The Minimax Rate of Perturbed Second-Order Calibration

Kamil Ciosek* **Banafsheh Rafiee** **Sina Ghiassian** **Nicolò Felicioni**
Spotify

Abstract

Second-order calibration error quantifies how closely a higher-order predictor's epistemic-uncertainty estimate matches the conditional variance of the label probability on its level sets. We characterize the minimax rate of estimating the second-order calibration error for binary classification in the regime where a small perturbation is applied to the classifier outputs. Our procedure is simple: add independent bandwidth-h sech noise to the score coordinates, then regress $Y^{(1)}$ and $Y^{(1)}Y^{(2)}$ on the perturbed score using low-degree polynomials. Crucially, the sech perturbation makes the calibration functions *analytic* in a suitable strip. The resulting estimator has error $O_h(\log^{3/2} n/\sqrt{n})$, with explicit constants. In the same setting, a matching $\Omega(1/\sqrt{n})$ lower bound establishes minimax optimality up to logarithmic factors. As a corollary, we give a finite-sample guarantee for perturbed second-order Platt scaling, yielding a post-hoc procedure that recalibrates both the mean prediction and the epistemic-variance estimate of the perturbed higher-order predictor. Along the way, we give an explicit two-moment formulation of second-order calibration error and relate it quantitatively to the formulation of Ahdritz et al. [2025]. Our experiments confirm the predicted rate and the quality of the recalibrated uncertainties.

1 Introduction

Modern predictors often output not just a probability but also an uncertainty estimate. A deep ensemble, a Bayesian neural network, or an evidential model produces, for an input x, a mean prediction $m(x)$ together with a variance-like disagreement statistic $\sigma^2(x)$ intended to capture epistemic uncertainty (for example the empirical variance in ensemble outputs). Whether σ^2 actually has this meaning is a separate question from how it was produced. Ahdritz et al. [2025] gave a precise characterization: σ^2 captures epistemic uncertainty exactly when the predictor is *higher-order calibrated*. A calibrated σ^2 separates inherent ambiguity (aleatoric uncertainty) from resolvable ignorance (epistemic uncertainty). We propose to measure such second-order calibration (and thus the quality of epistemic uncertainty) using the second-order calibration error, in analogy to the well-known first-order calibration error [Błasiok et al., 2023, Lee et al., 2023].

This raises two basic statistical questions. How many 2-snapshots are needed to *estimate* the second-order calibration error of a given predictor, and at what rate? And how can one *correct* a miscalibrated (m, σ^2) so that its epistemic uncertainty acquires real-world semantics, with finite-sample guarantees? A fundamental problem is that some regularization is necessary: without additional structure, the worst-case estimation error need not vanish as more data are observed (see Appendix D). Ahdritz et al. [2025] address this by replacing exact level sets with a fixed partition of the score space.

We replace the partition by perturbing the classifier output using a sech kernel, an intervention that is controlled by just one hyperparameter (bandwidth) and that is arguably less invasive. Calibration

*Correspondence to: kciosek@gmail.com

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

functions are then approximated by regressing the labels on polynomial features of perturbed scores. Crucially, this specific perturbation makes the calibration functions of the perturbed classifier *analytic* in a suitable complex strip. Leveraging standard tools, we then show that this polynomial regression estimates the calibration error at rate $\tilde{O}(1/\sqrt{n})$ with explicit constants; moreover, using the (standard) Le Cam technique, we provide a matching $\Omega(1/\sqrt{n})$ lower bound, proving minimax-optimality up to logarithmic factors. The same construction then yields a post-hoc recalibration procedure with a finite-sample second-order calibration guarantee for the perturbed predictor, which we call second-order Platt scaling.

Contributions. We make the following contributions.

1. **Moment formulation**: we give an explicit binary two-moment calibration error, defined by conditioning on the predictor's score rather than choosing finite buckets.
2. **Analyticity** (lemma 1): for fixed bandwidth h, the sech-perturbed calibration functions are analytic in a complex strip.
3. **Fixed-h estimation at $\tilde{O}(1/\sqrt{n})$** (proposition 1): polynomial regression estimates $\mathrm{CE}_2^{\mathrm{pert}}$ at rate $O_h(\log^{3/2} n/\sqrt{n})$ from 2-snapshots.
4. **Rate tightness for the same target**: proposition 2 gives an $\Omega(1/\sqrt{n})$ lower bound for fixed-h sech-perturbed calibration, matching the upper bound up to logarithmic factors.
5. **Second-order Platt scaling** (proposition 3): remapping through the estimated calibration functions approximately recalibrates the perturbed predictor at the same $\tilde{O}_h(1/\sqrt{n})$ rate.

2 Setup

We study binary classification with $Y \in \{0,1\}$. Nature generates i.i.d. data; $f^*(x) = P(Y{=}1 \mid X{=}x)$. We make no assumptions on f^* or the marginal of X i.e., we are in the distribution-free setting. We have n i.i.d. 2-snapshots $(X_i, Y_i^{(1)}, Y_i^{(2)})$ with $Y_i^{(1)}, Y_i^{(2)} \mid X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(f^*(X_i))$.

Definition 1 (Higher-order predictor and calibration)**.** *A higher-order predictor maps $x \mapsto S(x) = (m(x), \sigma^2(x))$ with $m \in [0,1]$, $\sigma^2 \in [0, m(1{-}m)]$. Its calibration functions are*

$$\eta_1(s) = \mathbb{E}[Y \mid S(X) = s], \qquad \eta_2(s) = \mathbb{E}[f^*(X)^2 \mid S(X) = s], \tag{1}$$

and the second-order calibration error is

$$\mathrm{CE}_2 = \mathbb{E}[|\eta_1(S) - m|] + \mathbb{E}[|\eta_2(S) - v|] \tag{2}$$

with $v = m^2 + \sigma^2$.

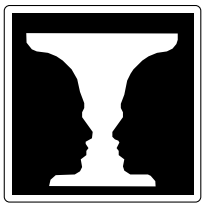
Intuitions. We will now provide intuitions about the usefulness of definition 1. The law of total variance can always be used to write

$$\underbrace{\eta_1(s)(1-\eta_1(s))}_{\text{total}} = \underbrace{\mathbb{E}[f^*(X)(1{-}f^*(X)) \mid S{=}s]}_{\text{aleatoric}} + \underbrace{\mathrm{Var}(f^*(X) \mid S{=}s)}_{\text{epistemic}}. \tag{3}$$

When $\mathrm{CE}_2 = 0$ i.e., for a second-order calibrated classifier, we can further write

$$\mathrm{Var}(f^*(X) \mid S{=}s) = \underbrace{\mathbb{E}[(f^*(X))^2 \mid S{=}s]}_{\eta_2(s)=v} - \underbrace{\mathbb{E}[f^*(X) \mid S{=}s]^2}_{\eta_1(s)=m} = v - m^2 = \sigma^2$$

This means that calibration allows us to interpret σ^2 as a measure of epistemic uncertainty. A high-level comparison between first and second-order calibration is given in Table 1. Note that while the definition of second-order calibration given here is ostensibly different from the definition due to Ahdritz et al. [2025], the two versions are in fact very closely related (see Appendix F for a discussion). See Figure 1 for further illustrations of calibration functions and how they relate to epistemic and aleatoric uncertainty.[2]



ambiguous
Rubin figure



vase
only



two faces
only

Set A: $S = (\frac{1}{2}, 0)$

True label-probability: $f^*(x) = 1/2$.

$$\eta_1 = 1/2, \qquad \eta_2 = 1/4$$

hence

$$\sigma^2 = \eta_2 - \eta_1^2 = 1/4 - 1/4 = 0.$$

(no epistemic uncertainty)

Set B: $S = (\frac{1}{2}, \frac{1}{4})$

Half the time $f^*(x) = 1$, half the time $f^*(x) = 0$.

$$\eta_1 = 1/2, \qquad \eta_2 = \tfrac{1}{2}(1^2 + 0^2) = 1/2$$

hence

$$\sigma^2 = \eta_2 - \eta_1^2 = 1/2 - 1/4 = 1/4.$$

(high epistemic uncertainty)

Figure 1: Why second-order calibration is useful. Both sets have the same mean prediction $m = 1/2$, so a first-order calibrated predictor treats them identically. But Set A's Rubin figures are genuinely 50:50 (aleatoric, $\sigma^2 = 0$), while Set B mixes clear vases with clear faces that the classifier hasn't seen enough data to tell apart (epistemic, $\sigma^2 = 1/4$). A calibrated σ^2 distinguishes inherent ambiguity from resolvable ignorance.

	First-order	**Second-order (this paper)**
Output	probability s	(m, σ^2)
Data	labels Y_i	2-snapshots $(Y_i^{(1)}, Y_i^{(2)})$
Measure	$\eta(s) = \mathbb{E}[Y \mid S{=}s]$	$\eta_1(s), \eta_2(s)$
Diagnose	$\mathrm{CE}_1 = \mathbb{E}[\lvert\eta - s\rvert]$	CE_2
Fix	$s \mapsto \eta(s)$	$(m, \sigma^2) \mapsto (\eta_1, \eta_2 - \eta_1^2)$
Guarantee	calibrated probabilities	calibrated epistemic uncertainty
Rate	$\tilde{O}(1/\sqrt{n})$	$\tilde{O}(1/\sqrt{n})$

Table 1: Comparison of first and second-order calibration (rates hold for the sech-perturbed predictor).

Goals. Definition 1 applies to any higher-order predictor. Our statistical goal is to estimate this error for the fixed-h sech-perturbed predictor (defined in section 3), and to recalibrate that same predictor using the calibration functions in (1).

3 The Sech Perturbation, Analyticity and Estimation

Key insight: analyticity. We crucially leverage the bandwidth-h sech perturbation kernel of Ciosek et al. [2026], previously only used to enforce smoothness (we define the kernel formally in Definition 2 in the following section). Crucially, this specific perturbation makes the resulting calibration functions analytic in a complex strip. The classical Bernstein ellipse theorem then gives exponentially accurate polynomial approximation, yielding the $O_h(\log^{3/2} n/\sqrt{n})$ bound with polynomial degree $l = O_h(\log n)$. We note that, while Definition 1 is universal, our bounds assume that the classifier outputs have the sech perturbation applied (with a fixed bandwidth, see the following section).

[2]The observation that ambiguity is connected to epistemic uncertainty has been made by Osband et al. [2023] in the context of epistemic neural networks, a specific way of obtaining uncertainty estimates.

3.1 The perturbation scheme

Similarly to Ciosek et al. [2026], we perturb the predictor's scores by adding noise from a sech kernel. This is a design choice: the perturbed predictor is the one we deploy and recalibrate. Ciosek et al. [2026] have shown that, for realistic values of the bandwidth h, this perturbation has a very small impact on classifier performance (as measured by AUROC). We describe a practical process for choosing the bandwidth h in Appendix C.

Definition 2 (Sech perturbation)**.** *For a score $s \in [a, b]$ and bandwidth $h > 0$, the perturbation kernel is*

$$k_h(t \mid s) = \frac{\operatorname{sech}\left(\frac{t-s}{h}\right)}{Z(s,h)} \cdot \mathbf{1}_{[a,b]}(t), \qquad Z(s,h) = \int_a^b \operatorname{sech}\left(\frac{u-s}{h}\right) \mathrm{d}u. \tag{4}$$

The perturbed score $S \sim k_h(\cdot \mid s)$ satisfies $S \in [a, b]$.

Write

$$S_{\text{orig}}(x) = \big(m_{\text{orig}}(x), \sigma^2_{\text{orig}}(x)\big)$$

for the original predictor. After perturbation, we write $S(x) = (m(x), \sigma^2(x))$. Throughout the remainder of the paper, **unadorned notation refers to this perturbed, deployed score**. We apply a product perturbation on $\mathcal{R} = [0, 1] \times [0, \frac{1}{4}]$; that is, we draw the two coordinates independently:

Step 1. Draw $m \sim k_h(\cdot \mid m_{\text{orig}})$ on $[0, 1]$.

Step 2. Draw $\sigma^2 \sim k_h(\cdot \mid \sigma^2_{\text{orig}})$ on $[0, \frac{1}{4}]$, independently.

The rectangle $\mathcal{R}$ contains the feasible region

$$\mathcal{F} = \{(m_{\text{orig}}, \sigma^2_{\text{orig}}) : \sigma^2_{\text{orig}} \le m_{\text{orig}}(1 - m_{\text{orig}})\},$$

so every original score lies in $\mathcal{R}$. The perturbed score (m, σ^2) may fall outside $\mathcal{F}$, but always remains in $\mathcal{R}$. For perturbed scores outside $\mathcal{F}$, we use the same conditional moments and the same error formula (2), now on all of $\mathcal{R}$. All estimation and recalibration guarantees below concern this perturbed score.

3.2 The analyticity lemma

The following lemma, proved in Appendix E.2, is the kernel-specific step in our analysis. It shows that the denominator of the normalized sech smoothing is zero-free in a complex strip, and hence that the perturbed calibration functions are analytic. The subsequent polynomial approximation and regression arguments use standard tools. This zero-free argument does not automatically extend to Gaussian perturbations, because analyticity of the numerator and denominator alone does not exclude complex zeros of the denominator.

