}[p^* \mid \hat{p} = p] = \int_0^1 q\, \pi(q \mid p)\, dq$. Substituting (13) into the posterior distribution, canceling the common factor $(1/(\tau \cdot p(1-p)))$ in the numerator and denominator, and changing variables via $\theta = \mathrm{logit}(q)$, with $g(\theta)$ as in Assumption 2, yields (12). □

4.2 Bayes-Optimal Calibration Function (Main Theorem)

The following main theorem shows that, under a Gaussian prior assumption, the Bayes-optimal calibration function is expressed as a sigmoid integral over a Gaussian posterior distribution.

Assumption 3 (Gaussian Prior Distribution)**.** $\theta^* \sim \mathcal{N}(m, s^2)$, where m is the prior center of the logit and s^2 represents the prior uncertainty. Conjugacy with the Gaussian likelihood (Assumption 1) ensures that the posterior distribution is Gaussian in closed form (see Theorem 11).

Theorem 11 (Bayes-Optimal Calibration Function)**.** Under the logit-noise model (Assumption 1) and the Gaussian prior (Assumption 3), the Bayes-optimal calibration function for the squared-error loss L_{SE} is uniquely given by

$$\tilde{p}_{\mathrm{Bayes}}(\hat{\theta}) = \mathbb{E}[\sigma(\theta^*) \mid \hat{\theta}] = \mathbb{E}_{Z \sim \mathcal{N}(\mu_{\mathrm{post}},\, \sigma^2_{\mathrm{post}})}[\sigma(Z)], \tag{14}$$

where the posterior parameters are

$$\sigma^2_{\mathrm{post}} = \left(\frac{1}{s^2} + \frac{1}{\tau^2}\right)^{-1}, \qquad \mu_{\mathrm{post}} = \sigma^2_{\mathrm{post}} \left(\frac{m}{s^2} + \frac{\hat{\theta}}{\tau^2}\right), \tag{15}$$

and, since $\hat{p} = \sigma(\hat{\theta})$, substituting $\hat{\theta} = \mathrm{logit}(\hat{p})$ yields $\tilde{p}_{\mathrm{Bayes}}(\hat{p})$.

Proof. We proceed in three steps.

Step 1 (posterior distribution of θ^*) By Assumption 1, the likelihood is $p(\hat{\theta} \mid \theta^*) = (1/\tau)\varphi((\hat{\theta} - \theta^*)/\tau)$. By conjugacy with the Gaussian prior $\theta^* \sim \mathcal{N}(m, s^2)$, we have

$$\theta^* \mid \hat{\theta} \sim \mathcal{N}(\mu_{\mathrm{post}},\, \sigma^2_{\mathrm{post}}), \tag{16}$$

where μ_{post} and σ^2_{post} are given by (15).

Step 2 (optimality under the squared-error loss) By the same completion-of-the-square argument as in Proposition 7, the minimizer of the conditional Bayes risk is uniquely $\tilde{p}_{\mathrm{Bayes}}(\hat{\theta}) = \mathbb{E}[p^* \mid \hat{\theta}]$.

Step 3 (application to the logit-noise model) Combining $p^* = \sigma(\theta^*)$ with the result of Step 1,

$$\tilde{p}_{\text{Bayes}}(\hat{\theta}) = \mathbb{E}[\sigma(\theta^*) \mid \hat{\theta}] = \mathbb{E}_{Z\sim\mathcal{N}(\mu_{\text{post}},\,\sigma^2_{\text{post}})}[\sigma(Z)]. \qquad \square$$

Remark 12 (Limit of a Flat Prior)**.** Letting $s^2 \to \infty$ in Assumption 3 corresponds to a flat (noninformative) prior $g(\theta) \propto 1$ in logit space. The posterior parameters then degenerate to $\mu_{\text{post}} \to \hat{\theta}$ and $\sigma^2_{\text{post}} \to \tau^2$, giving

$$\tilde{p}_{\text{Bayes}}(\hat{\theta}) = \mathbb{E}_{Z\sim\mathcal{N}(\hat{\theta},\,\tau^2)}[\sigma(Z)]. \tag{17}$$

The prior $g(\theta) \propto 1$ is improper (not normalizable), but because the likelihood is Gaussian, the posterior $p(\theta^* \mid \hat{\theta}) = \mathcal{N}(\hat{\theta}, \tau^2)$ is properly normalized.

4.3 Isomorphism with Temperature Scaling

The Bayes-optimal estimator of Theorem 11 is the Gaussian-sigmoid integral $\mathbb{E}[\sigma(Z)]$ ($Z \sim \mathcal{N}(\mu, \sigma^2)$), which admits no closed form. We establish the correspondence in functional form with temperature scaling through an approximation of this Gaussian-sigmoid integral. The following lemma states the classical approximation introduced by MacKay (1992a; 1992b) for marginalization in Bayesian neural networks, and it is the key to the temperature-scaling isomorphism of this paper.

Lemma 13 (MacKay Approximation of the Gaussian-Sigmoid Integral)**.** For $Z \sim \mathcal{N}(\mu, v^2)$,

$$\mathbb{E}[\sigma(Z)] \approx \sigma\left(\frac{\mu}{\sqrt{1 + \pi v^2/8}}\right). \tag{18}$$

Proof. We use the probit approximation $\sigma(z) \approx \Phi(\lambda z)$, where Φ is the standard normal distribution function and $\lambda = \sqrt{\pi/8}$. For $Z \sim \mathcal{N}(\mu, v^2)$ we have $\lambda Z \sim \mathcal{N}(\lambda\mu, \lambda^2 v^2)$, and

$$\mathbb{E}[\sigma(Z)] \approx \mathbb{E}[\Phi(\lambda Z)] = \Phi\left(\frac{\lambda\mu}{\sqrt{1 + \lambda^2 v^2}}\right) \approx \sigma\left(\frac{\mu}{\sqrt{1 + \pi v^2/8}}\right).$$

The second equality follows from the integral formula for the probit function against a Gaussian distribution. The final step applies $\Phi(\lambda z) \approx \sigma(z)$ once more. $\square$

Proposition 14 (Isomorphism with Temperature Scaling)**.** Under the MacKay approximation (Lemma 13) applied to the flat-prior limit of Theorem 11 ((17) in Remark 12), the Bayes-optimal estimator is expressed as

$$\tilde{p}_{\text{Bayes}}(\hat{p}) \approx \sigma\left(\frac{\text{logit}(\hat{p})}{T}\right), \qquad T = \sqrt{1 + \frac{\pi\tau^2}{8}}. \tag{19}$$

This is the same functional form as temperature scaling with temperature parameter T, and its correspondence with the noise variance τ^2 is given by

$$\tau^2 = \frac{8}{\pi}(T^2 - 1). \tag{20}$$

Proof. By Remark 12, $\tilde{p}_{\text{Bayes}}(\hat{p}) = \mathbb{E}_{Z\sim\mathcal{N}(\text{logit}(\hat{p}),\,\tau^2)}[\sigma(Z)]$. Applying Lemma 13 with $\mu = \text{logit}(\hat{p})$ and $v^2 = \tau^2$ gives

$$\tilde{p}_{\text{Bayes}}(\hat{p}) \approx \sigma\left(\frac{\text{logit}(\hat{p})}{\sqrt{1 + \pi\tau^2/8}}\right) = \sigma\left(\frac{\text{logit}(\hat{p})}{T}\right).$$

Equation (20) follows immediately by solving the definition of T for τ^2. $\square$

Remark 15 (Scope of the MacKay Approximation)**.** The isomorphism of Proposition 14 depends on the MacKay approximation. This approximation tends to incur larger error when v (the flat-prior limit of τ) is large, or when μ (i.e., $\text{logit}(\hat{p})$) is far from 0. To assess the approximation, including the strongly

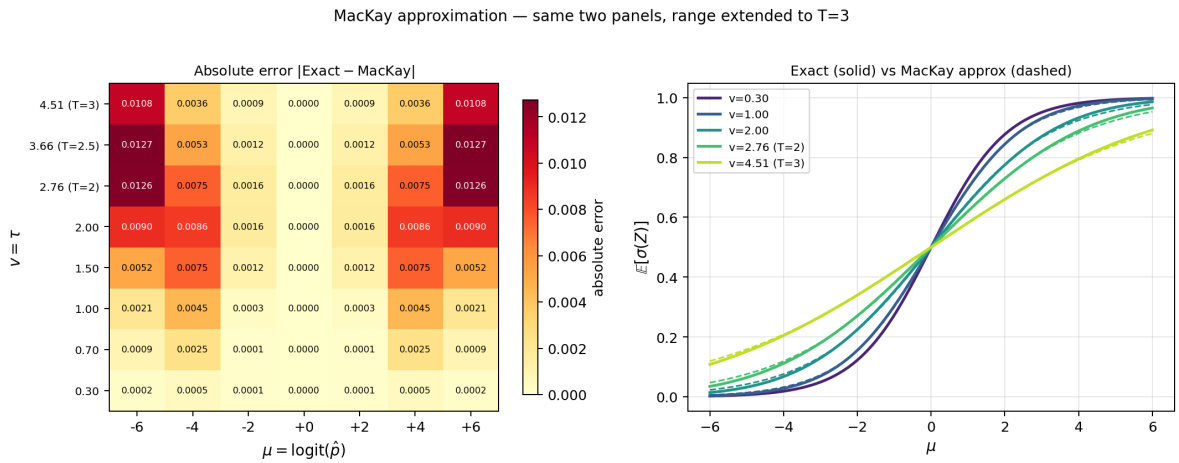

Figure 1: Grid evaluation of the absolute error of the MacKay approximation, $|\mathbb{E}_{Z\sim\mathcal{N}(\mu,v^2)}[\sigma(Z)] - \sigma(\mu/\sqrt{1+\pi v^2/8})|$ (the left grid comprises 7 points $\mu \in \{-6,-4,\ldots,6\}$ and 8 points $v = \tau \in \{0.3, 0.7, 1.0, 1.5, 2.0, 2.76, 3.66, 4.51\}$; extended to temperature $T \leq 3$). **Left**: error heatmap in (μ, v) space (values are absolute error; the temperatures $T = 2, 2.5, 3$ corresponding to the vertical axis $v = \tau$ are also annotated). The error is zero at $\mu = 0$ and increases with high temperature and large $|\mu|$; the maximum within this grid is approximately 0.013. **Right**: comparison of the exact value (solid) and the MacKay approximation (dashed). The two curves nearly coincide where v is small, but visibly deviate from each other at $v = 4.51$ ($T = 3$). This error component corresponds to (I-a) in Definition 8 and delineates the structural limitation of the temperature-scaling isomorphism (Proposition 14) of this paper.

miscalibrated regime (large T, i.e., large τ) that is practically important for temperature scaling, we evaluate the absolute error $|\mathbb{E}[\sigma(Z)] - \sigma(\mu/\sqrt{1+\pi v^2/8})|$ on a grid over $\mu \in [-6, 6]$, $v = \tau \in [0.3, 4.51]$ (temperature $T = \sqrt{1+\pi\tau^2/8} \in [1.0, 3.0]$), shown in Figure 1. The error is structurally zero at $\mu = 0$ and is bounded in μ (tending to 0 as $|\mu| \to \infty$). The peak of the error shifts toward large $|\mu|$ (extremely overconfident predictions) as v increases, and the worst-case absolute error, maximized over μ, is approximately 0.009 at $T = 1.5$, 0.013 at $T = 2$, and 0.015 at $T = 3$: it increases monotonically with temperature but does not diverge (these are worst-case values obtained by extending μ beyond the display range $|\mu| \leq 6$ of Figure 1; the maximum within the displayed grid is approximately 0.013). That is, even over the practical temperature range ($T = 2$ to 3), the approximation error remains within about 1.5% on the probability scale. This approximation error corresponds to (I-a) in Definition 8 and remains a structural component that does not vanish even under conditions where the assumption holds by construction (quantified in Section 7.2). The results of this paper therefore state that a temperature-scaling-type form arises naturally as an approximation to the Bayes-optimal estimator, not that the two estimation procedures are identical. Temperature scaling is typically a procedure that empirically optimizes T on a validation set, whereas what this paper provides is the correspondence in functional form.

Corollary 16 (Approximation under a Gaussian Prior)**.** Applying Lemma 13 to $Z \sim \mathcal{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ of Theorem 11 gives

$$\tilde{p}_{\text{Bayes}}(\hat{\theta}) \approx \sigma\left(\frac{\mu_{\text{post}}}{\sqrt{1+\pi\sigma^2_{\text{post}}/8}}\right). \tag{21}$$

Substituting $\hat{\theta} = \text{logit}(\hat{p})$ into (15) gives the approximate form of $\tilde{p}_{\text{Bayes}}(\hat{p})$.

5 Extension to the Brier Score

We show that, even when the loss function is changed to the Brier score, the Bayes-optimal estimator is identical to that under the squared-error loss. This coincidence arises because the Brier score is a squared-type, strictly proper scoring rule.

Lemma 17 (Decomposition of the Brier Score)**.** For $Y \mid p^* \sim \text{Bernoulli}(p^*)$,

$$\mathbb{E}[(\tilde{p} - Y)^2 \mid p^*] = (\tilde{p} - p^*)^2 + p^*(1 - p^*). \tag{22}$$

Proof. $\mathbb{E}[(\tilde{p} - Y)^2 \mid p^*] = p^*(1 - \tilde{p})^2 + (1 - p^*)\tilde{p}^2 = (\tilde{p} - p^*)^2 + p^*(1 - p^*)$. □

Lemma 18 (Tower Property of Conditional Expectation)**.** Assume $Y \perp\!\!\!\perp \hat{p} \mid X$ (Y and $\hat{p}$ are independent conditional on X). Then

$$\mathbb{E}[Y \mid \hat{p} = p] = \mathbb{E}[p^* \mid \hat{p} = p] = C(p). \tag{23}$$

Proof. By the law of iterated expectations, $\mathbb{E}[Y \mid \hat{p} = p] = \mathbb{E}[\mathbb{E}[Y \mid X, \hat{p}] \mid \hat{p} = p] = \mathbb{E}[\mathbb{E}[Y \mid X] \mid \hat{p} = p] = \mathbb{E}[p^* \mid \hat{p} = p]$. □

Proposition 19 (Bayes Optimality under the Brier Score)**.** Under the Brier score loss $L_{\text{BS}}(\tilde{p}, Y) = (\tilde{p} - Y)^2$, the Bayes-optimal estimator that minimizes the conditional Bayes risk $R_{\text{BS}}(\tilde{p} \mid \hat{\theta}) = \mathbb{E}[(\tilde{p} - Y)^2 \mid \hat{\theta}]$ given the observation $\hat{\theta}$ coincides with that of the squared-error loss, and is given by

$$\tilde{p}_{\text{Brier}}(\hat{\theta}) = \mathbb{E}[p^* \mid \hat{\theta}]. \tag{24}$$

In particular, under the logit-noise model (Assumption 1) and the flat prior (Remark 12), we have

$$\tilde{p}_{\text{Brier}}(\hat{\theta}) = \mathbb{E}_{Z \sim \mathcal{N}(\hat{\theta},\, \tau^2)}[\sigma(Z)], \tag{25}$$

and the temperature-scaling approximation of Proposition 14 applies directly.

Proof. Under Assumption 1, the noise ε is independent of Y given θ^*, so $Y \perp\!\!\!\perp \hat{\theta} \mid p^*$; this is the same conditional independence that underlies Lemma 18. Hence, by the law of iterated expectations, $\mathbb{E}[(\tilde{p} - Y)^2 \mid p^*, \hat{\theta}] = \mathbb{E}[(\tilde{p} - Y)^2 \mid p^*]$, and Lemma 17 gives

$$\begin{aligned} R_{\text{BS}}(\tilde{p} \mid \hat{\theta}) &= \mathbb{E}[\mathbb{E}[(\tilde{p} - Y)^2 \mid p^*] \mid \hat{\theta}] \\ &= \mathbb{E}[(\tilde{p} - p^*)^2 \mid \hat{\theta}] + \mathbb{E}[p^*(1 - p^*) \mid \hat{\theta}]. \end{aligned}$$

The second term does not depend on $\tilde{p}$, so the minimization reduces to $\min_{\tilde{p}} \mathbb{E}[(\tilde{p} - p^*)^2 \mid \hat{\theta}]$, which gives the same solution $\mathbb{E}[p^* \mid \hat{\theta}]$ as in Theorem 11. The expression under the flat prior follows from Remark 12. □

Remark 20 (Relation to Proper Scoring Rules)**.** The Brier score is an example of a strictly proper scoring rule (Gneiting & Raftery, 2007; Buja et al., 2005). In the framework of Gneiting and Raftery (2007), the expectation of a strictly proper scoring rule is uniquely minimized at the true probability distribution, and the minimizer is a posterior mean under the corresponding Bregman divergence. Buja et al. (2005) further systematize proper scoring rules in terms of a weight function, a link function, and a Bregman distance, providing a framework that treats log loss and the squared-error loss from a unified viewpoint. Proposition 19 gives a concrete instance, in the logit-noise model, of the fact that for squared-type proper scoring rules the minimizer reduces to a posterior mean.