Lemma 1 (Analyticity of perturbed calibration functions)**.** *Let S_{orig} be any higher-order predictor with calibration functions $\eta_1^{\text{orig}}, \eta_2^{\text{orig}}$. Let η_1, η_2 be the calibration functions of its* sech*-perturbed predictor with bandwidth $h > 0$.*

- ***(1D)*** *If $S_{\text{orig}}(x) = m_{\text{orig}}(x) \in [0, 1]$ is a scalar score, then η_1 is analytic in the strip $\{z \in \mathbb{C} : |\operatorname{Im}(z)| < h\pi/2\} \cap \{z \in \mathbb{C} : 0 \le \operatorname{Re}(z) \le 1\}$.*
- ***(2D)*** *For the full score $S_{\text{orig}}(x) = (m_{\text{orig}}(x), \sigma^2_{\text{orig}}(x))$ with product perturbation, fix any $0 < \rho < h\pi/4$. Then, for each $j = 1, 2$, η_j extends jointly holomorphically to $\Omega_\rho := \{z \in \mathbb{C}^2 : |\operatorname{Im} z_1| < \rho,\ |\operatorname{Im} z_2| < \rho\}$. Moreover, on the closed product sub-strip $|\operatorname{Im} z_r| \le \rho$, we have $|\eta_j(z)| \le M_\rho$, where $M_\rho := \sec(2\rho/h)$.*

3.3 Polynomial Estimation

We state a classical polynomial-approximation result, proved in Appendix E.3 in our notation.

Lemma 2 (Bernstein ellipse theorem on $[0, 1]$)**.** *Let $f : [0, 1] \to \mathbb{R}$ be the restriction of a function analytic in the strip $|\operatorname{Im}(z)| < \rho$ around $[0, 1]$, with $\sup_{|\operatorname{Im}(z)|<\rho} |f(z)| \le M$. Then for every integer $l \ge 0$, there exists a polynomial q_l of degree l with*

$$\sup_{t \in [0,1]} |f(t) - q_l(t)| \le 2M \, \frac{\theta^{-l}}{\theta - 1},$$

Algorithm 1 Estimating fixed-h perturbed second-order calibration error

Require: Original predictor S_{orig}, 2-snapshot data $\{(X_i, Y_i^{(1)}, Y_i^{(2)})\}_{i=1}^n$, bandwidth h, and polynomial degree l

for $i = 1, \dots, n$ **do**
  $(m_{\mathrm{orig},i}, \sigma^2_{\mathrm{orig},i}) \leftarrow S_{\mathrm{orig}}(X_i)$
  Draw $m_i \sim k_h(\cdot \mid m_{\mathrm{orig},i})$ on $[0,1]$
  Draw $\sigma_i^2 \sim k_h(\cdot \mid \sigma^2_{\mathrm{orig},i})$ on $[0, \frac{1}{4}]$, independently
  $S_i \leftarrow (m_i, \sigma_i^2)$, $P_i \leftarrow Y_i^{(1)} Y_i^{(2)}$, and $v_i \leftarrow m_i^2 + \sigma_i^2$
end for
Fit $\hat{\eta}_1$ by least squares of $Y_i^{(1)}$ on S_i over the class $\mathcal{P}_l$
Fit $\hat{\eta}_2$ by least squares of P_i on S_i over the class $\mathcal{P}_l$
$\widehat{\mathrm{CE}}_2 \leftarrow \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \left(|\hat{\eta}_1(S_i) - m_i| + |\hat{\eta}_2(S_i) - v_i| \right)$
return $\widehat{\mathrm{CE}}_2, \hat{\eta}_1, \hat{\eta}_2$

where $\theta = 2\rho + \sqrt{4\rho^2 + 1} > 1$ is the Bernstein ellipse parameter.

This yields the following corollary (proven in Appendix E.4)

Corollary 1 (Polynomial approximation of η)**.** *Fix any $0 < \rho < h\pi/4$, and define*

$$\theta_\rho := 2\rho + \sqrt{4\rho^2 + 1}, \qquad M_\rho := \sec(2\rho/h).$$

For every integer $\ell \geq 0$ and each $j = 1, 2$, there exists a polynomial $q_{j,\ell}(m, \sigma^2)$ of degree at most ℓ in each coordinate such that, on $\mathcal{R} = [0,1] \times [0, 1/4]$,

$$0 \leq q_{j,\ell}(m, \sigma^2) \leq 1, \qquad \sup_{(m,\sigma^2) \in \mathcal{R}} |\eta_j(m, \sigma^2) - q_{j,\ell}(m, \sigma^2)| \leq B_\rho \theta_\rho^{-\ell},$$

where $B_\rho := \frac{16 M_\rho \theta_\rho}{(\theta_\rho - 1)^2}$.

Corollary 1 immediately gives us a way to leverage analyticity to measure the quality of the polynomial approximation in our specific use-case. Indeed, for fixed bandwidth h, we choose $\rho = h\pi/8$, so that $\theta_h := \frac{h\pi}{4} + \sqrt{1 + \frac{h^2\pi^2}{16}} > 1$, $M_h = \sqrt{2}$, and $B_h := \frac{16\sqrt{2}\,\theta_h}{(\theta_h - 1)^2}$ are constants. Choosing $l = \lceil 2 \log n / \log \theta_h \rceil$ gives $\varepsilon_l \leq B_h \theta_h^{-l} \leq B_h n^{-2}$, negligible compared to the estimation error.

4 Our Algorithm and Estimator

Algorithm 1 summarizes our core procedure. The constrained polynomial regression class is defined below.

For any bounded function u on $\mathcal{R}$, write $\|u\|_\infty := \sup_{(m,\sigma^2)\in\mathcal{R}} |u(m, \sigma^2)|$. Let $\mathcal{P}_l$ denote the class

$$\mathcal{P}_l := \left\{ \mathrm{clip}_{[0,1]}(p) : p(m, \sigma^2) = \beta^\top \phi(m, \sigma^2),\ \|p\|_\infty \leq 1 \right\},$$

where ϕ is the tensor product polynomial basis of degree l in each variable, so ϕ has $d = (l+1)^2$ components. The unit ball of this finite-dimensional polynomial space is compact, and clipping is continuous in the sup norm. Therefore $\mathcal{P}_l$ is compact and the squared-loss ERM minimizers below exist. Every function in $\mathcal{P}_l$ takes values in $[0,1]$.

Given 2-snapshots $(X_i, Y_i^{(1)}, Y_i^{(2)})_{i=1}^n$ with perturbed scores $S_i = (m_i, \sigma_i^2)$ and products $P_i = Y_i^{(1)} Y_i^{(2)}$:

$$\hat{\eta}_1 = \arg\min_{f \in \mathcal{P}_l} \ \frac{1}{n} \sum_{i=1}^{n} \big(Y_i^{(1)} - f(S_i) \big)^2, \tag{5}$$

$$\hat{\eta}_2 = \arg\min_{f \in \mathcal{P}_l} \ \frac{1}{n} \sum_{i=1}^{n} \big(P_i - f(S_i) \big)^2. \tag{6}$$

Both regressions use the same constrained ERM class $\mathcal{P}_l$, with responses $Y_i^{(1)}$ and P_i, respectively. Proposition 1 analyzes this constrained ERM (the experiments instead use ridge regression followed by output clipping as a practical surrogate).

The CE estimator is:

$$\widehat{\mathrm{CE}}_2 = \frac{1}{n}\sum_{i=1}^{n}|\hat{\eta}_1(S_i) - m_i| + \frac{1}{n}\sum_{i=1}^{n}|\hat{\eta}_2(S_i) - (m_i^2 + \sigma_i^2)|. \tag{7}$$

5 Main Theoretical Result

We now characterize the estimation rate for the calibration error of the fixed-h perturbed predictor.

Proposition 1 (Estimation of $\mathrm{CE}_2^{\mathrm{pert}}$ at near-parametric rate)**.** *Fix the perturbation bandwidth $h > 0$ and let $\theta_h := \frac{h\pi}{4} + \sqrt{1 + \frac{h^2\pi^2}{16}}$. Set the polynomial degree to $l = \lceil 2\log n / \log \theta_h \rceil$, so that $d = (l+1)^2 = O_h(\log^2 n)$. Then the estimator* (7) *satisfies*

$$\mathbb{E}\big[|\widehat{\mathrm{CE}}_2 - \mathrm{CE}_2^{\mathrm{pert}}|\big] = O_h\left(\frac{\log^{3/2} n}{\sqrt{n}}\right), \tag{8}$$

where $\mathrm{CE}_2^{\mathrm{pert}}$ is the second-order calibration error of the perturbed predictor. The constant depends on h, but not on the score distribution or on the original calibration functions.

Origin of the logarithmic gap. For fixed h, analyticity gives approximation error that decreases exponentially with the polynomial degree l. Thus $l = O_h(\log n)$ makes the approximation error negligible. In two dimensions, the tensor polynomial class then has

$$d = (l+1)^2 = O_h(\log^2 n)$$

coefficients. The covering-number bounds used for regression and for evaluating the empirical calibration error contribute a further $\log n$ factor. The resulting estimation terms are therefore

$$O\left(\sqrt{\frac{d\log n}{n}}\right) = O_h\left(\frac{\log^{3/2} n}{\sqrt{n}}\right).$$

The lower bound is only $\Omega(1/\sqrt{n})$. We do not know whether the $\log^{3/2} n$ factor is information-theoretically necessary; it may be an artifact of the proof technique.

Moreover, we can use essentially the same technique to obtain a novel result about first-order calibration; this is reflected in the remark below.

Remark 1 (First-Order Calibration)**.** *While Proposition 1 does not immediately concern first-order calibration,*[3] *the same technique estimates $\mathrm{CE}_1 = \mathbb{E}[|\mathbb{E}[Y \mid m] - m|]$ at rate $O_h(\log^{3/2} n/\sqrt{n})$.*

Comparison with bucketing. The bucketing rate depends on the specific setting we are considering. For a predictor that is constant on a fixed partition with B cells, the natural estimator of the aggregate, probability-weighted second-order calibration error has expected error $O(\sqrt{B/n})$. Thus, when B is fixed, the partition-based target has the usual $n^{-1/2}$ rate, with a constant depending on B. A different bucketing rate of $n^{-1/4}$ arises in a different setting: using an increasingly fine regular partition to approximate an unpartitioned Lipschitz target in two dimensions. Our theorem concerns a third target, $\mathrm{CE}_2^{\mathrm{pert}}$ for fixed h. The sech perturbation makes this target analytic, giving $O_h(\log^{3/2} n/\sqrt{n})$. We therefore claim introducing the new perturbed setting rather than faster estimation in the fixed-partition setting.

6 Lower Bound and Minimax Optimality for a Fixed Perturbation

To complement the upper bound, we give a lower bound for the same fixed-h sech-perturbed setting. Propositions 1 and 2 therefore characterize its minimax rate up to logarithmic factors. The following proposition is proved in Appendix G.

[3] In first-order calibration, conditioning is only with respect to the mean, ignoring the epistemic variance.

Proposition 2 (Minimax lower bound for fixed-h perturbed CE_2)**.** *There exists an absolute constant $c > 0$ such that, for every sech bandwidth $h > 0$, every $n \geq 1$, and every estimator $\widehat{\mathrm{CE}}_2$, two distributions P_0, P_1 in the setup of the main paper exist with*

$$\max_{b\in\{0,1\}} \mathbb{E}_{P_b}\big[\,\big|\widehat{\mathrm{CE}}_2 - \mathrm{CE}_2^{\mathrm{pert}}(P_b)\big|\,\big] \;\geq\; \frac{c}{\sqrt{n}}.$$

The constant c does not depend on h.

7 Second-Order Platt Scaling

We now describe the second-order analogue of Platt scaling.

7.1 The recalibration map

Given the true calibration functions η_1, η_2, define the *recalibrated predictor*:

$$T(x) = \big(\underbrace{\eta_1(S(x))}_{m'(x)},\; \underbrace{\eta_2(S(x)) - \eta_1(S(x))^2}_{(\sigma^2)'(x)}\big), \tag{9}$$

where $S(x)$ is the perturbed score.

Proposition 3 (Exact second-order Platt scaling)**.** *If η_1, η_2 are the true calibration functions, then T is perfectly second-order calibrated with respect to its own level sets:*

$$\mathbb{E}[Y \mid T(X) = (m', (\sigma^2)')] = m', \tag{10}$$

$$\mathbb{E}[f^*(X)^2 \mid T(X) = (m', (\sigma^2)')] = (m')^2 + (\sigma^2)'. \tag{11}$$

Proof. **First moment.** By the tower property:

$$\mathbb{E}[Y \mid T(X) = (m', (\sigma^2)')] = \mathbb{E}[\;\underbrace{\mathbb{E}[Y \mid S(X)]}_{=\eta_1(S(X))=m'(X)}\;\mid T(X) = (m', (\sigma^2)')].$$

On the event $\{T(X) = (m', (\sigma^2)')\}$, the first component $m'(X) = m'$ is constant. So the expression equals m'.

Second moment. Similarly:

$$\mathbb{E}[f^*(X)^2 \mid T(X) = (m', (\sigma^2)')] = \mathbb{E}[\eta_2(S(X)) \mid T(X) = (m', (\sigma^2)')].$$

Now $\eta_2(S(X)) = \eta_1(S(X))^2 + (\eta_2(S(X)) - \eta_1(S(X))^2) = m'(X)^2 + (\sigma^2)'(X)$. On the conditioning event, this equals $(m')^2 + (\sigma^2)'$. □

7.2 Approximate guarantee from finite data

For the estimated calibration functions from Section 4, we first enforce consistency of the two predicted moments:

$$\begin{aligned} a(s) &= \hat{\eta}_1(s), \qquad b(s) = \hat{\eta}_2(s), \\ b'(s) &= \min\{a(s), \max\{a(s)^2, b(s)\}\}, \\ \hat{T}(x) &= \big(a(S(x)),\, b'(S(x)) - a(S(x))^2\big). \end{aligned} \tag{12}$$

Since $0 \leq a(s) \leq 1$, clipping the second moment to $[a(s)^2, a(s)]$ ensures that the reported variance lies in $[0, a(s)(1 - a(s))]$. Thus $\hat{T}$ always satisfies the feasibility requirement in Definition 1.