6 Examples and Interpretation

We illustrate the main theorem with a concrete example, and present auxiliary results that give an intuitive account of the shrinkage mechanism discussed in Section 8. Section 6.1 derives properties of the calibration function under a uniform prior, and Section 6.2 reinterprets the calibration problem as an inverse problem.

6.1 Analysis under a Uniform Prior

Corollary 21 (Calibration Function under a Uniform Prior)**.** When the true posterior probability p^* follows a uniform distribution on $(0,1)$ ($\pi(p^*) = 1$), Assumption 2 gives $g(\theta) = \sigma(\theta)(1-\sigma(\theta))$, and Proposition 10 yields

$$C(p) = \frac{\displaystyle\int_{-\infty}^{\infty} \sigma(\theta) \cdot \varphi\left(\frac{\mathrm{logit}(p) - \theta}{\tau}\right) \cdot \sigma(\theta)(1-\sigma(\theta))\, d\theta}{\displaystyle\int_{-\infty}^{\infty} \varphi\left(\frac{\mathrm{logit}(p) - \theta}{\tau}\right) \cdot \sigma(\theta)(1-\sigma(\theta))\, d\theta}. \tag{26}$$

Proposition 22 (Symmetry and Special Values under a Uniform Prior)**.** Under a uniform prior distribution, the calibration function satisfies the following properties:

1. (Invariance at the midpoint) $C(0.5) = 0.5$.
2. (Reflection symmetry) $C(1-p) = 1 - C(p)$ for all $p \in (0,1)$.
3. (Limiting values) $\lim_{p\to 0} C(p) = 0$ and $\lim_{p\to 1} C(p) = 1$.

Proof. (1) Since $\mathrm{logit}(0.5) = 0$, $\varphi(-x) = \varphi(x)$, and $\sigma(-\theta) = 1 - \sigma(\theta)$, the function $f(\theta) = \varphi(\theta/\tau) \cdot \sigma(\theta)(1 - \sigma(\theta))$ is even. Using the change of variables $s = -\theta$ gives $\int \sigma(\theta) f(\theta)\, d\theta = \int (1-\sigma(s)) f(s)\, ds$, so the numerator equals 1/2 of the denominator, hence $C(0.5) = 0.5$.

(2) Since $\mathrm{logit}(1-p) = -\mathrm{logit}(p)$ and $\sigma(-\theta)(1-\sigma(-\theta)) = \sigma(\theta)(1-\sigma(\theta))$, the denominator of $C(1-p)$ coincides with that of $C(p)$, and its numerator equals the denominator minus the numerator of $C(p)$.

(3) As $p \to 0$, $\mathrm{logit}(p) \to -\infty$, so the likelihood concentrates in the region $\theta \to -\infty$, where $\sigma(\theta) \to 0$; hence $C(p) \to 0$. By reflection symmetry (2), $\lim_{p\to 1} C(p) = 1$. □

6.2 Reformulation as an Inverse Problem

The calibration problem can be naturally understood as the following pair of forward and inverse problems.

Proposition 23 (Representation as an Inverse Problem)**.** The calibration problem under the logit-noise model is expressed as inverse inference for the following forward problem:

$$\textbf{Forward problem:} \quad \theta^* \xrightarrow{+\varepsilon} \hat{\theta} \xrightarrow{\sigma} \hat{p}, \tag{27}$$

$$\textbf{Inverse problem:} \quad \hat{p} \xrightarrow{\mathrm{logit}} \hat{\theta} \xrightarrow{\text{posterior inference}} \theta^* \mid \hat{\theta} \xrightarrow{\sigma} p^*. \tag{28}$$

Here, the Bayes-optimal estimator is given as the "posterior expectation of the inverse map" into probability space,

$$\tilde{p}_{\mathrm{Bayes}} = \mathbb{E}[\sigma(\theta^*) \mid \hat{\theta}]. \tag{29}$$

Remark 24 (Nonlinearity and the Shrinkage Effect)**.** In solving the inverse problem, the nonlinearity of σ (convex for $\theta < 0$, concave for $\theta > 0$) plays an essential role. In general,

$$\mathbb{E}[\sigma(\theta^*) \mid \hat{\theta}] \neq \sigma(\mathbb{E}[\theta^* \mid \hat{\theta}]) = \sigma(\mu_{\mathrm{post}}),$$

and this discrepancy is the source of the shrinkage effect.

Specifically, when the mass of the conditional distribution concentrates in the concave region $\theta > 0$ (i.e., $\hat{p} > 0.5$), the conditional version of Jensen's inequality gives approximately

$$\mathbb{E}[\sigma(\theta^*) \mid \hat{\theta}] \leq \sigma(\mu_{\mathrm{post}}),$$

so $\hat{p}$ is corrected toward 0.5. When $\hat{p} < 0.5$, the inequality reverses and a similar shrinkage occurs.

7 Numerical Verification

We numerically verify the correction behavior predicted by the main results (Theorem 11 and Propositions 14 and 19) under two data conditions. Section 7.1 lays out the intent of the verification design, Section 7.2 treats the condition under which the assumption holds by construction, Section 7.3 treats the condition under which the assumption holds empirically through a learner, Section 7.4 examines setting dependence and the scope of applicability, Section 7.5 examines the dependence on the number of training samples, and Section 7.6 gives an overall summary.

7.1 Verification Design

The claim of this paper is that, under Assumption 1, a temperature-scaling-type correction arises as an approximation to the Bayes-optimal estimator. The verification must therefore observe both that the theoretical behavior is reproduced under conditions where the assumption holds, and how deviations from the assumption affect correction performance. We accordingly separate the sources of error into (I-a), (I-b), and (I-c) of Definition 8 and set up two experimental conditions.

We emphasize here an important premise common to both conditions below. In both conditions, the true logit $\theta^* = \mathrm{logit}(p^*)$ is analytically known, and the noise standard deviation is estimated from its residual as $\hat{\tau} = \mathrm{std}(\hat{\theta} - \theta^*)$. All numerical verification in this paper therefore relies on the idealized condition that θ^* is observable, and this procedure cannot be applied as is to real data for which θ^* is unavailable. Extending it to real data requires a separately developed method for estimating τ, for example via confidence binning, and we revisit this as a principal limitation of this paper in Section 8.

In the synthetic-data condition (Condition S), we sample the true logit from $\theta^* \sim \mathrm{Uniform}[-B, B]$ and directly generate $\hat{\theta} = \theta^* + \varepsilon$ ($\varepsilon \sim \mathcal{N}(0, \tau^2)$). Because Assumption 1 holds by definition, (I-c) is zero by construction, and the residual is explained solely by (I-a) (the MacKay approximation error) and (I-b) (the finite-sample error of $\hat{\tau}$).

In the learner-mediated condition (Condition L), we generate Gaussian-mixture data with $M = 4$ classes in $D = 4$ dimensions and obtain the predicted probability $\hat{p}$ via logistic regression. Here noise is not injected externally but arises naturally from the fact that the learner cannot perfectly reproduce the true posterior probability, so (I-c) is generally nonzero, and the empirical validity of the assumption must be confirmed through residual diagnostics.

We use a common set of evaluation metrics for both conditions. We compare three correction methods: temperature scaling using the true noise τ (a-true), temperature scaling using the estimated $\hat{\tau}$ (a-est), and nonparametric bin correction, i.e., binning of the predicted value (c). The evaluation metrics are the squared error $\mathrm{MSE}(\tilde{p}, p^*)$, the RMSE between the calibration curve and the theoretical curve, and the RMSE of the post-correction curve from the diagonal. In figures and tables we additionally label three calibration curves: the pre-correction curve (α), the theoretical curve predicted by Proposition 14 with the true τ (β-true), and the post-correction curve obtained by applying (a-true) (γ).

7.2 Verification under the Synthetic-Data Condition (Condition S)

We evaluate at $N = 20{,}000$ (10,000 calibration and 10,000 test samples), $B = 6$, and four levels $\tau \in \{0.3, 0.7, 1.0, 1.5\}$. Table 1 shows the main numerical results (the pre- and post-correction calibration curves are illustrated for $\tau = 1.5$ in Figures 2 and 3).