Corollary 2. *Under the assumptions of Proposition 1, the recalibrated predictor $\hat{T}$ satisfies*

$$\mathbb{E}[\mathrm{CE}_2(\hat{T})] = O_h\left(\frac{\log^{3/2} n}{\sqrt{n}}\right).$$

See Appendix E.8 for proof.

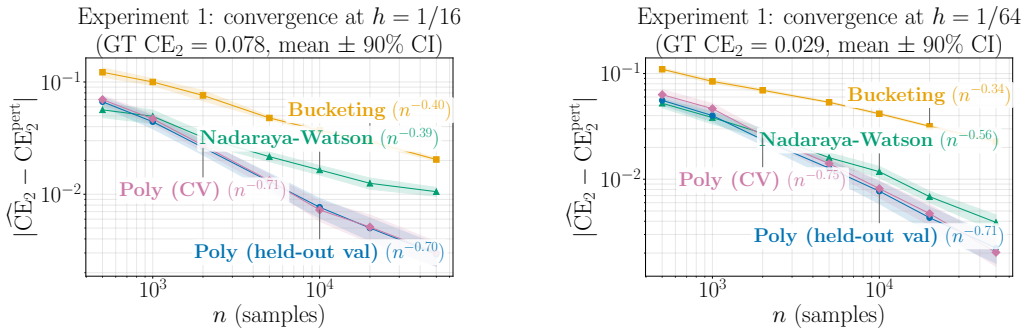


Figure 2: **Experiment 1 (rate).** $|\widehat{\mathrm{CE}}_2 - \mathrm{CE}_2^{\mathrm{pert}}|$ vs. n at h=1/16 (left) and h=1/64 (right); mean ± Student-t 90% CI across 20 seeds, log–log slopes in legend.

8 Experiments

Our four experiments illustrate the method's empirical behavior. Experiment 1 shows favorable finite-sample scaling of the practical estimator; Experiment 2 confirms that the resulting recalibrator $\hat{T}$ improves calibration in the spirit of Corollary 2; Experiment 3 isolates *why* that calibration is worth doing: its *magnitude*, not just its ranking, controls downstream decisions; and Experiment 4 reproduces the same effect on real labels. Experiments 1-3 use synthetic data with known ground truth, Experiment 4 uses the public Weather Sentiment AMT dataset.

Experiment 1: the predicted rate. The DGP is a 10-component Gaussian mixture in $\mathbb{R}^4$ with $f^*(x)=\sigma(w^\top x)$; the predictor is a 5-member MLP ensemble whose mean and unbiased variance are sech-perturbed coordinate-wise. Ground-truth $\mathrm{CE}_2^{\mathrm{pert}}$ is computed by quasi-Monte Carlo (replicate spread $< 2{\cdot}10^{-5}$). We compare 2D bucketing, Nadaraya-Watson with an isotropic Gaussian kernel, and our polynomial estimator (held-out and CV-tuned variants); baseline bin counts and bandwidths follow textbook MISE rules with constants tuned on a held-out split. Figure 2 shows that the polynomial estimator dominates at every n, with empirical log–log slope ≈ -0.70 at both $h \in \{1/16, 1/64\}$ - at least as steep as the predicted $-1/2$ rate and steeper than the -0.33 to -0.55 slopes of bucketing and NW (consistent with their distribution-free Lipschitz rate). The two polynomial variants are indistinguishable, so the advantage is not an artefact of held-out tuning. Setup details in App. B.1.

Experiment 2: the recalibrated σ^2 tracks the truth. We replace the predictor with a 20-member ensemble deliberately undertrained so that its raw σ^2 carries only a faint epistemic signal; the recalibrator is a tensor-Chebyshev ridge regression of degree 12 fit on $2{\cdot}10^4$ 2-snapshots. Ground-truth conditional variance at perturbed scores is obtained by sech-smoothing the empirical (p, p^2) joint of the DGP. The raw σ^2 is essentially uncorrelated with truth (Pearson 0.10), but after applying $\hat{T}$ the scatter concentrates on the diagonal (Pearson 0.74); see Figure 3. Across 20 calibration seeds, CE_2 drops from 0.37 to 0.04 ± 0.01, demonstrating improved calibration with the practical recalibrator. Setup details in App. B.2.

Experiment 3: magnitudes, not just rankings. A clinic refers borderline patients ($m \in [0.35, 0.65]$) at cost c=0.06 for a noisy specialist panel; the per-patient gain $g(\theta)=2\theta^2 - 2\theta + 0.5 - c$ has conditional mean $2\eta_2(S) - 2\eta_1(S) + 0.5 - c$, so any post-hoc procedure that calibrates only η_1 is misspecified for the Bayes-optimal score. The borderline cohort splits into an aleatoric subpopulation ($\theta{\approx}0.5$, small σ^2) and a hidden-subtype one ($\theta \in \{0.12, 0.88\}$, large σ^2), both centred at $m{\approx}0.5$. Figure 4 reports realised gain versus referral threshold τ. 2-D second-order Platt tracks the oracle across the full $\tau \geq 0$ range. The raw (m, σ^2) plug-in matches at τ=0 but *collapses* near $\tau{\approx}0.06$: it ranks patients correctly but its score *magnitudes* are biased low on the epistemic group, so any cost-aware threshold drops the entire group. First-order methods cannot separate the subpopulations and refer essentially no one. The lesson is exactly what Proposition 3 promises and ranking-only metrics would hide: a calibrated σ^2 matters numerically. Setup details in App. B.3.

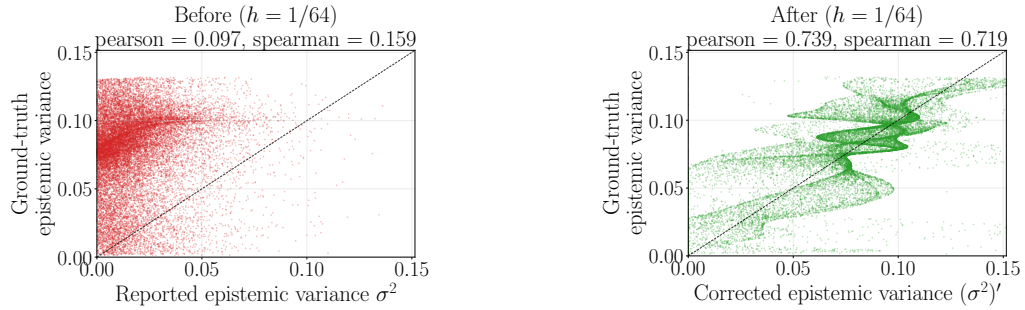


Figure 3: **Experiment 2 (recalibration).** Reported σ^2 (left) and corrected $(\sigma^2)'$ (right) versus the ground-truth conditional variance on a held-out split, at $h{=}1/64$. Dashed line is $y{=}x$.

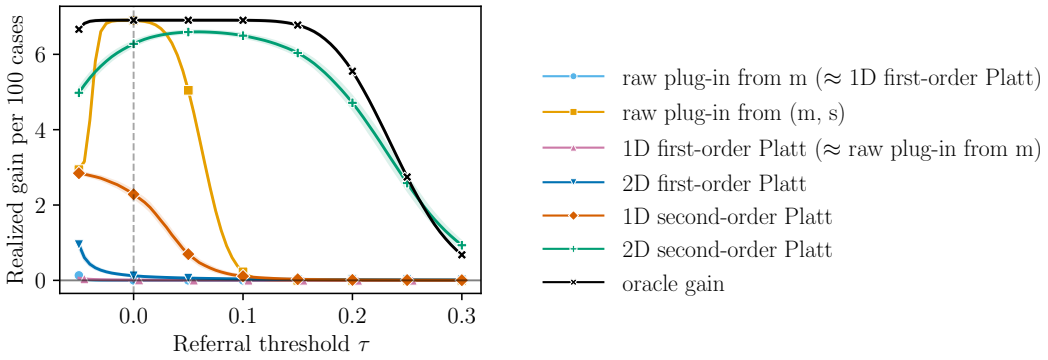


Figure 4: **Experiment 3 (decision utility).** Realised gain per 100 borderline patients vs. referral threshold τ (mean ± 1.96 SEM, 200 repeats). 2-D second-order Platt tracks the oracle; the raw (m, σ^2) plug-in collapses near $\tau{\approx}0.06$.

Experiment 4: the same effect on real labels. We move to the public Weather Sentiment AMT dataset (300 tasks, 20 worker votes each, one-vs-rest binarisation), with an audit task whose downstream value provably depends on η_2 and not just η_1: an auditor must rank items by the probability that two fresh workers disagree, $g(\theta){=}2\theta(1-\theta)$, whose calibrated form $2\eta_1(S) - 2\eta_2(S)$ requires η_2. Items are mixed in equal proportion from an aleatoric cohort ($\theta \approx 1/2$) and a hidden-subtype cohort (θ near 0 or 1); the score $S = (m, s) \in [0, 1]^2$ is constructed in an XOR pattern so that the 1-D marginals of m and of s are *identical* between cohorts, leaving any 1-D calibrator information-theoretically blind. At a 5% audit budget, 2-D second-order Platt finds 46.4 disagreements per 100 audited items, versus 30.9 for 2-D first-order Platt, $\sim$26 for all five 1-D and raw baselines, and 49.8 for the oracle; the paired AUC lift of 2-D second-order over each non-oracle baseline has a 95% CI strictly above zero (per-baseline win fractions ≥ 0.97). On real labels, in a task whose value provably depends on η_2, second-order Platt is the only practical method that approaches the oracle. Full audit-yield curves and per-method statistics in App. B.4 (Fig. 5).

9 Related Work

Higher-order calibration. Ahdritz et al. [2025] introduced higher-order calibration on predictor level sets and developed partition-based methods using k-snapshots. Their Theorem 2.6 provides a finite-sample guarantee for estimating the snapshot distribution within a partition cell. We study the average binary two-moment error of a fixed-h sech-perturbed predictor under global i.i.d. sampling. Appendix F relates our moment criterion to their Wasserstein criterion on the same fixed partition.

First-order calibration measurement. ECE is conventionally estimated by binning [Naeini et al., 2015, Guo et al., 2017], but binned estimators are biased and inconsistent [Vaicenavicius et al., 2019, Kumar et al., 2019]. Kumar et al. [2019] give a debiased estimator with finite-sample bounds,

Roelofs et al. [2022] empirically compare variants, and Widmann et al. [2019] obtain consistent multi-class estimators via matrix-valued kernels. Błasiok et al. [2023] unify this literature through the distance to calibration, identifying which measures are polynomially related to it. Closest in spirit to our work is the SmoothECE of Błasiok and Nakkiran [2023], a first-order calibration measure built from Nadaraya–Watson regression with a reflected-Gaussian kernel. We differ in two essential respects: we study the two-dimensional second-order error, and we use the sech kernel of Ciosek et al. [2026], which makes the fixed-h perturbed calibration functions analytic rather than merely smooth. For this perturbed target, Bernstein-ellipse approximation gives exponentially accurate polynomial approximation and an $O_h(\log^{3/2} n/\sqrt{n})$ estimation rate. By contrast, generic two-dimensional local regression under only Lipschitz regularity has the $n^{-1/4}$ bias-variance rate described in Section 5. This comparison concerns estimators of the same perturbed target, not the fixed-partition target of Ahdritz et al. [2025]. Ciosek et al. [2026] establish an $n^{-1/3}$ rate for the first-order error and do not exploit analyticity. Lee et al. [2023] study the complementary problem of *testing* first-order calibration and prove minimax optimality of their T-Cal procedure under a smoothness assumption.

Distribution-free impossibility and regularization. Gupta et al. [2020] show that distribution-free calibration guarantees require the score level sets to form at most a countable partition, ruling out generic continuous-output schemes (under the assumptions of their framework). For the related evaluation problem, Lee et al. [2023, Proposition 9] show that first-order calibration testing is impossible under continuity alone, without stronger regularity. The reduction in Appendix D transfers this obstruction to unrestricted unperturbed second-order calibration. Bucketing, structural assumptions on the calibration functions, and deployment-time perturbation are alternative ways to make the problem tractable; our results use the last of these.

Post-hoc calibration. Platt scaling [Platt et al., 1999], isotonic regression [Zadrozny and Elkan, 2002], beta calibration [Kull et al., 2017], and temperature scaling [Guo et al., 2017] are the standard first-order recalibration methods: each fits a low-dimensional map from scores to corrected probabilities. Our second-order Platt scaling is, to our knowledge, the first post-hoc procedure that recalibrates both a mean prediction and an epistemic-variance estimate with provable finite-sample bounds on the calibration error of the recalibrated classifier without requiring score bucketing.

Multicalibration. A complementary line of work strengthens first-order calibration along the subgroup axis rather than the moment axis. Hébert-Johnson et al. [2018] require the mean prediction to be calibrated simultaneously on every subpopulation in a rich, computationally-identifiable class. Our work is orthogonal: we calibrate the second moment on the predictor's own level sets rather than the first moment across many subgroups. Combining the two, i.e., developing multicalibrated higher-order predictors is a natural direction left to future work.

Additional Related Work We provide additional pointers to related work in appendix A.

10 Conclusion

For every fixed bandwidth, the sech perturbation makes the calibration functions of the perturbed predictor analytic. Exploiting this structure with polynomial regression yields $O_h(\log^{3/2} n/\sqrt{n})$ estimation of $\mathrm{CE}_2^{\mathrm{pert}}$, with a matching $\Omega(1/\sqrt{n})$ lower bound up to logarithmic factors. The result concerns the perturbed predictor that is evaluated, recalibrated, and deployed; it is distinct from both unperturbed calibration and calibration relative to a fixed partition. Moreover, remapping the perturbed predictor through the estimated calibration functions produces an approximately second-order calibrated predictor whose epistemic uncertainty has finite-sample guarantees. We briefly explore remaining open questions in Appendix I.

References

Gustaf Ahdritz, Aravind Gollakota, Parikshit Gopalan, Charlotte Peale, and Udi Wieder. Provable uncertainty decomposition via higher-order calibration. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TId1SHe8JG.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Jarosław Błasiok and Preetum Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. *arXiv preprint arXiv:2309.12236*, 2023.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740, 2023.

Kamil Ciosek, Nicolò Felicioni, Sina Ghiassian, Juan Elenter, Francesco Tonolini, David Gustafsson, Eva Garcia-Martin, Carmen Barcena Gonzalez, and Raphaëlle Bertrand-Lalo. Measuring uncertainty calibration. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4AjfwNnWAV.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR, 2018.

H. Fawzi. Numerical Analysis – Lecture 13, 2023. URL https://www.damtp.cam.ac.uk/user/hf323/M23-II-NA/Lect13.pdf. Lecture Notes.

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.

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.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty with proper scoring rules, 2024. URL https://arxiv.org/abs/2404.12215.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457 – 506, 2019. URL https://api.semanticscholar.org/CorpusID:216465307.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pages 623–631. PMLR, 2017.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in neural information processing systems*, 32, 2019.

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

Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335):1–72, 2023.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36:2795–2823, 2023.

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.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

Patrick Rebeschini. Bernstein's Concentration Inequalities. Fast Rates, 2021. URL `https://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/22/material/lecture07.pdf`. Lecture Notes.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4036–4054. PMLR, 2022.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-order uncertainty quantification: A distance-based approach. *arXiv preprint arXiv:2312.00995*, 2023.

Yusuf Sale, Alireza Javanmardi, and Eyke Hüllermeier. Aleatoric and epistemic uncertainty in conformal prediction. 2025.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*, pages 3459–3467. PMLR, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. *Advances in neural information processing systems*, 32, 2019.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in artificial intelligence*, pages 2282–2292. PMLR, 2023.

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.

A Further Related Work

Second-order uncertainty measures. A complementary line of work asks *what* second-order quantity ought to be measured, rather than how to calibrate it. Wimmer et al. [2023] examine entropy-based decompositions of aleatoric and epistemic uncertainty, Sale et al. [2023] propose Wasserstein-based alternatives, and Hofman et al. [2024] develop scoring-rule decompositions. These works concern what to measure; ours concerns how to calibrate it. The general aleatoric/epistemic distinction we exploit is reviewed in Hüllermeier and Waegeman [2021] and Hüllermeier and Waegeman [2019].

Bayesian and evidential uncertainty decomposition. Several lines of work produce mean-and-uncertainty outputs without calibration guarantees: Bayesian active learning by disagreement [Houlsby et al., 2011], variance decompositions for Bayesian neural networks [Depeweg et al., 2018, Kendall and Gal, 2017], and deep ensembles [Lakshminarayanan et al., 2017]. A complementary family bypasses ensembling by directly parameterising a higher-order output distribution: prior networks [Malinin and Gales, 2018] and evidential deep learning [Sensoy et al., 2018] both place a Dirichlet on the simplex and read epistemic uncertainty from its concentration. Second-order Platt scaling is agnostic to the source of (m, σ^2): it can recalibrate the output of any of these methods post hoc.

Distribution-free uncertainty quantification. Our analysis is distribution-free in the sense that no assumption is placed on f^* or on the marginal of X; the only structural input is the sech perturbation, which is a deployment-side design choice rather than an assumption on the data. A different distribution-free paradigm is conformal prediction [Vovk et al., 2005, Angelopoulos and Bates, 2023], which produces per-sample prediction sets with a marginal coverage guarantee on the label. Conformal set size reflects total predictive uncertainty and does not, on its own, separate aleatoric from epistemic components; recent work that attempts such a split [Sale et al., 2025] does so by importing an external decomposition and using conformal machinery to calibrate it. Second-order Platt scaling is complementary: it targets the moments of the conditional label distribution on the predictor's level sets, yielding a calibrated epistemic-uncertainty scalar rather than a coverage-valid set.

B Details of Experiments

B.1 Experiment 1: Additional Details

DGP and predictor. $X \in \mathbb{R}^4$ is drawn from a 10-component isotropic Gaussian mixture (centres $\sim \mathcal{N}(0, 1.5^2 I)$, unit-variance components); $f^*(x) = \sigma(w^\top x)$ with $w \sim \mathcal{N}(0, I/2)$. The predictor is a 5-member MLP ensemble ($4 \to 64 \to 64 \to 1$, ReLU, BCE, Adam at 10^{-3}, 30 epochs, batch 256) trained once on a fresh sample of 5,000; the score is its mean and unbiased variance, clipped to the feasible region. For each coordinate, with original score s and truncation interval $[a, b]$, draw $U \sim \mathrm{Unif}[0, 1]$ and set

$$\alpha = \arctan\left(e^{(a-s)/h}\right), \qquad \beta = \arctan\left(e^{(b-s)/h}\right), \qquad T = s + h \log \tan(\alpha + U(\beta - \alpha)).$$

Then T has density $k_h(\cdot \mid s)$ on $[a, b]$. The two coordinates use independent draws of U.

Ground truth. For each h, $\mathrm{CE}_2^{\mathrm{pert}}$ is computed by depositing a scrambled-Sobol QMC sample of size 2^{18} on a 1025×257 tensor grid, applying the truncated-sech kernel matrices exactly, and integrating the pointwise loss against the perturbed-score density (trapezoidal rule). Two independent Sobol replicates agree to $\leq 2 \cdot 10^{-5}$, much smaller than any estimator error.

Baselines. All estimators target η_1 by regressing $Y^{(1)}$ on the perturbed score and η_2 by regressing $Y^{(1)}Y^{(2)}$. *Bucketing* uses $K \times K$ axis-aligned cells with $K = \mathrm{clip}(\lceil c\,(n/h^2)^{1/4} \rceil, 2, 200)$, the standard 2D-histogram MSE balance for an $O(1/h)$-Lipschitz target. *Nadaraya–Watson* uses an isotropic Gaussian kernel (truncated at $5b$) of bandwidth $b = c\,h^{1/2} n^{-1/4}$ on the σ^2-rescaled rectangle (multiplying σ^2 by 4 so both axes have unit support). The constant c is selected from $\{0.25, 0.5, 1, 2, 4, 8\}$ on a held-out split by squared-error loss; bucketing dyadically extends the grid if the optimum lies on the boundary.

Practical polynomial estimator. The experiments use the ridge-plus-clipping surrogate described after (6), not the constrained ERM analyzed in Proposition 1. We use tensor Chebyshev features of degree l, fit by ridge with λ scaled by $\mathrm{tr}(G)/d$, switching to dual form when $(l+1)^2 > n$. The candidate degree schedule starts at $l_{\max} = \min(l_{\text{paper}}, L_{\text{cap}})$ with $l_{\text{paper}}(n, h) = \lceil 2 \log n / \log \theta_h \rceil$, where

$$\theta_h = \frac{h\pi}{4} + \sqrt{1 + \frac{h^2\pi^2}{16}},$$

using the theoretical degree from Proposition 1 as an upper cap, and per-h caps $L_{\text{cap}} \in \{44, 88\}$ for $h \in \{1/16, 1/64\}$, then halves down to a floor of 4. The ridge multiplier is selected from $\{10^{-6}, \ldots, 1\}$. *val*: (l, α) chosen on a held-out hyper split. *ridge, CV*: 5-fold CV on the validation split, with closed-form leave-one-out predictions reported on the same split. Per-coordinate predictions clipped to $[0, 1]$.

Sampling and aggregation. For each of 20 seeds we draw three i.i.d. splits (train, hyper, valid) of size 50,000; perturbations are redrawn per h, raw scores and labels are shared across h. For each $n \in \{500, 10^3, 2{\cdot}10^3, 5{\cdot}10^3, 10^4, 2{\cdot}10^4, 5{\cdot}10^4\}$ the estimators use only the first n samples of each split (so per-seed n-curves are coupled). Bands are Student-t 90% intervals across seeds; reported slopes are least-squares fits in $\log_{10}(n)$ vs. $\log_{10}$(mean error).

B.2 Experiment 2: implementation details

Ensemble. 20 MLPs (`4-64-64-1`, ReLU), each trained for 60 epochs of BCE with Adam (lr 10^{-3}, batch 256) on the same 24 (x, y) pairs from the DGP, varying only the initialisation seed. The small per-member training set keeps the disagreement variance informative. Outputs are clipped to the feasible set $\sigma^2 \leq m(1-m)$.

Ground-truth surfaces. For each h, we draw 2^{18} scrambled Sobol points from the DGP and bilinearly deposit $(1, p, p^2)$ on a 1025×257 tensor grid over $[0, 1] \times [0, 1/4]$. Convolving each axis with the 1D sech kernel at bandwidth h yields gridded η_1, η_2 and a perturbed score density; the ground-truth conditional variance and CE_2 at h are read off these surfaces, with bilinear interpolation at arbitrary perturbed scores.

Recalibrator. (M, Σ) is affinely mapped to $[-1, 1]^2$ and lifted to a tensor-Chebyshev basis of degree 12 (169 features). The two regressions for $\hat{\eta}_1, \hat{\eta}_2$ share this basis and are solved by Cholesky on the 169×169 Gram matrix, with only a numerical ridge floor of $10^{-12} \cdot \text{tr}(G)/169$ (no cross-validation, no degree search). Predictions are projected onto the feasible moment region $\hat{\eta}_1 \in [0, 1]$, $\hat{\eta}_2 \in [\hat{\eta}_1^2, \hat{\eta}_1]$ before forming $(\sigma^2)'$.

Variance across seeds. The CE_2 intervals reported in the main text are mean $\pm$ standard deviation over 20 independent draws of the $n{=}2{\cdot}10^4$ calibration set.

B.3 Experiment 3 details

DGP. Four-group mixture with weights $(0.25, 0.25, 0.35, 0.15)$. Groups 0/1 are clear negatives/positives concentrated at $\theta \approx 0.08$ and $\theta \approx 0.92$ (small σ^2); they contribute negligible mass after the borderline filter and exist only so the unfiltered population has full support over m. The two groups that populate the borderline cohort:

- *Group 2 (aleatoric, weight* 0.35*):* $m \sim \mathcal{N}(0.5,\, 0.035^2)$, $\sigma^2 \sim \mathcal{N}(0.010,\, 0.002^2)$, $\theta \sim \mathcal{N}(0.5,\, 0.040^2)$.
- *Group 3 (hidden subtype, weight* 0.15*):* $m \sim \mathcal{N}(0.5,\, 0.035^2)$, $\sigma^2 \sim \mathcal{N}(0.060,\, 0.010^2)$; subtype $b \sim \text{Unif}\{0, 1\}$, then $\theta \sim \mathcal{N}(0.12,\, 0.030^2)$ if $b = 0$, else $\mathcal{N}(0.88,\, 0.030^2)$.

All draws are independently clipped so that $m, \theta \in [0, 1]$ and $\sigma^2 \in [0, 1/4]$. Within the borderline cohort, $\eta_1 \approx 0.5$ for both groups while $\eta_2 \approx 0.25$ versus $\eta_2 \approx 0.4$, so the two cannot be separated from m alone.

Calibrators. Tensor-product Chebyshev features ($T_0, \ldots, T_d$ on $2m-1 \in [-1, 1]$ and on $8\sigma^2-1 \in [-1, 1]$) with $d_{1\text{D}} = 6$ and per-axis $d_{2\text{D}} = 4$ (25 features), fit by ridge ERM with ridge 10^{-4}. Calibration labels are 2-snapshots $Y^{(1)}, Y^{(2)} \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$; first-order calibrators regress $Y^{(1)}$ and set $\hat{\eta}_2 := \hat{\eta}_1^2$, second-order calibrators jointly regress $(Y^{(1)},\, Y^{(1)}Y^{(2)})$. Predicted moments are projected onto $\{\hat{\eta}_1 \in [0, 1],\ \hat{\eta}_1^2 \leq \hat{\eta}_2 \leq \hat{\eta}_1\}$ (Jensen plus the Bernoulli upper bound $\theta^2 \leq \theta$). Each method's patient score is $2\hat{\eta}_2 - 2\hat{\eta}_1 + 0.5 - c$.

Methods. Seven curves in Figure 4: (i) raw plug-in from m ($\hat{\eta}_1 := m$, $\hat{\eta}_2 := m^2$); (ii) raw plug-in from (m, σ^2) ($\hat{\eta}_1 := m$, $\hat{\eta}_2 := m^2 + \sigma^2$, then project); (iii)/(iv) 1D/2D first-order Platt; (v)/(vi) 1D/2D second-order Platt; (vii) oracle, scoring with the true θ.