Table 1 shows the following. First, regarding the effect of (I-b), $\hat{\tau}$ agrees with τ to within a relative error of less than 1% for all τ, and the MSEs of (a-true) and (a-est) coincide within the displayed digits (the difference is on the order of 10^{-5} in the supplementary data). Under a calibration sample size of 10,000, (I-b) can be regarded as effectively zero in practice.

The magnitude of (I-a) is confirmed by the fact that the RMSE of the post-correction curve from the diagonal increases monotonically in τ, from 3.2×10^{-3} at $\tau = 0.3$ to 1.09×10^{-2} at $\tau = 1.5$. This directly demonstrates that the structural limitation of the MacKay approximation noted in Remark 15 manifests as τ increases.

Table 1: Key numerical results for the synthetic-data condition (Condition S, $N = 20{,}000$). $T_{\text{true}} = \sqrt{1 + \pi\tau^2/8}$. $\hat{\tau}$ is the standard deviation of the calibration residuals, and $R(\gamma, \text{diag})$ is the RMSE between the (a-true)-corrected curve and the diagonal. Each value is the mean $\pm$ standard deviation over 50 independent seeds (T_{true} is a deterministic function of τ).

τ	T_{true}	$\hat{\tau}$	MSE_{raw}	$\text{MSE}_{(a\text{-true})}$	$R(\gamma, \text{diag})$
0.3	1.018	0.300 ± 0.002	0.00125 ± 0.00003	0.00124 ± 0.00003	0.00315 ± 0.00070
0.7	1.092	0.700 ± 0.004	0.00665 ± 0.00016	0.00650 ± 0.00015	0.00681 ± 0.00158
1.0	1.180	1.000 ± 0.006	0.01325 ± 0.00031	0.01271 ± 0.00029	0.00850 ± 0.00199
1.5	1.372	1.499 ± 0.009	0.02832 ± 0.00067	0.02623 ± 0.00059	0.01086 ± 0.00207

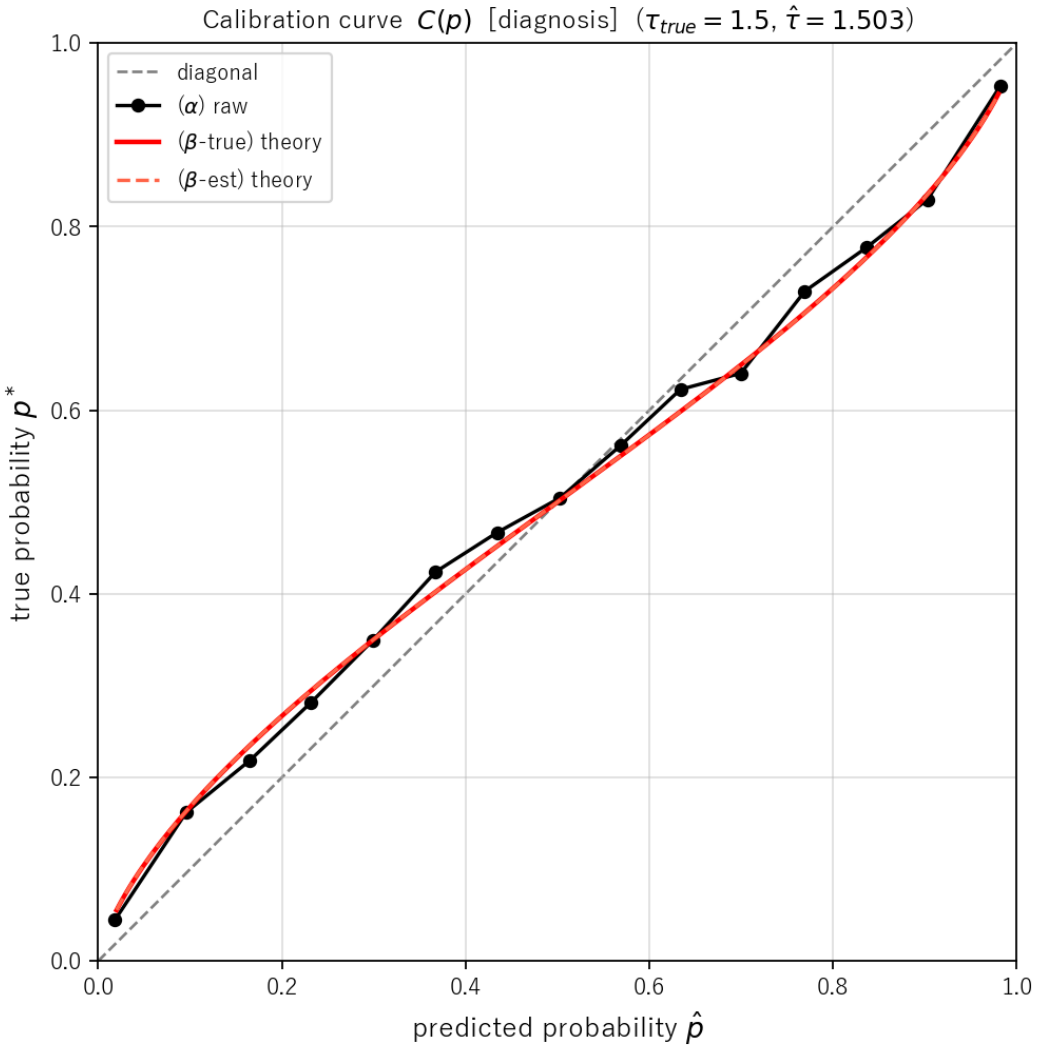

Figure 2: Comparison of the pre-correction curve (α) and the theoretically predicted curve (β-true) at $\tau = 1.5$. Their agreement shows that the functional form of Proposition 14 captures the structure of the empirical data.

We note that Figure 1 shows the pointwise absolute error of the MacKay approximation itself, whereas the present metric $R(\gamma, \text{diag})$ represents that error aggregated into the empirical post-correction calibration

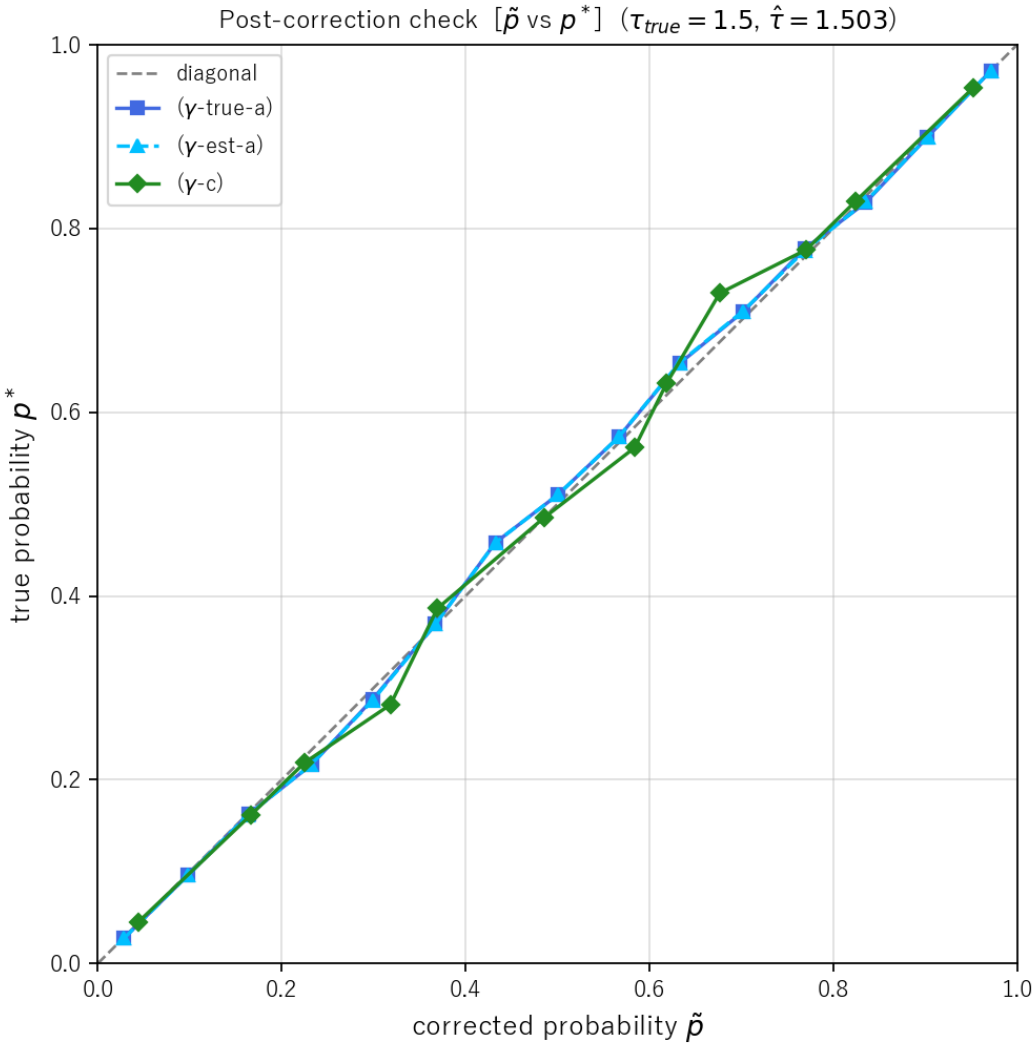

Figure 3: Approach of the post-correction curve (γ) to the diagonal at $\tau = 1.5$. (a-true) and (a-est) nearly coincide, showing that (I-b) is negligible. The residual from the diagonal corresponds to (I-a).