Replication. 200 independent repeats with seeds $0, \ldots, 199$; pre-filter sizes $N_{\text{cal}} = 3000$, $N_{\text{eval}} = 8000$ (about half pass the borderline filter under the group weights); threshold grid of 71 points in $[-0.05, 0.30]$. Bands in Fig. 4 are ± 1.96 SEM across repeats. We did not tune the polynomial degrees, ridge, or filter range; the values above were our first choice. Source code is included with the submission.

B.4 Experiment 4: details

Data. Weather Sentiment AMT: per-(worker, task) categorical responses with majority-vote gold. We binarise as one-vs-rest on the gold class; the figure uses class 1. Tasks with fewer than 5 votes are dropped, leaving 300 tasks with 20 votes each.

Cohort, provisional label, and θ. Let $\hat{p}_i$ be the empirical positive rate of the binary votes for item i. The aleatoric cohort is $\{i : |\hat{p}_i - \frac{1}{2}| \le 0.22\}$, the hidden-subtype cohort is $\{i : \hat{p}_i \le 0.12 \text{ or } \hat{p}_i \ge 0.88\}$, and the two are balanced to equal size. Each item is assigned a provisional binary label prov_i: aleatoric items use $\text{Bernoulli}(\frac{1}{2})$; hidden items use the empirical majority for half (chosen uniformly) and its complement for the other half. The agreement probability of a fresh worker is then $\theta_i = \hat{p}_i$ if $\text{prov}_i = 1$, else $1 - \hat{p}_i$. Aleatoric items therefore have $\theta_i \approx \frac{1}{2}$, hidden items θ_i near 0 or 1; both cohorts have $\mathbb{E}\theta_i \approx \frac{1}{2}$ but very different $\mathbb{E}\theta_i^2$.

Score construction (XOR mode). Let $\text{lo} = \frac{1}{2} - s_{\text{gap}}/2$, $\text{hi} = \frac{1}{2} + s_{\text{gap}}/2$ with $s_{\text{gap}} = 0.35$. Draw $b_i \sim \text{Bernoulli}(\frac{1}{2})$ and set $m_i = \text{hi}$ if $b_i = 1$ else lo; set $s_i = \text{hi}$ if $\mathbb{1}[\text{kind}_i = \text{aleatoric}] = b_i$, else lo. Add independent Gaussian noise of std $\sigma_m = 0.06$, $\sigma_s = 0.08$ and clip to $[0, 1]$. By construction the marginal of m and the marginal of s are the same Gaussian-noisy two-mode mixture in both cohorts, while the joint pattern is a cohort-aligned XOR.

Calibrators. For each calibration item, $\widehat{\eta}_1$ is fit to a_i/n_i and $\widehat{\eta}_2$ to $a_i(a_i - 1)/(n_i(n_i - 1))$, the unbiased all-pairs estimators of θ_i and θ_i^2. Feature maps: 1-D is $[1, m, \ldots, m^4]$; 2-D is the total-degree-3 tensor basis in (m, s), 10 features. We use ridge regression with $\lambda = 10^{-3}$. Predictions are clipped to $\widehat{\eta}_1 \in [0, 1]$ and projected to the admissible region $\widehat{\eta}_1^2 \le \widehat{\eta}_2 \le \widehat{\eta}_1$.

Methods. Eight ranking scores: raw $2m(1 - m)$; raw s; raw $1 - s$; 1-D first-order $2\widehat{\eta}_1^{\text{1D}}(1 - \widehat{\eta}_1^{\text{1D}})$; 1-D second-order $2\widehat{\eta}_1^{\text{1D}} - 2\widehat{\eta}_2^{\text{1D}}$; 2-D first-order $2\widehat{\eta}_1^{\text{2D}}(1 - \widehat{\eta}_1^{\text{2D}})$; 2-D second-order $2\widehat{\eta}_1^{\text{2D}} - 2\widehat{\eta}_2^{\text{2D}}$ (primary); oracle $g(\theta_i) = 2\theta_i(1 - \theta_i)$, shown for upper-bound reference only.

Splits and reporting. Per repeat, items are partitioned disjointly into selection (25%), calibration (35%), evaluation (rest). Cohort thresholds, s_{gap}, noise scales, score-mode, and target class are chosen on the selection split by a held-out audit-AUC lift criterion; the final figure uses 1000 fresh draws of the calibration/evaluation split at the chosen configuration $(\text{extreme_thr}, \text{ale_width}, \sigma_m, \sigma_s, s_{\text{gap}}) = (0.12, 0.22, 0.06, 0.08, 0.35)$. Yield at audit fraction f is $100 \cdot \overline{g(\theta_i)}$ over the top-$\lceil f n_{\text{ev}} \rceil$ evaluation items by score; bands are $\pm 1.96\,\text{SEM}$ across repeats. Source code reproducing all numbers and figures is included with the submission.

Per-method audit-yield summary. At a 5% audit budget the seven non-oracle methods are well separated into three tiers: 2-D second-order Platt at 46.4 disagreements per 100 audited items; 2-D first-order Platt at 30.9; and the five 1-D and raw baselines (raw m, raw s, raw $1 - s$, 1-D first-order, 1-D second-order) clustered around $\sim$26. The oracle $g(\theta_i) = 2\theta_i(1 - \theta_i)$ achieves 49.8. The paired AUC lift of 2-D second-order Platt over each of the six non-oracle baselines has a 95% CI strictly above zero, with per-baseline win fractions ≥ 0.97 across the 1000 evaluation repeats — i.e., the 2-D-second-order ordering audit-AUC dominates each baseline's ordering on essentially every draw, not just on average. The collapse of all 1-D and raw curves onto a common envelope is expected by construction: the XOR score design makes the cohort indicator information-theoretically unrecoverable from any single coordinate, so no 1-D calibrator can do better than chance at distinguishing aleatoric from hidden-subtype items.

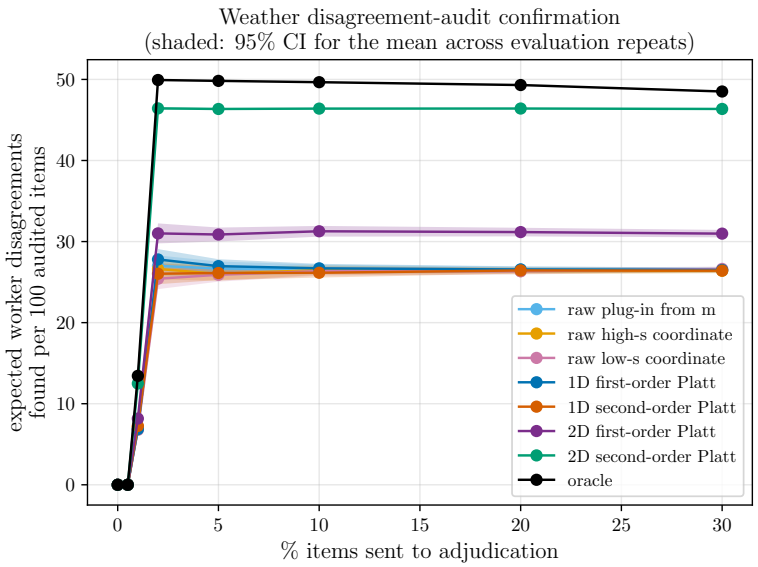


Figure 5: **Experiment 4: full audit-yield curves.** Yield (expected fresh-worker disagreements found per 100 audited items) vs. audit budget on the Weather Sentiment AMT cohort, over 1000 evaluation repeats. 2-D second-order Platt is the only non-oracle method that exploits the joint (m, s) structure required by the decision problem; all 1-D and raw baselines collapse onto a common curve because the 1-D marginals are identical between the aleatoric and hidden-subtype cohorts by construction. Bands are ± 1.96 SEM.

C Choosing the perturbation bandwidth

In practice, we recommend choosing h from a small, prespecified dyadic grid, for example

$$\mathcal{H} = \{2^{-4}, 2^{-5}, \ldots, 2^{-10}\}.$$

On a held-out selection split, evaluate the paired change in AUROC, or preferably the relevant deployment utility, relative to the unperturbed predictor. Because the deployed predictor is randomized, this comparison can be averaged over several perturbation draws. Select the largest h whose upper confidence bound on performance degradation is below a prespecified tolerance. When no application-specific tolerance is available, a simple default is the one-standard-error rule: choose the largest h whose performance is within one standard error of the unperturbed predictor.

This rule reflects the statistical tradeoff. Smaller h changes the predictor less, but narrows the analyticity strip and increases the required polynomial degree and finite-sample constants. Larger h makes estimation easier, but perturbs the predictor more. Performance and calibration estimates at neighboring grid values provide a direct sensitivity check.

If h is selected on data independent of the calibration sample, Proposition 1 applies conditionally on the selected value. For a grid of M prespecified values, the fixed-h tail bounds in the proof can also be made simultaneous by applying each at failure probability δ/M and taking a union bound. This replaces $\log(1/\delta)$ by $\log(M/\delta)$, adding only a term of order $\sqrt{\log M/n}$ to the high-probability estimation error.

After selecting h, the theoretical degree

$$l_{\text{th}} = \left\lceil \frac{2 \log n}{\log \theta_h} \right\rceil$$

can be used as an upper cap. The degree and ridge parameter can then be selected by held-out validation or cross-validation, as in Appendix B.1.

D Why Structure Beyond Continuity Is Necessary

Over the unrestricted class of data-generating distributions, the minimax error for estimating the unperturbed CE_2 does not vanish. To see how the first-order obstruction transfers, restrict to $f^*(X) \in \{0, 1\}$ and set

$$\sigma^2(X) = m(X)(1 - m(X)).$$

In this construction, conditioning on $S = (m, \sigma^2)$ is equivalent to conditioning on m, while $v = m$ and $\eta_2(S) = \eta_1(S)$. Therefore

$$\mathrm{CE}_2 = 2\mathbb{E}[|\mathbb{E}[Y \mid m] - m|] = 2\mathrm{CE}_1.$$

A uniformly vanishing estimator of unrestricted unperturbed CE_2 would thus give a uniformly consistent estimator, and hence a consistent test, for unrestricted first-order calibration. This contradicts the first-order testing impossibility under continuity alone [Lee et al., 2023, Proposition 9].

E Proofs

E.1 Auxiliary lemma

Lemma 3 (Positive real part)**.** *For any $a \in \mathbb{R}$ and $|b| < \pi/2$:*

$$\mathrm{Re}\big(\mathrm{sech}(a + ib)\big) = \frac{\cosh(a)\cos(b)}{|\cosh(a + ib)|^2} > 0.$$

Proof. $\cosh(a + ib) = \cosh(a)\cos(b) + i\sinh(a)\sin(b)$. So

$$\mathrm{Re}(1/\cosh(a + ib)) = \mathrm{Re}(\overline{\cosh(a + ib)})/|\cosh(a + ib)|^2 = \cosh(a)\cos(b)/|\cosh(a + ib)|^2.$$

Since $\cosh(a) > 0$ for all real a and $\cos(b) > 0$ for $|b| < \pi/2$, the result follows. □

E.2 Proof of analyticity of calibration functions (lemma 1)

Proof. We first give the argument for η_1 in 1D, and then give the separate 2D argument.

The perturbed calibration function is $\eta_1(t) = N(t)/D(t)$ (see appendix D.1.1 of the work by Ciosek et al. [2026]). Here

$$N(t) = \int_0^1 \eta_1^{\mathrm{orig}}(s)\,\mathrm{sech}\left(\frac{t - s}{h}\right)\,\mathrm{d}\tilde{p}(s), \tag{13}$$

$$D(t) = \int_0^1 \mathrm{sech}\left(\frac{t - s}{h}\right)\,\mathrm{d}\tilde{p}(s), \tag{14}$$

Here $d\tilde{p}(s) := Z(s, h)^{-1}\, dP_{\mathrm{orig}}(s)$, where P_{orig} is the original score law. This is a finite positive measure because $Z(\cdot, h)$ is continuous and strictly positive on $[0, 1]$. No density assumption is needed.

Analyticity of N and D. For each fixed $s \in [0, 1]$, the function $z \mapsto \mathrm{sech}((z - s)/h)$ is analytic in $|\mathrm{Im}(z)| < h\pi/2$ (since $\mathrm{sech}(w) = 1/\cosh(w)$ has its nearest poles at $w = \pm i\pi/2$, i.e., at $z = s \pm ih\pi/2$). Moreover, for $z = t + ib$ with $t \in [0, 1]$ and $|b| \le h\pi/2 - \delta$ (any $\delta > 0$):

$$|\mathrm{sech}((z - s)/h)| = \frac{1}{|\cosh((t - s)/h + ib/h)|} \le \frac{1}{\cos(b/h) \cdot 1} \le \frac{1}{\cos(\pi/2 - \delta/h)},$$

which is a uniform bound in s and t. Since $0 \le \eta_1^{\mathrm{orig}} \le 1$ and $\tilde{p}$ is a finite measure, the integrands are dominated on compact subsets of the strip by an integrable constant. Integration therefore preserves analyticity, so $N(z)$ and $D(z)$ are analytic in the strip.

Non-vanishing of D. For $z = t + ib$ with $|b| < h\pi/2$:

$$\mathrm{Re}\big(D(t + ib)\big) = \int_0^1 \mathrm{Re}\left(\mathrm{sech}\left(\frac{t + ib - s}{h}\right)\right)\,\mathrm{d}\tilde{p}(s) > 0,$$

by Lemma 3: the integrand is strictly positive and $\tilde{p}$ has positive mass. Therefore $\mathrm{Re}\, D(z) > 0$, so $D(z) \neq 0$ and $\eta_1 = N/D$ is analytic in the strip.

2D extension. Fix $0 < \rho < h\pi/4$. Let $u = (u_1, u_2) = (m, \sigma^2)$ and $z = (z_1, z_2)$. Absorb the original score law and the two truncation normalizers into a finite positive measure $d\mu(u)$ on the feasible set $\mathcal{F} \subseteq [0,1] \times [0, 1/4]$. For $j = 1, 2$ write

$$K_z(u) := \prod_{r=1}^{2} \operatorname{sech}\left(\frac{z_r - u_r}{h}\right), \qquad D(z) := \int_{\mathcal{F}} K_z(u)\, d\mu(u), \qquad N_j(z) := \int_{\mathcal{F}} \eta_j^{\text{orig}}(u) K_z(u)\, d\mu(u).$$

Then $\eta_j(z) = N_j(z)/D(z)$ on the real rectangle. On compact subsets of

$$\Omega_\rho := \{z \in \mathbb{C}^2 : |\operatorname{Im} z_1| < \rho, |\operatorname{Im} z_2| < \rho\},$$

the kernels are uniformly bounded and holomorphic, so dominated convergence gives joint holomorphy of N_j and D.

It remains to show that D does not vanish and that N_j/D is bounded. If $w = a + ib$ with $|b| < \pi/2$, then

$$|\arg \operatorname{sech}(w)| = |\arg \cosh(w)| = |\arctan\{\tanh(a)\tan(b)\}| \le |b|.$$

Hence, for z with $|\operatorname{Im} z_r| \le \rho$,

$$|\arg K_z(u)| \le \frac{|\operatorname{Im} z_1| + |\operatorname{Im} z_2|}{h} \le \frac{2\rho}{h} < \frac{\pi}{2},$$

and therefore

$$\operatorname{Re} K_z(u) \ge \cos(2\rho/h)|K_z(u)|.$$

Consequently

$$\operatorname{Re} D(z) = \int_{\mathcal{F}} \operatorname{Re} K_z(u)\, d\mu(u) \ge \cos(2\rho/h) \int_{\mathcal{F}} |K_z(u)|\, d\mu(u) > 0,$$

so $D(z) \neq 0$. Since $0 \le \eta_j^{\text{orig}} \le 1$,

$$|N_j(z)| \le \int_{\mathcal{F}} |K_z(u)|\, d\mu(u) \le M_\rho \operatorname{Re} D(z) \le M_\rho |D(z)|, \qquad M_\rho := \sec(2\rho/h).$$

Thus, on the closed product sub-strip $|\operatorname{Im} z_r| \le \rho$,

$$|\eta_j(z)| \le M_\rho.$$

□

E.3 Bernstein Ellipse Theorem in our notation (lemma 2 in main text)

Proof. Define

$$g(x) := f\left(\frac{x+1}{2}\right), \qquad x \in [-1, 1].$$

Since f is analytic in the strip

$$\{z \in \mathbb{C} : |\mathrm{Im}(z)| < \rho\},$$

the function g is analytic in the strip

$$S := \{x \in \mathbb{C} : |\mathrm{Im}(x)| < 2\rho\},$$

because

$$\mathrm{Im}\left(\frac{x+1}{2}\right) = \frac{\mathrm{Im}(x)}{2},$$

so $x \in S$ implies

$$\left|\mathrm{Im}\left(\frac{x+1}{2}\right)\right| < \rho.$$

Moreover, $|g(x)| \le M$ on S.

Now let $B(a)$ be the Bernstein ellipse from Theorem 3.17 of the lecture notes by Fawzi [2023]:

$$B(a) = \left\{x + iy \in \mathbb{C} : \frac{x^2}{\cosh^2(a\pi)} + \frac{y^2}{\sinh^2(a\pi)} < 1\right\}.$$

If $z = x + iy \in B(a)$, then

$$\frac{y^2}{\sinh^2(a\pi)} < 1,$$

hence

$$|\mathrm{Im}(z)| = |y| < \sinh(a\pi).$$

Choose $a > 0$ so that

$$\sinh(a\pi) = 2\rho.$$

Then every $z \in B(a)$ satisfies

$$|\mathrm{Im}(z)| < 2\rho,$$

so

$$B(a) \subset S.$$

Therefore g is analytic on $B(a)$ and satisfies

$$\sup_{z \in B(a)} |g(z)| \leq M.$$

By the Bernstein ellipse Theorem [Fawzi, 2023, Theorem 3.17], for every $l \geq 0$ there exists a polynomial p_l of degree l such that

$$\sup_{x \in [-1,1]} |g(x) - p_l(x)| \leq 2M \, \frac{c^{l+1}}{1-c}, \qquad c = e^{-a\pi}.$$

Now define

$$q_l(t) := p_l(2t-1), \qquad t \in [0,1].$$

Then q_l is a polynomial of degree l, and with $x = 2t - 1$,

$$\sup_{t \in [0,1]} |f(t) - q_l(t)| = \sup_{x \in [-1,1]} |g(x) - p_l(x)|.$$

Finally, set $\theta := e^{a\pi}$. Since $\sinh(a\pi) = 2\rho$, we have

$$\theta = e^{a\pi} = \sinh(a\pi) + \sqrt{1 + \sinh^2(a\pi)} = 2\rho + \sqrt{4\rho^2 + 1},$$

and therefore $c = \theta^{-1}$. Hence

$$2M \, \frac{c^{l+1}}{1-c} = 2M \, \frac{\theta^{-(l+1)}}{1-\theta^{-1}} = 2M \, \frac{\theta^{-l}}{\theta - 1}.$$

This proves the claim. □

E.4 Proof of corollary 1 in the main text

Proof. Indeed, map $\mathcal{R}$ to $[-1,1]^2$ by $x = 2m - 1$ and $y = 8\sigma^2 - 1$. The mapped function is bounded by M_ρ on a product of Bernstein ellipses with common parameter θ_ρ; for the second coordinate the available strip is larger, so using the first-coordinate value θ_ρ is conservative. The tensor Chebyshev coefficients satisfy $|a_{kl}| \leq 4M_\rho \theta_\rho^{-(k+l)}$. Hence the tensor Chebyshev truncation $p_{j,\ell}$ satisfies

$$\|p_{j,\ell} - \eta_j\|_\infty \leq a_\ell, \qquad a_\ell := \frac{8M_\rho\theta_\rho}{(\theta_\rho - 1)^2}\theta_\rho^{-\ell},$$

where the sup norm is over $\mathcal{R}$.

Since $\eta_j \in [0,1]$, this implies $p_{j,\ell} \in [-a_\ell, 1 + a_\ell]$ on $\mathcal{R}$. Define

$$q_{j,\ell} := \frac{p_{j,\ell} + a_\ell}{1 + 2a_\ell}.$$

This is a polynomial of the same degree and satisfies $0 \leq q_{j,\ell} \leq 1$ on $\mathcal{R}$. Moreover,

$$\begin{aligned} \|q_{j,\ell} - \eta_j\|_\infty &\leq \frac{\|p_{j,\ell} - \eta_j\|_\infty + a_\ell \|1 - 2\eta_j\|_\infty}{1 + 2a_\ell} \\ &\leq \frac{2a_\ell}{1 + 2a_\ell} \leq 2a_\ell = B_\rho \theta_\rho^{-\ell}. \end{aligned}$$

□

E.5 Covering Number Lemma

Lemma 4. *Let*

$$V_l := \operatorname{span}\{u_1^i u_2^j : 0 \le i, j \le l\}$$

be the degree-l tensor polynomial space on $\mathcal{R} = [0,1] \times [0,1/4]$, equipped with

$$\|p\|_\infty := \sup_{u \in \mathcal{R}} |p(u)|.$$

Let

$$B_l := \{p \in V_l : \|p\|_\infty \le 1\}, \qquad \mathcal{P}_l := \{\operatorname{clip}_{[0,1]}(p) : p \in B_l\}.$$

Then, for every $0 < \varepsilon \le 1$,

$$N(\varepsilon, \mathcal{P}_l, \|\cdot\|_\infty) \le \left(\frac{3}{\varepsilon}\right)^{(l+1)^2}.$$

Equivalently,

$$\log N(\varepsilon, \mathcal{P}_l, \|\cdot\|_\infty) \le (l+1)^2 \log\left(\frac{3}{\varepsilon}\right).$$

Proof. The space V_l has dimension $d = (l+1)^2$. The standard volumetric bound for the unit ball of a d-dimensional normed space gives

$$N(\varepsilon, B_l, \|\cdot\|_\infty) \le \left(1 + \frac{2}{\varepsilon}\right)^d \le \left(\frac{3}{\varepsilon}\right)^d,$$

where the last inequality uses $0 < \varepsilon \le 1$.

The scalar clipping map is 1-Lipschitz. Therefore, for any $p, q \in B_l$,

$$\|\operatorname{clip}_{[0,1]}(p) - \operatorname{clip}_{[0,1]}(q)\|_\infty \le \|p - q\|_\infty.$$

Consequently, clipping every element of an ε-net of B_l produces an ε-net of $\mathcal{P}_l$ of the same size. Thus

$$N(\varepsilon, \mathcal{P}_l, \|\cdot\|_\infty) \le N(\varepsilon, B_l, \|\cdot\|_\infty) \le \left(\frac{3}{\varepsilon}\right)^{(l+1)^2}.$$

Taking logarithms proves the second statement. □

E.6 More Auxiliary Lemmas

Lemma 5. *We have*

$$\mathbb{E}[h_{\hat{f}}] \le 4\varepsilon_l^2 + \frac{406\, d \log(3n) + 408}{n}.$$

Proof. To keep constants explicit, we use the covering bound from lemma 4 in the form

$$N(\delta, \mathcal{P}_l, \|\cdot\|_\infty) \le \left(\frac{3}{\delta}\right)^d, \qquad 0 < \delta \le 1,$$

with $d = (l+1)^2$.

Set

$$\delta_n := \frac{1}{n},$$

and let $\mathcal{G}_n \subset \mathcal{P}_l$ be a δ_n-cover of $\mathcal{P}_l$ in $\|\cdot\|_\infty$. Then

$$|\mathcal{G}_n| \le N(\delta_n, \mathcal{P}_l, \|\cdot\|_\infty) \le (3n)^d.$$

For every $f \in \mathcal{P}_l$, choose $g_f \in \mathcal{G}_n$ such that

$$\|f - g_f\|_\infty \le \delta_n.$$

Recall that

$$h_f(S,Z) = (f(S) - Z)^2 - (f^*(S) - Z)^2.$$

If $f, g \in \mathcal{P}_l$, then

$$h_f - h_g = (f - g)(f + g - 2Z),$$

hence, since $f, g \in [0,1]$ and $Z \in \{0,1\}$,

$$|h_f - h_g| \le 2|f - g| \le 2\|f - g\|_\infty.$$

Therefore

$$|Ph_f - Ph_g| \le 2\|f - g\|_\infty, \qquad |P_n h_f - P_n h_g| \le 2\|f - g\|_\infty,$$

where $P_n h := \frac{1}{n}\sum_{i=1}^n h(S_i, Z_i)$. In particular, for the ERM $\hat{f}$, since $P_n h_{\hat{f}} \le 0$,

$$P_n h_{g_{\hat{f}}} \le P_n h_{\hat{f}} + 2\delta_n \le \frac{2}{n},$$

and

$$Ph_{\hat{f}} \le Ph_{g_{\hat{f}}} + 2\delta_n.$$

Now fix $g \in \mathcal{G}_n$, and write

$$\mu_g := Ph_g = \mathbb{E}[h_g].$$

By Step 2(b), $h_g \in [-4, 4]$, so

$$(\mu_g - h_g) \le \mu_g + 4 \le 8.$$

Applying Bernstein's inequality [Rebeschini, 2021, Corollary 7.3] to the random variables

$$X_i := \mu_g - h_g(S_i, Z_i),$$

which satisfy $X_i - \mathbb{E}X_i \le 8$, gives for every $t > 0$,

$$\Pr(\mu_g - P_n h_g \ge t) \le \exp\left(-\frac{nt^2}{2(\mathrm{Var}(h_g) + \frac{8}{3}t)}\right).$$

By Step 2(c),

$$\mathrm{Var}(h_g) \le \mathbb{E}[h_g^2] \le 32\mu_g + 64\varepsilon_l^2.$$

Hence

$$\Pr(\mu_g - P_n h_g \ge t) \le \exp\left(-\frac{nt^2}{2(32\mu_g + 64\varepsilon_l^2 + \frac{8}{3}t)}\right).$$

Suppose now that

$$\mu_g \ge \frac{4}{n} \qquad \text{and} \qquad \mu_g \ge 4\varepsilon_l^2.$$

Then $t := \mu_g - \frac{2}{n}$ satisfies $t \ge \mu_g/2$, and therefore

$$\Pr\left(P_n h_g \le \frac{2}{n}\right) \le \exp\left(-\frac{n(\mu_g/2)^2}{2(32\mu_g + 64\varepsilon_l^2 + \frac{8}{3}\mu_g)}\right).$$

Using $64\varepsilon_l^2 \le 16\mu_g$, the denominator is at most

$$2\left(32\mu_g + 16\mu_g + \frac{8}{3}\mu_g\right) = \frac{304}{3}\mu_g,$$

so

$$\Pr\left(P_n h_g \le \frac{2}{n}\right) \le \exp\left(-\frac{3n\mu_g}{1216}\right) \le \exp\left(-\frac{n\mu_g}{406}\right).$$