curve; although the two correspond to the same component, their numerical values do not coincide because the aggregation methods differ.

The magnitude of the shrinkage effect (Remark 24) is quantified by the relative difference between $\mathrm{MSE}_{\mathrm{raw}}$ and $\mathrm{MSE}_{(a\text{-true})}$, which increases monotonically from nearly 0% at $\tau = 0.3$ to about 7% at $\tau = 1.5$. This trend is consistent with the theoretical limiting behavior in which the correction approaches the identity map as $\tau \to 0$ and collapses to the constant 0.5 as $\tau \to \infty$.

7.3 Verification under the Learner-Mediated Condition (Condition L)

We generate N samples from a Gaussian mixture with $M = 4$ classes in $D = 4$ dimensions whose class means are the standard basis vectors e_m (var $= 1.0$), convert them into a 2v2 task (classes $0, 1$ versus $2, 3$), and run the experiment with 1,000 training, 3,000 calibration, and 3,000 test samples per class. The true logit $\theta^* = \mathrm{logit}(p^*)$ is analytically computable from the density ratio. We train logistic regression on the training data after applying a resampling procedure that flattens the histogram of θ^* (θ-flattening, 24 bins,

Table 2: Diagnostics of the residual $r = \hat{\theta} - \theta^*$ under the learner-mediated condition (Condition L; calibration data, $N_{\text{cal}} = 12{,}000$). Each value is the mean $\pm$ standard deviation over 50 independent seeds. The "Consistency with assumption" column indicates consistency with Assumption 1.

Metric	Mean $\pm$ SD	Consistency with assumption
$\mathbb{E}[r]$	-0.003 ± 0.026	≈ 0 (holds)
$\text{std}(r) = \hat{\tau}$	0.444 ± 0.008	Holds at the mean/variance level
$\text{skew}(r)$	0.009 ± 0.069	≈ 0 (symmetry holds)
Excess kurtosis $\text{kurt}(r)$	$+2.03 \pm 0.20$	Deviation from normality (heavy tails)
Slope of r vs. θ^*	-0.124 ± 0.039	Deviation from homoscedasticity

Table 3: Comparison of ECE and MSE with and without θ-flattening resampling (Condition L, test data, $N_{\text{test}} = 12{,}000$). (a-est): temperature scaling using the estimated $\hat{\tau}$; (c): bin correction. Each value is the mean $\pm$ 95% CI over 50 independent seeds.

Metric	Method	Without θ-flat.	With θ-flat.
ECE	raw	0.0051 ± 0.0007	0.0078 ± 0.0009
	(a-est)	0.0064 ± 0.0011	0.0062 ± 0.0008
	(c)	0.0026 ± 0.0002	0.0027 ± 0.0002
MSE	raw	0.0074 ± 0.0001	0.0075 ± 0.0001
	(a-est)	0.0074 ± 0.0001	0.0075 ± 0.0001
	(c)	0.0074 ± 0.0001	0.0075 ± 0.0001

keep ratio 0.7), and compute the residual $r = \hat{\theta} - \theta^*$ on the calibration and test data. Table 2 shows the values from the residual diagnostics.

The residual diagnostics show that Assumption 1 holds at the level of mean and variance, while homoscedasticity with respect to θ^* is systematically violated, with the slope of r against θ^* consistently negative across the 50 seeds (-0.124 ± 0.039), and normality is also violated, with an excess kurtosis of $+2.03$ (Figure 4). These are concrete manifestations of (I-c) in Definition 8.

We show the correction performance in Table 3 (mean $\pm$ 95% CI over 50 independent seeds). Without θ-flattening, $\text{ECE}_{\text{raw}} = 0.0051$, and with θ-flattening, $\text{ECE}_{\text{raw}} = 0.0078$; both are small, at or below 0.01, indicating that there is little room for raw correction in this experimental setting. For MSE, there is almost no difference across conditions and methods, at 0.0074–0.0075. The effect of temperature scaling (a-est) is asymmetric across conditions.

A paired t-test on the within-trial ECE difference (a-est $-$ raw) (a one-sample t-test on the per-trial difference; consistency also confirmed with a Wilcoxon signed-rank test) shows a significant decrease of -0.0016 (95% CI ± 0.0010, a relative change of about -21%, two-sided $p = 0.002$) with θ-flattening, whereas without it there is a significant increase of $+0.0013$ (95% CI ± 0.0010, about $+26\%$ relative to the raw ECE, two-sided $p = 0.013$, with 62% of trials worsening). We note that none of the p-values reported in this paper are corrected for multiple comparisons, and we interpret them as descriptive rather than confirmatory indicators. It should also be noted that, in both cases, the absolute magnitudes are small relative to the level of ECE (~ 0.005–0.008), and the trial-to-trial variability is also large. This asymmetry suggests that θ-flattening has the effect of bringing the training distribution of LR closer to a uniform prior, thereby better satisfying the premise of temperature-scaling correction (Assumption 1). We note that (c) (bin correction) consistently gives the smallest ECE (≈ 0.0026) under both conditions. This condition is positioned as a case in which a partial deviation from the assumption (the presence of (I-c)) and the condition-dependence of the room for correction are observed simultaneously.

7.4 Setting Dependence and Scope of Applicability (Condition L)

We presented the main result of this paper under a single setting of Condition L (var = 1.0, 2v2, $M = D = 4$); we now confirm its robustness and scope of applicability by varying the class overlap var and the problem scale $M = D$ (each setting uses 50 independent seeds, mean $\pm$ 95% CI). Because the data-generating process

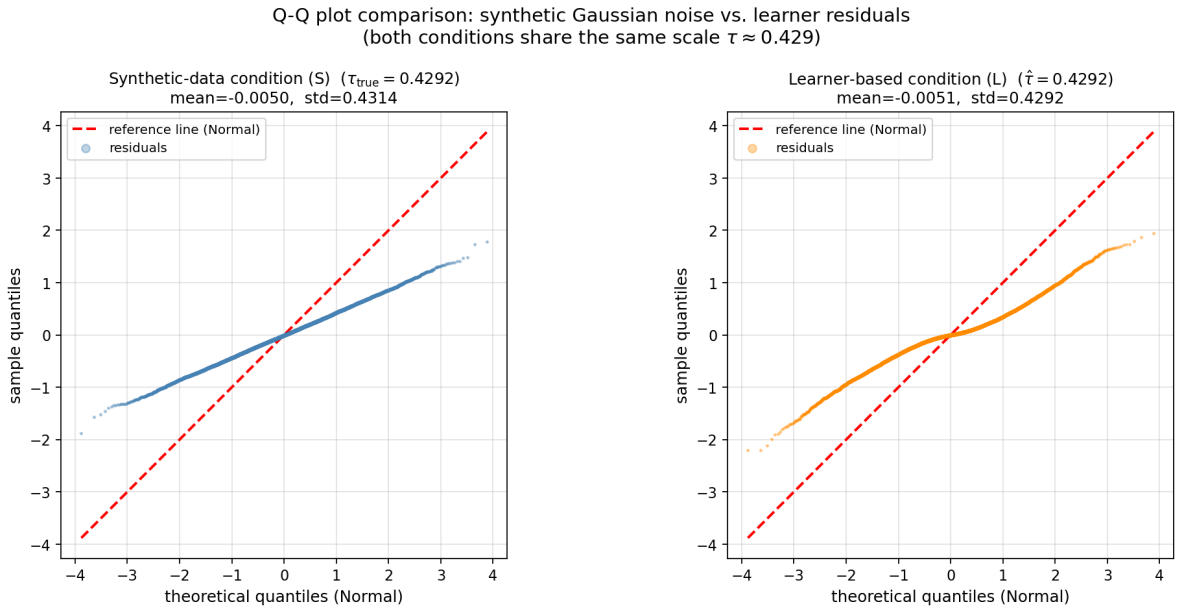

Figure 4: Comparison of residual Q–Q plots at the same $\hat{\tau} \approx 0.43$. Left: the synthetic-data condition (Condition S, in which the assumption holds by definition). Right: the learner-mediated condition (Condition L, in which the assumption holds empirically through a learner). The LR residuals deviate from the normal line, with a flattened center and steeper shoulders, and reach more extreme values in both tails than a Gaussian of the same scale, exhibiting a heavy-tailed (leptokurtic) distribution consistent with the positive excess kurtosis +2.03 in Table 2. This heavy-tailed shape signifies a shape deviation (I-c) from Assumption 1, indicating the existence of a residual component that temperature scaling cannot fully correct.

Table 4: Results of Condition L with varying class overlap var (M=D=4, 2v2, mean ± 95% CI over 50 independent seeds). slope$_{\hat{\theta}}$ is the slope of the regression of $\hat{\theta}$ on θ^* (the assumption holds when this equals 1).

var	slope$_{\hat{\theta}}$	$\hat{\tau}$	ECE$_{\text{raw}}$	ECE$_{(a\text{-est})}$	ECE$_{(c)}$
0.1	0.52	4.68	0.0148 ± 0.0008	0.168 ± 0.002	0.0083 ± 0.0003
0.5	0.80	0.83	0.0093 ± 0.0012	0.0139 ± 0.0015	0.0054 ± 0.0003
1.0	0.87	0.44	0.0078 ± 0.0009	0.0063 ± 0.0008	0.0028 ± 0.0002
2.0	0.92	0.24	0.0069 ± 0.0011	0.0066 ± 0.0010	0.0015 ± 0.0001

places each class mean at a standard basis vector e_m, we do not vary the dimension independently but link it to $M = D$, restricting to even numbers of classes so that the half-and-half binarization is well-defined.

Class overlap and scope of applicability Table 4 shows the results of varying var. As var decreases and the classes separate, the regression slope of $\hat{\theta}$ on θ^* moves away from 1 (0.92 → 0.52), $\hat{\tau}$ and the non-normality of the residuals increase, and Assumption 1 is strongly violated. Correspondingly, the effect of temperature scaling (a-est) transitions from neutral (little room for correction) at var = 2.0, through improvement at var = 1.0, to a breakdown at var = 0.1 that worsens ECE by roughly a factor of 10. In contrast, the nonparametric correction (c) consistently gives the smallest ECE across all values of var. This empirically corroborates the method-selection framework (Section 8.4) of measuring the degree to which the assumption holds via residual diagnostics and using nonparametric correction in regions where it holds poorly. We note that, at var = 0.1, the true logit θ^* becomes extremely large and the linear classifier's

Table 5: Results of Condition L with varying problem scale $M{=}D$ (var=1.0, half-and-half binarization, mean $\pm$ 95% CI over 50 independent seeds).

$M{=}D$	$\text{slope}_{\hat{\theta}}$	ECE_{raw}	$\text{ECE}_{(a\text{-est})}$	$\text{ECE}_{(c)}$
4 (2v2)	0.87	0.0078 ± 0.0009	0.0063 ± 0.0008	0.0028 ± 0.0002
6 (3v3)	0.82	0.0084 ± 0.0011	0.0062 ± 0.0008	0.0022 ± 0.0002
8 (4v4)	0.79	0.0089 ± 0.0010	0.0062 ± 0.0008	0.0019 ± 0.0001

prediction $\hat{\theta}$ is compressed relative to it (slope ≈ 0.5); this is the primary cause of the breakdown and delineates the lower bound of the scope of applicability for this data-generating process.

Problem scale and task Table 5 shows the results for $M = D \in \{4, 6, 8\}$ (half-and-half binarization). Even as the scale increases, (a-est) consistently improves ECE (stable at ≈ 0.0062), and (c) remains the smallest. Temperature-scaling-type correction is robust to problem scale. Changing the binarization to 1v3 ($M = 4$, var $= 1.0$) causes (a-est) to slightly worsen ECE ($0.0073 \pm 0.0009 \to 0.0081 \pm 0.0009$), exhibiting a weak, split-dependent asymmetry of the same kind as the condition without θ-flattening in Table 3.

7.5 Dependence on the Number of Training Samples

We sweep the number of training samples per class $n_{\text{train}} \in \{5, 10, 20, 50, 100, 200, 500, 1000\}$ and examine how the behavior of (I-c) under the learner-mediated condition and the relative merits of the correction methods vary. All other settings are identical to Section 7.3. The results are shown in Figure 5.

For small n_{train}, $\hat{\tau}$ is large, and temperature scaling (a-est) improves MSE on average (at $n_{\text{train}} = 5$, MSE $0.054 \pm 0.010 \to 0.045 \pm 0.007$, mean $\pm$ 95% CI over 50 independent seeds); however, in this regime the 95% CI of the nonparametric correction (c) does not overlap with that of (a-est), and (c) consistently falls below it (at $n_{\text{train}} = 5$, (a-est) 0.045 ± 0.007 vs. (c) 0.031 ± 0.004). This is because, in the regime dominated by the shape deviation (I-c) from the assumption, the assumption-free correction (c) has the advantage. In contrast, for $n_{\text{train}} \geq 500$, the difference between (a-est) and (c) shrinks to below 4×10^{-4}, and temperature scaling achieves accuracy comparable to the nonparametric method. This shows that, in the regime where $\hat{\tau}$ is stable and the absolute value of the residual slope is small, the correction based on Proposition 14 works effectively in practice. We note that the residual diagnostic quantities (the slope of r vs. θ^*, excess kurtosis) remain bounded away from 0 throughout the entire scanned range of n_{train}, coinciding at $n_{\text{train}} = 1000$ with the values in Table 2 (slope -0.124 ± 0.039, excess kurtosis $+2.03 \pm 0.20$). That is, the shape deviation (I-c) caused by the learner is not resolved by increasing the number of training samples, and persists regardless of n.

7.6 Summary

The results of Section 7.2 show that, under conditions where Assumption 1 holds by construction, the temperature-scaling-type correction behaves as predicted by Proposition 14, and that the error remaining after correction can be quantified as (I-a). Sections 7.3 and 7.5 show that, although (I-c) emerges under a learner-mediated condition and persists regardless of the number of training samples, the estimated noise level $\hat{\tau}$ stabilizes as the number of training samples increases, and temperature scaling converges to an accuracy comparable to the nonparametric correction. Section 7.4 further shows that the conclusions are preserved under varying class overlap and problem scale, while a condition with strongly separated classes (var $= 0.1$) marks the lower bound of the scope of applicability. Taken together, the practical role of the theoretical results of this paper is as follows: they function as predicted within the range where the assumption holds, and should be used complementarily with nonparametric correction according to the degree of deviation. However, all of the numerical verification in this paper relies on the idealized condition that the true logit θ^* is known (Section 7.1); applying it to real data requires a separately developed method for estimating τ (Section 8).