Now fix $u \ge 0$, and define

$$r(u) := \max\left\{\frac{4}{n},\ 4\varepsilon_l^2,\ \frac{406\big(d\log(3n) + u\big)}{n}\right\}.$$

If $\mu_g \geq r(u)$, then

$$\Pr\left(P_n h_g \leq \frac{2}{n}\right) \leq \exp\left(-\frac{nr(u)}{406}\right) \leq \exp(-d\log(3n) - u) = (3n)^{-d}e^{-u}.$$

Taking a union bound over all $g \in \mathcal{G}_n$, and using $|\mathcal{G}_n| \leq (3n)^d$, we obtain

$$\Pr\left(\exists g \in \mathcal{G}_n : \ Ph_g \geq r(u) \text{ and } P_n h_g \leq \frac{2}{n}\right) \leq e^{-u}.$$

On the complementary event, the particular net point $g_{\hat{f}}$ associated with $\hat{f}$ cannot satisfy $Ph_{g_{\hat{f}}} \geq r(u)$, because we already know $P_n h_{g_{\hat{f}}} \leq 2/n$. Thus

$$Ph_{g_{\hat{f}}} < r(u),$$

and hence

$$Ph_{\hat{f}} \leq Ph_{g_{\hat{f}}} + \frac{2}{n} < r(u) + \frac{2}{n}.$$

We have proved that for every $u \geq 0$,

$$\Pr\left(Ph_{\hat{f}} > \frac{2}{n} + \max\left\{\frac{4}{n},\ 4\varepsilon_l^2,\ \frac{406\big(d\log(3n) + u\big)}{n}\right\}\right) \leq e^{-u}.$$

Since

$$\max\left\{\frac{4}{n},\ 4\varepsilon_l^2,\ \frac{406\big(d\log(3n) + u\big)}{n}\right\} \leq 4\varepsilon_l^2 + \frac{406\, d\log(3n)}{n} + \frac{406u}{n},$$

we get

$$\Pr\left(Ph_{\hat{f}} > 4\varepsilon_l^2 + \frac{406\, d\log(3n) + 2}{n} + \frac{406u}{n}\right) \leq e^{-u}.$$

Integrating this tail bound,

$$\mathbb{E}[Ph_{\hat{f}}] \leq 4\varepsilon_l^2 + \frac{406\, d\log(3n) + 2}{n} + \frac{406}{n}\int_0^\infty e^{-u}\, du,$$

hence

$$\mathbb{E}[h_{\hat{f}}] = \mathbb{E}[Ph_{\hat{f}}] \leq 4\varepsilon_l^2 + \frac{406\, d\log(3n) + 408}{n}.$$

This is the desired fast-rate bound. □

E.7 Proof of Main Result (Proposition 1)

Proof. Since both $\hat{\eta}_1$ and $\hat{\eta}_2$ are ERM in the same class $\mathcal{P}_l$ with $\{0,1\}$-valued responses, we give the argument once for a generic calibration function η estimated by $\hat{f} = \arg\min_{f\in\mathcal{P}_l}(1/n)\sum(Z_i - f(S_i))^2$, where $Z_i \in \{0,1\}$.

Step 1: Approximation error. By Corollary 1, there exists a degree-l polynomial q_l satisfying $0 \leq q_l \leq 1$ on $\mathcal{R}$ and

$$\|q_l - \eta\|_\infty \leq \varepsilon_l \leq B_h\theta_h^{-l} \leq B_h n^{-2}.$$

Since $\|q_l\|_\infty \leq 1$ and $\mathrm{clip}_{[0,1]}(q_l) = q_l$, we have $q_l \in \mathcal{P}_l$. Set $f^* := q_l$.

Step 2: Excess risk via the Bernstein condition. Define the excess loss $h_f(S,Z) = (f(S) - Z)^2 - (f^*(S) - Z)^2$. We verify three properties:

(a) *Excess risk equals excess L^2 error:* $\mathbb{E}[h_f] = \|f - \eta\|^2_{L^2(P)} - \|f^* - \eta\|^2_{L^2(P)}$.

This follows by expanding: $h_f = (f - f^*)(f + f^* - 2Z)$, so $\mathbb{E}[h_f \mid S] = (f - f^*)(f + f^* - 2\eta) = (f - \eta)^2 - (f^* - \eta)^2$.

(b) *Boundedness:* $|h_f| \le |f - f^*| \cdot |f + f^* - 2Z| \le 1 \cdot 4 = 4$, since $f, f^* \in [0,1]$ and $Z \in \{0,1\}$.

(c) *Bernstein condition:* $\mathbb{E}[h_f^2] \le 16\|f - f^*\|_{L^2(P)}^2$, since $h_f^2 \le 16(f - f^*)^2$. Furthermore,

$$\begin{aligned}\|f - f^*\|_{L^2}^2 &= \|(f - \eta) - (f^* - \eta)\|_{L^2}^2 \\ &\le 2\|f - \eta\|_{L^2}^2 + 2\|f^* - \eta\|_{L^2}^2 \\ &= 2\mathbb{E}[h_f] + 2\|f^* - \eta\|_{L^2}^2 + 2\|f^* - \eta\|_{L^2}^2 \\ &\le 2\mathbb{E}[h_f] + 4\varepsilon_l^2.\end{aligned}$$

So $\mathbb{E}[h_f^2] \le 32\mathbb{E}[h_f] + 64\varepsilon_l^2$.

Step 3: Fast rate with explicit constants. Since, by Step 2(a),

$$\mathbb{E}[h_{\hat{f}}] = \|\hat{f} - \eta\|_{L^2(P)}^2 - \|f^* - \eta\|_{L^2(P)}^2,$$

and $\|f^* - \eta\|_{L^2(P)} \le \varepsilon_l$, it follows from lemma 5 that

$$\mathbb{E}\|\hat{f} - \eta\|_{L^2(P)}^2 \le 5\varepsilon_l^2 + \frac{406\, d \log(3n) + 408}{n}.$$

Step 4: From L^2 error to CE_2. Apply the bound from Step 3 separately to the two calibration functions η_1, η_2. For $j \in \{1, 2\}$, let

$$A_n := 5\varepsilon_l^2 + \frac{406\, d \log(3n) + 408}{n},$$

so that

$$\mathbb{E}\big[\|\hat{\eta}_j - \eta_j\|_{L^2(P)}^2\big] \le A_n.$$

By Cauchy–Schwarz and then Jensen's inequality,

$$\mathbb{E}\big[\|\hat{\eta}_j - \eta_j\|_{L^1(P)}\big] \le \mathbb{E}\big[\|\hat{\eta}_j - \eta_j\|_{L^2(P)}\big] \le \Big(\mathbb{E}\big[\|\hat{\eta}_j - \eta_j\|_{L^2(P)}^2\big]\Big)^{1/2} \le \sqrt{A_n}.$$

Now write

$$\widetilde{v} := \widetilde{m}^{\,2} + \widetilde{\sigma}_f^2.$$

The perturbed second-order calibration error is

$$CE_2^{\mathrm{pert}} = \mathbb{E}\big[|\eta_1(\widetilde{S}) - \widetilde{m}|\big] + \mathbb{E}\big[|\eta_2(\widetilde{S}) - \widetilde{v}|\big].$$

Define the plug-in population analogue

$$CE_2^{\mathrm{plug}} := \mathbb{E}\big[|\hat{\eta}_1(\widetilde{S}) - \widetilde{m}|\big] + \mathbb{E}\big[|\hat{\eta}_2(\widetilde{S}) - \widetilde{v}|\big].$$

Using the reverse triangle inequality,

$$\big||a - q| - |b - q|\big| \le |a - b|,$$

we obtain

$$\big|CE_2^{\mathrm{plug}} - CE_2^{\mathrm{pert}}\big| \le \|\hat{\eta}_1 - \eta_1\|_{L^1(P)} + \|\hat{\eta}_2 - \eta_2\|_{L^1(P)}.$$

Taking expectations and using the previous bound gives

$$\mathbb{E}\big[\big|CE_2^{\mathrm{plug}} - CE_2^{\mathrm{pert}}\big|\big] \le 2\sqrt{A_n}.$$

That is,

$$\mathbb{E}\big[\big|CE_2^{\mathrm{plug}} - CE_2^{\mathrm{pert}}\big|\big] \le 2\left(5\varepsilon_l^2 + \frac{406\, d \log(3n) + 408}{n}\right)^{1/2}.$$

Finally, with

$$l = \left\lceil \frac{2 \log n}{\log \theta_h} \right\rceil, \qquad d = (l+1)^2,$$

and $\varepsilon_l \le B_h \theta_h^{-l}$ from Corollary 1, we have

$$\varepsilon_l^2 \le B_h^2\, \theta_h^{-2l} \le B_h^2 n^{-4},$$

and therefore

$$\mathbb{E}\big[\big|CE_2^{\mathrm{plug}} - CE_2^{\mathrm{pert}}\big|\big] \le 2\left(5B_h^2 n^{-4} + \frac{406\,(l+1)^2 \log(3n) + 408}{n}\right)^{1/2}.$$

In particular,

$$\mathbb{E}\big[\big|CE_2^{\mathrm{plug}} - CE_2^{\mathrm{pert}}\big|\big] \le 2\left(5B_h^2 n^{-4} + \frac{406\left(\frac{2\log n}{\log \theta_h} + 2\right)^2 \log(3n) + 408}{n}\right)^{1/2}.$$

This proves the $O_h(\log^{3/2} n/\sqrt{n})$ bound for $\mathrm{CE}_2^{\mathrm{plug}}$.

Step 5: From population to empirical CE. The preceding steps bound $\mathbb{E}\big[\big|\mathrm{CE}_2^{\mathrm{plug}} - \mathrm{CE}_2^{\mathrm{pert}}\big|\big]$, where $\mathrm{CE}_2^{\mathrm{plug}} = \mathbb{E}\big[|\hat\eta_1(\tilde S) - \tilde m|\big] + \mathbb{E}\big[|\hat\eta_2(\tilde S) - \tilde v|\big]$ is evaluated under the *population* measure. The estimator (7) uses the *empirical* measure over the same sample, so we must control the difference.

For any fixed $f \in \mathcal{P}_l$ the map $s \mapsto |f(s) - \tilde m|$ takes values in $[0,1]$. Define the function class

$$\mathcal{G}_l \;=\; \big\{\,(s,\tilde m,\tilde v) \mapsto |f(s)-\tilde m| + |g(s)-\tilde v| \;:\; f,g \in \mathcal{P}_l\big\}.$$

Every member of $\mathcal{G}_l$ is bounded in $[0, \frac{9}{4}]$. We now bound its covering number.

For any fixed constant c, the map $t \mapsto |t-c|$ is 1-Lipschitz, so for any $f_1, f_2 \in \mathcal{P}_l$ and any c,

$$\sup_s \big|\,|f_1(s)-c| - |f_2(s)-c|\,\big| \;\le\; \sup_s |f_1(s) - f_2(s)| \;=\; \|f_1 - f_2\|_\infty.$$

Therefore, if $\{f^{(1)}, \dots, f^{(N_1)}\}$ is an ϵ-cover of $\mathcal{P}_l$ in $\|\cdot\|_\infty$, then for every $f \in \mathcal{P}_l$ there exists $f^{(k)}$ with $\sup_s \big|\,|f(s)-c| - |f^{(k)}(s)-c|\,\big| \le \epsilon$.

Now consider a pair $(f,g) \in \mathcal{P}_l \times \mathcal{P}_l$ and choose respective ϵ-approximants $f^{(k)}, g^{(j)}$. By the triangle inequality applied to the sum defining $\mathcal{G}_l$,

$$\big\|\,(|f(\cdot)-\tilde m| + |g(\cdot)-\tilde v|) - (|f^{(k)}(\cdot)-\tilde m| + |g^{(j)}(\cdot)-\tilde v|)\,\big\|_\infty \;\le\; \|f-f^{(k)}\|_\infty + \|g-g^{(j)}\|_\infty \;\le\; 2\epsilon.$$

Hence an ϵ-cover of $\mathcal{P}_l$ of size N_1 induces a 2ϵ-cover of $\mathcal{G}_l$ of size N_1^2. Substituting $\epsilon/2$ and applying lemma 4:

$$\log N(\epsilon,\, \mathcal{G}_l,\, \|\cdot\|_\infty) \;\le\; 2\log N\big(\tfrac{\epsilon}{2},\, \mathcal{P}_l,\, \|\cdot\|_\infty\big) \;\le\; 2(l+1)^2 \log\frac{6}{\epsilon} \;=\; 2d\log\frac{6}{\epsilon}.$$

We now bound

$$U \;:=\; \sup_{h \in \mathcal{G}_l} \left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z)]\right|,$$

where $Z_i = (S_i, \tilde m_i, \tilde v_i)$ are i.i.d.