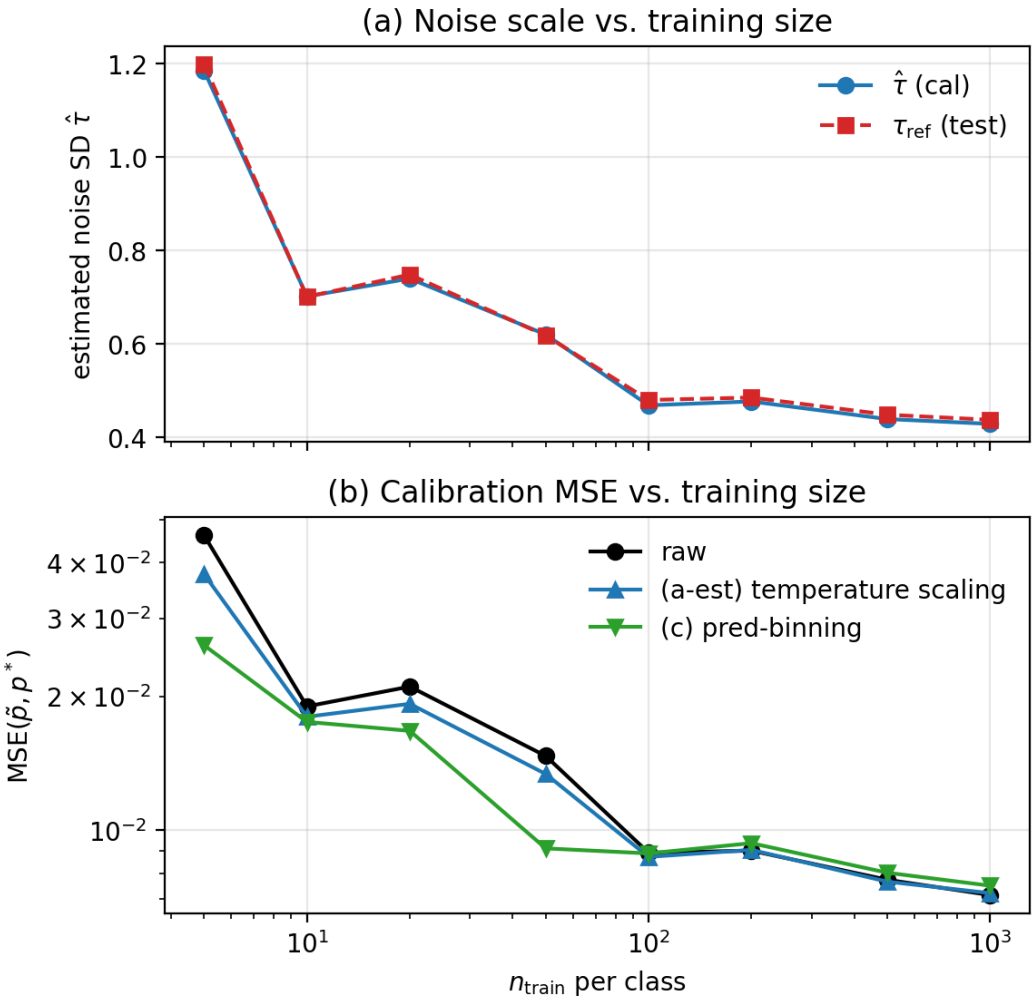

Figure 5: Dependence on the number of training samples (Condition L). (a) $\hat{\tau}$ decreases monotonically as n_{train} increases, converging to approximately 0.43 at $n_{\text{train}} = 1000$. (b) MSE decreases monotonically for raw, (a-est), and (c) alike, but for small n, (c) falls below (a-est) (see the text for details).

8 Discussion

8.1 Qualitative Properties of the Bayes-Optimal Estimator

The approximation (19) of Proposition 14 clarifies the qualitative behavior of the Bayes-optimal estimator. Since the inverse temperature parameter $\kappa(\tau) = T^{-1} = (1 + \pi\tau^2/8)^{-1/2}$ always satisfies $\kappa(\tau) < 1$ for $\tau > 0$, under the approximation (19) it holds that $\tilde{p}_{\text{Bayes}} < \hat{p}$ if $\hat{p} > 0.5$, $\tilde{p}_{\text{Bayes}} > \hat{p}$ if $\hat{p} < 0.5$, and $\tilde{p}_{\text{Bayes}} = 0.5$ (a fixed point) if $\hat{p} = 0.5$. This property, whereby the Bayes-optimal estimator shrinks the predicted probability toward 0.5 in the presence of noise, is qualitatively consistent with correcting the output of an overconfident model (quantitative verification on deep learning models lies beyond the empirical scope of this paper). The shrinkage operation governed by $\kappa(\tau)$ can be interpreted, within the calibration–sharpness bivariate evaluation framework of Kuleshov and Deshpande (2022), as an operation that improves calibration while appropriately preserving sharpness.

The limiting behavior is also consistent with the theory. In the limit $\tau \to 0$, where noise vanishes, $\kappa(\tau) \to 1$ and the correction approaches the identity map ($\tilde{p}_{\text{Bayes}} \to \hat{p}$). This matches the intuition that, without noise, calibration correction is unnecessary. Conversely, in the limit $\tau \to \infty$, where noise dominates, $\kappa(\tau) \to 0$

and all predictions converge to 0.5 ($\tilde{p}_{\text{Bayes}} \to 0.5$). This corresponds to a situation in which the observation carries no information whatsoever.

8.2 A Generative-Model Interpretation of Temperature Scaling

Proposition 14 gives temperature scaling, which is widely used in practice, an interpretation from the standpoint of a generative model. Via the correspondence $\tau^2 = (8/\pi)(T^2 - 1)$, one can back out the implicitly assumed noise level τ in logit space from a temperature parameter T estimated on a validation set. For example, $\tau = 1.0$ corresponds to $T \approx 1.18$ (Table 1), i.e., a temperature parameter of $T \approx 1.18$ implicitly assumes Gaussian noise with standard deviation 1.0 in logit space.

We stress, however, that this correspondence holds only within (i) the flat-prior assumption and (ii) the MacKay approximation, and does not assert that the two estimation procedures are identical. The results of Section 7.2 show that, as τ increases, the MacKay approximation error (I-a) increases monotonically from 3.2×10^{-3} at $\tau = 0.3$ to 1.09×10^{-2} at $\tau = 1.5$; in the strong-noise regime, this correspondence should therefore be regarded as a functional-form reference rather than an exact identity.

As for a theoretical characterization of temperature scaling, Mattei and Loureiro (2026) give a geometric and information-theoretic interpretation as an "information projection onto a prescribed-entropy set" and the "unique accuracy-preserving linear scaler." The generative-model interpretation of this paper is complementary to theirs, providing a characterization along a different axis: "Bayes-optimal correction in the presence of logit noise."

8.3 Unification under Squared-Type Losses

The combination of Theorem 11 and Proposition 19 shows that the Bayes-optimal estimators under the squared-error loss (a loss against the true probability p^*) and the Brier score (a loss against the observed label Y) reduce to the same posterior mean $\mathbb{E}[p^* \mid \hat{\theta}]$. This coincidence is a structural consequence of the fact that, in the decomposition of the Brier score (Lemma 17), the term that does not depend on p^* does not contribute to the optimization over $\tilde{p}$. Consequently, the analysis in this paper based on the logit-noise model applies consistently regardless of whether the squared-error loss or the Brier score is used as the evaluation metric.

8.4 Choosing between Temperature Scaling and Nonparametric Correction

The dependence on the number of training samples in Section 7.5 shows that the relative merits of temperature scaling and nonparametric correction (binning of the predicted value) transition according to n_{train}. For small n_{train}, the shape deviation (I-c) from the assumption dominates, and the assumption-free nonparametric correction outperforms temperature scaling. This is also consistent with the theoretical result of Chidambaram and Ge (2023) that the performance of TS degrades asymptotically under distributions with large class-support overlap. On the other hand, for large n_{train}, the influence of (I-c) shrinks and the accuracy of the two methods nearly coincides. This observation suggests positioning the two methods as complementary rather than competing. The theoretical results of this paper provide a method-selection framework in which (i) the empirical degree to which the assumption holds is quantified via residual diagnostics, and (ii) temperature scaling is chosen where this degree is high and nonparametric correction where it is low.

We emphasize that this framework is a diagnostic aid, not a guarantee that temperature scaling helps whenever it is applied. In the without-θ-flattening condition of Table 3 (Section 7.3), temperature scaling using the estimated $\hat{\tau}$ *increased* ECE on average, worsening it in 62% of trials, even though the mean and variance of the residual satisfied Assumption 1. This is a concrete instance of the more general point that satisfying the assumption at the level of mean and variance alone does not guarantee that the correction helps; the shape deviation (I-c) can still dominate and make the correction actively harmful, and residual diagnostics for higher-order deviations (skewness, kurtosis, and the θ^*-dependent slope), not just $\hat{\tau}$, should be checked before applying temperature scaling.

We note that the two methods also contrast in computational cost. Temperature scaling can be calibrated with only a one-dimensional search over the single parameter T, and at inference time it is a lightweight method that reduces to the scalar operation $\sigma(\mathrm{logit}(\hat{p})/T)$. The Bayes-optimal estimator (14) itself, in contrast, requires numerical quadrature of the Gaussian-sigmoid integral $\mathbb{E}[\sigma(Z)]$, and the method-selection procedure based on residual diagnostics proposed in this paper also entails additional computation of diagnostic statistics (excess kurtosis, the θ^*-dependent slope, and so on). However, all of these computations amount to a one-time preprocessing step performed at the calibration stage, and the inference-time computational cost can be kept on par with temperature scaling. Backing out τ from T via the temperature-scaling isomorphism (Proposition 14) also offers the practical advantage of operating with lightweight temperature scaling throughout, without going through numerical quadrature.

8.5 Limitations and Future Directions

The theoretical framework of this paper has several limitations, the most consequential of which concerns applicability to real data. All of the numerical verification in this paper relies on the idealized condition that the true logit θ^* is analytically known (Section 7.1), so that the noise level can be estimated as the residual standard deviation $\hat{\tau} = \mathrm{std}(\hat{\theta} - \theta^*)$. On real data, θ^* is never observed, and this estimator cannot be computed as is; consequently, this paper does not provide a validated procedure for deciding, from real data alone, when temperature scaling is trustworthy as an approximate Bayes-optimal correction. Developing such a procedure is an important task for bridging theory and practice, and it is a prerequisite for any practical use of the correspondence established here. One possible direction is to construct the empirical accuracy rate as a pseudo-value for p^* via confidence binning and to estimate $\hat{\tau}$ from its residual against the predicted logit, but the reliability of this substitution has not been established and is left for future work. An extension to a sample-dependent $\tau(x)$ relates to the lineage of Adaptive Temperature Scaling proposed by Balanya et al. (2024), and is a promising direction toward input-dependent confidence estimation.

The framework also has two further limitations. First, the generative assumption of additive Gaussian noise in logit space does not capture all of the miscalibration exhibited by real deep learning models. As observed in Section 7.3, under a learner-mediated condition the residuals exhibit a heavy-tailed distribution and a systematic bias with respect to θ^*. Non-Gaussian, non-additive extensions that explicitly treat such (I-c) are left for future work. Second, this paper is restricted to binary classification; generalizing to multi-class classification requires additional considerations, such as a multivariate Gaussian model in logit space.

9 Conclusion

In this study, we derived the Bayes-optimal calibration function under the logit-noise model and obtained three main theoretical results.

First, we showed that the Bayes-optimal calibration function for the squared-error loss is uniquely characterized as the posterior mean $\mathbb{E}[\sigma(\theta^*) \mid \hat{\theta}]$ (Theorem 11). This result is expressed as a sigmoid integral over a Gaussian posterior distribution, obtained by exploiting the conjugacy of the Gaussian prior. Second, we showed that the MacKay approximation to the Bayes-optimal estimator is expressed in the same functional form as temperature scaling, and we established a quantitative correspondence between the temperature parameter T and the noise variance τ^2 (Proposition 14). This gives temperature scaling a theoretical interpretation grounded in a generative model. This isomorphism, however, holds in the limit of a flat (non-informative) prior in logit space, and this dependence is an important caveat when interpreting the claims of this paper. We further showed that the Bayes-optimal estimator reduces to the same posterior mean under the Brier score as well (Proposition 19), confirming theoretical unification under squared-type losses.

For numerical verification, we used two conditions: synthetic data in which the assumption holds by definition, and data in which the assumption holds empirically through logistic regression. Under the idealized condition, we confirmed that the post-correction residual is quantified as the MacKay approximation error (I-a) and increases monotonically from $\tau = 0.3$ to $\tau = 1.5$. Under a learner-mediated condition, we showed that the assumption holds at the level of mean and variance, while normality and homoscedasticity hold only partially, which constrains the effective range of temperature-scaling-type correction. We also con-

firmed that, although the shape deviation from the assumption persists regardless of the number of training samples, temperature scaling converges to an accuracy comparable to nonparametric correction as the number of training samples increases. We further confirmed that these conclusions are preserved across multiple settings with varying class overlap and problem scale, and we delineated a condition with strongly separated classes as the lower bound of the scope of applicability for temperature scaling.

Future work includes extension to non-Gaussian noise models, generalization to multi-class classification, and the development of a method for estimating the noise parameter on real data.

References

Sergio A. Balanya, Juan Maroñas, and Daniel Ramos. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Computing and Applications*, 36:8073–8095, 2024. doi: 10.1007/s00521-024-09505-4.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009. doi: 10.1002/qj.456.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, The Wharton School, University of Pennsylvania (Buja, Shen) and Department of Statistics, University of Washington (Stuetzle), November 2005. URL `https://sites.stat.washington.edu/wxs/Learning-papers/paper-proper-scoring.pdf`.

Muthu Chidambaram and Rong Ge. On the limitations of temperature scaling for distributions with overlaps, 2023. URL `https://arxiv.org/abs/2306.00740`.

Muthu Chidambaram, Holden Lee, Colin Mcswiggen, and Semon Rezchikov. How flawed is ECE? An analysis via logit smoothing. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8417–8435. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/chidambaram24a.html`.

Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Expectation consistency for calibration of neural networks, 2023. URL `https://arxiv.org/abs/2303.02644`.

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. doi: 10.2307/2987588.

Martin Ferianc, Ondrej Bohdal, Timothy Hospedales, and Miguel Rodrigues. Navigating noise: A study of how noise influences generalisation and calibration of neural networks, 2023. URL `https://arxiv.org/abs/2306.17630`. Originally titled "Impact of Noise on Calibration and Generalisation of Neural Networks"; v2 (2024) renamed to current title.

Futoshi Futami and Masahiro Fujisawa. Information-theoretic generalization analysis for expected calibration error. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL `https://proceedings.neurips.cc/paper/2024/file/9961e42624a6c083279303767c73269d-Paper-Conference.pdf`.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

Sooyong Jang, Radoslav Ivanov, Insup Lee, and James Weimer. Confidence calibration with bounded error using transformations, 2021. URL `https://arxiv.org/abs/2102.12680`.

Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11683–11693. PMLR, 2022. URL https://proceedings.mlr.press/v162/kuleshov22a.html.

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2015)*, volume 9284 of *Lecture Notes in Computer Science*, pp. 68–85. Springer, 2015. doi: 10.1007/978-3-319-23528-8_5.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 3792–3803, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html.

Gendo Kumoi, Hideki Yagi, Manabu Kobayashi, Masayuki Goto, and Shigeichi Hirasawa. Performance Evaluation of Error-Correcting Output Coding Based on Noisy and Noiseless Binary Classifiers. *International Journal of Neural Systems*, 33(2):2350004, 2023. doi: 10.1142/S0129065723500041.

David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992a. doi: 10.1162/neco.1992.4.3.448.

David J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992b. doi: 10.1162/neco.1992.4.5.720.

Pierre-Alexandre Mattei and Bruno Loureiro. The well-tempered classifier: Some elementary properties of temperature scaling, 2026. URL https://arxiv.org/abs/2602.14862.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://arxiv.org/abs/2002.09437.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. URL https://papers.nips.cc/paper/8717-when-does-label-smoothing-help.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Binary classifier calibration: Bayesian non-parametric approach, 2014. URL https://arxiv.org/abs/1401.2955. Selection over Bayesian Binnings (SBB) and Averaging over Bayesian Binnings (ABB).

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 2901–2907, 2015. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR 2017 Workshop Track*, 2017. URL https://arxiv.org/abs/1701.06548.

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.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4036–4054. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/roelofs22a.html.

Yung-Chen Tang, Pin-Yu Chen, and Tsung-Yi Ho. Neural clamping: Joint input perturbation and temperature scaling for neural network calibration. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://openreview.net/forum?id=qSFToMqLcq. arXiv:2209.11604.

Linwei Tao, Haolan Guo, Minjing Dong, and Chang Xu. Consistency calibration: Improving uncertainty calibration via consistency among perturbed neighbors, 2024. URL https://arxiv.org/abs/2410.12295.

Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. URL https://arxiv.org/abs/1905.11001.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. Evaluating model calibration in classification, 2019. URL https://arxiv.org/abs/1902.06977.

Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 178–190. PMLR, 2020. URL https://proceedings.mlr.press/v108/wenger20a.html.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pp. 609–616. Morgan Kaufmann, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 694–699. ACM, 2002. doi: 10.1145/775047.775151.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11117–11128. PMLR, 2020. URL https://proceedings.mlr.press/v119/zhang20k.html.

Lijun Zhang, Roger Grosse, Richard Zemel, Raquel Urtasun, and Geoffrey Hinton. When and how mixup improves calibration. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26352–26374. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhang22f.html.