Discretisation. Fix $\epsilon = 1/n$ and let $\mathcal{N} \subset \mathcal{G}_l$ be an $(1/n)$-cover of $\mathcal{G}_l$ in $\|\cdot\|_\infty$. By the covering-number bound derived above,

$$|\mathcal{N}| \;\le\; N\big(\tfrac{1}{n},\, \mathcal{G}_l,\, \|\cdot\|_\infty\big) \;\le\; (6n)^{2(l+1)^2}.$$

Concentration on the cover. For any fixed $h \in \mathcal{N}$, the random variables $h(Z_1), \dots, h(Z_n)$ are i.i.d. in $[0, 9/4]$. By Hoeffding's inequality (with range $\frac{9}{4} - 0 = \frac{9}{4}$), for every $t > 0$:

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z)]\right| > t\right) \;\le\; 2\exp\!\left(-\frac{2nt^2}{(\frac{9}{4})^2}\right) = 2\exp\!\left(-\frac{32}{81}nt^2\right).$$

A union bound over $\mathcal{N}$ gives

$$\Pr\left(\max_{h\in\mathcal{N}}\left|\frac{1}{n}\sum_{i=1}^{n}h(Z_i)-\mathbb{E}[h(Z)]\right|>t\right) \;\le\; 2\,|\mathcal{N}|\,\exp\!\left(-\frac{32}{81}nt^2\right) \tag{15}$$

$$\le\; 2\,(6n)^{2(l+1)^2}\exp\!\left(-\frac{32}{81}nt^2\right). \tag{16}$$

Extension to all of $\mathcal{G}_l$. For any $h\in\mathcal{G}_l$, choose $h'\in\mathcal{N}$ with $\|h-h'\|_\infty\le 1/n$. Then

$$\left|\frac{1}{n}\sum_i h(Z_i)-\mathbb{E}[h(Z)]\right| \;\le\; \left|\frac{1}{n}\sum_i h'(Z_i)-\mathbb{E}[h'(Z)]\right|+2\|h-h'\|_\infty \;\le\; \left|\frac{1}{n}\sum_i h'(Z_i)-\mathbb{E}[h'(Z)]\right|+\frac{2}{n},$$

where the first inequality is the triangle inequality applied to both the empirical and population terms. Taking the supremum over h, and combining with (15), we obtain a tail bound for $U:=\sup_{h\in\mathcal{G}_l}\left|\frac{1}{n}\sum_{i=1}^{n}h(Z_i)-\mathbb{E}[h(Z)]\right|$:

$$\Pr\!\left(U>t+\frac{2}{n}\right) \;\le\; 2\,(6n)^{2(l+1)^2}\exp\!\left(-\frac{32}{81}nt^2\right), \qquad t>0. \tag{17}$$

Expectation bound. Using $\mathbb{E}[U]=\int_0^\infty \Pr(U>s)\,ds$ and the substitution $s=t+2/n$,

$$\mathbb{E}[U] \;\le\; \frac{2}{n}+\int_0^\infty \min\!\Big(1,\; 2(6n)^{2(l+1)^2}\,e^{-\frac{32}{81}nt^2}\Big)\,dt.$$

Define

$$t^* \;=\; \sqrt{\frac{\frac{81}{16}(l+1)^2\log(6n)+\frac{81}{32}\log 2}{n}},$$

chosen so that $2(6n)^{2(l+1)^2}\exp(-\frac{32}{81}n(t^*)^2)=1$. Splitting the integral at t^*:

$$\int_0^\infty \min(\cdots)\,dt \;\le\; t^*+\int_{t^*}^\infty 2(6n)^{2(l+1)^2}\,e^{-\frac{32}{81}nt^2}\,dt.$$

For the tail integral, note that for $t\ge t^*>0$ we have $1\le t/t^*$, so

$$\int_{t^*}^\infty e^{-\frac{32}{81}nt^2}\,dt \;\le\; \frac{1}{t^*}\int_{t^*}^\infty t\,e^{-\frac{32}{81}nt^2}\,dt \;=\; \frac{1}{t^*}\left[-\frac{81}{64n}e^{-\frac{32}{81}nt^2}\right]_{t^*}^\infty \;=\; \frac{81e^{-\frac{32}{81}n(t^*)^2}}{64nt^*}.$$

Multiplying by the prefactor and using the definition of t^*:

$$2(6n)^{2(l+1)^2}\cdot\frac{81e^{-\frac{32}{81}n(t^*)^2}}{64nt^*} \;=\; \frac{81}{64nt^*}.$$

Therefore

$$\mathbb{E}[U] \;\le\; \frac{2}{n}+t^*+\frac{81}{64nt^*} \;=\; \frac{2}{n}+\sqrt{\frac{\frac{81}{16}(l+1)^2\log(6n)+\frac{81}{32}\log 2}{n}}+\frac{81}{64\sqrt{n\big(\frac{81}{16}(l+1)^2\log(6n)+\frac{81}{32}\log 2\big)}}. \tag{18}$$

Application to $\widehat{\mathrm{CE}}_2$. Since $(\hat\eta_1,\hat\eta_2)\in\mathcal{P}_l\times\mathcal{P}_l$ regardless of the data realisation (the ERM always returns an element of the hypothesis class), the random function $z\mapsto|\hat\eta_1(z)-\tilde m|+|\hat\eta_2(z)-\tilde v|$ lies in $\mathcal{G}_l$, and the supremum U controls this particular instance. Concretely, $|\widehat{\mathrm{CE}}_2-\mathrm{CE}_2^{\mathrm{plug}}|\le U$, so (18) gives

$$\mathbb{E}\big[\big|\widehat{\mathrm{CE}}_2-\mathrm{CE}_2^{\mathrm{plug}}\big|\big] \;\le\; \frac{2}{n}+\sqrt{\frac{\frac{81}{16}(l+1)^2\log(6n)+\frac{81}{32}\log 2}{n}}+\frac{81}{64\sqrt{n\big(\frac{81}{16}(l+1)^2\log(6n)+\frac{81}{32}\log 2\big)}}..$$

Combining with Step 4. By the triangle inequality,

$$\mathbb{E}\big[\big|\widehat{\mathrm{CE}}_2-\mathrm{CE}_2^{\mathrm{pert}}\big|\big] \;\le\; \mathbb{E}\big[\big|\widehat{\mathrm{CE}}_2-\mathrm{CE}_2^{\mathrm{plug}}\big|\big]+\mathbb{E}\big[\big|\mathrm{CE}_2^{\mathrm{plug}}-\mathrm{CE}_2^{\mathrm{pert}}\big|\big].$$

The first term is bounded above. The second term was bounded in Step 4:

$$\mathbb{E}\big[\big|\mathrm{CE}_2^{\mathrm{plug}} - \mathrm{CE}_2^{\mathrm{pert}}\big|\big] \;\le\; 2\left(5B_h^2 n^{-4} + \frac{406(l+1)^2\log(3n)+408}{n}\right)^{1/2}.$$

With $l = \lceil 2\log n/\log\theta_h\rceil$, both terms are $O_h(\log^{3/2} n/\sqrt{n})$: the dominant contribution in each is the $\sqrt{(l+1)^2\log n\,/\,n}$ factor, and $(l+1)^2 = O_h(\log^2 n)$. This proves the rate claimed in equation (8). □

E.8 Proof of Corollary 2

Proof. Condition on the calibration data $\mathcal{D}$, including their perturbations. Let P denote the law of a fresh perturbed score S, independent of $\mathcal{D}$. Until the final averaging step, all expectations below are conditional on $\mathcal{D}$.

First control the effect of clipping the second moment. Fix a score s and write $u = \eta_1(s)$ and $v = \eta_2(s)$. The true moments satisfy $u^2 \le v \le u$. Let π_s be clipping to $[a(s)^2, a(s)]$. Since clipping is 1-Lipschitz,

$$|b'(s) - v| \le |b(s) - v| + |\pi_s(v) - v|.$$

If v lies inside this interval, the last term is zero. If $v < a(s)^2$, it is at most $|a(s)^2 - u^2| \le 2|a(s) - u|$; if $v > a(s)$, it is at most $|u - a(s)|$. Therefore

$$|b'(s) - \eta_2(s)| \le |\hat\eta_2(s) - \eta_2(s)| + 2|\hat\eta_1(s) - \eta_1(s)|.$$

The first and second moments reported by $\hat T$ are $a(S)$ and $b'(S)$, respectively. Both are functions of $\hat T$, since its second reported moment equals its variance plus its squared mean. The tower property and conditional Jensen inequality consequently give

$$\begin{aligned}\mathrm{CE}_2(\hat T) &\le \mathbb{E}[|\eta_1(S) - a(S)|] + \mathbb{E}[|\eta_2(S) - b'(S)|]\\ &\le 3\|\hat\eta_1 - \eta_1\|_{L^1(P)} + \|\hat\eta_2 - \eta_2\|_{L^1(P)}.\end{aligned}$$

Finally, take expectation over $\mathcal{D}$. Step 4 of the proof of Proposition 1 in Appendix E.7 bounds the expected value of each $L^1(P)$ error by $O_h(\log^{3/2} n/\sqrt{n})$, proving the claim. □

F Moments vs Wasserstein Distance for Defining Calibration

In this appendix we relate Definition 1 to the kth-order calibration framework of Ahdritz et al. [2025], specializing to binary labels and $k = 2$. For this comparison, both criteria use the same fixed partition. Our average moment error is quantitatively comparable to the corresponding average Wasserstein error, but does not generally control the worst-case error over all partition cells. Conditional expectations in Definition 1 are understood as versions defined almost surely under the score law; their values on null sets do not affect CE_2.

F.1 Preliminaries: symmetrized snapshots

A 2-*snapshot* for an instance x is a pair $(y_1, y_2) \in \{0,1\}^2$ with $y_1, y_2 \mid x \overset{\text{iid}}{\sim} f^*(x)$. Following Ahdritz et al. [2025], we identify each snapshot with its *empirical distribution* (normalized histogram):

$$\mathrm{Unif}(y_1, y_2)(y) \;:=\; \frac{1}{2}\sum_{i=1}^{2}\mathbf{1}[y_i = y], \qquad y \in \{0,1\}. \tag{19}$$

Each element of $\{0,1\}^2$ receives equal weight $1/k = 1/2$ in the histogram; the name Unif refers to this equal weighting, *not* to the resulting distribution being uniform over Y. Concretely,

$$(0,0) \mapsto (1,0), \qquad (0,1),(1,0) \mapsto \left(\tfrac{1}{2},\tfrac{1}{2}\right), \qquad (1,1) \mapsto (0,1),$$

where we write distributions as vectors (p_0, p_1) over $\{0,1\}$. Since the snapshot is i.i.d., the order of the tuple is immaterial, so the symmetrised snapshot space is

$$Y^{(2)} = \left\{0,\, \tfrac{1}{2},\, 1\right\} \;\subset\; \Delta Y,$$

a set of three atoms. Distributions over $Y^{(2)}$ are parametrised by a probability vector $(q_0, q_{1/2}, q_1)$.

F.2 The two definitions

Our definition (Definition 1, specialized to a partition). Fix a partition $[\cdot]$ of $\mathcal{X}$. Define the calibration functions

$$\eta_1(s) = \mathbb{E}[Y \mid S(X) = s], \qquad \eta_2(s) = \mathbb{E}[f^*(X)^2 \mid S(X) = s],$$

with $S(X) = (m(X), \sigma^2(X))$ and $v = m^2 + \sigma^2$. The second-order calibration error is

$$\mathrm{CE}_2 = \mathbb{E}\big[|\eta_1(S) - m|\big] + \mathbb{E}\big[|\eta_2(S) - v|\big]. \tag{20}$$

When the conditioning is taken with respect to the partition (i.e. $\eta_1([x]) = \mathbb{E}[f^*(X) \mid X \in [x]]$ and $\eta_2([x]) = \mathbb{E}[f^*(X)^2 \mid X \in [x]]$), the pointwise moment errors on a partition element $[x]$ are

$$\delta_1([x]) = \big|\eta_1([x]) - m([x])\big|, \qquad \delta_2([x]) = \big|\eta_2([x]) - v([x])\big|. \tag{21}$$

Ahdritz et al.'s definition 2.3. A 2-snapshot predictor $g\colon \mathcal{X} \to \Delta Y^{(2)}$ is *ε-second-order calibrated* with respect to a partition $[\cdot]$ if, for all $x \in \mathcal{X}$,

$$W_1\big(g(x),\ \mathrm{proj}_2 f^*([x])\big) \ \le\ \varepsilon, \tag{22}$$

where W_1 denotes the Wasserstein-1 distance on $\Delta\Delta Y$ with the ℓ_1 distance on ΔY as the ground metric, and $\mathrm{proj}_2 f^*([x])$ is the true distribution of symmetrised 2-snapshots drawn from $[x]$.

Remark 2 (Worst-case vs. average)**.** *Ahdritz et al. [2025] require* (22) *for* all *equivalence classes* $[x]$ *(worst-case). They note in footnote 4 that a natural relaxation requires* (22) *only with high probability over equivalence classes drawn according to the marginal* D_X*. We call this the* relaxed *variant: for a given* $\delta > 0$*,*

$$\Pr_{[x]\sim D_X}\Big[W_1\big(g(x),\ \mathrm{proj}_2 f^*([x])\big) > \varepsilon\Big] \ \le\ \delta. \tag{23}$$

Our quantity CE_2 *is an* average *(expectation under* D_X*) and therefore corresponds most naturally to this relaxed variant.*

F.3 Relating the two definitions

We now establish the precise relationship. The key is a pointwise sandwich between the Wasserstein error and the moment errors.

Lemma 6 (Pointwise sandwich)**.** *Fix a partition element* $[x]$*. Let* $\delta_1([x]), \delta_2([x])$ *be as in* (21) *and* $W_1([x])$ *as in* (22)*. Then*

$$\delta_1([x]) + \delta_2([x]) \ \le\ \frac{3}{2}\, W_1([x]) \ \le\ 3\big(\delta_1([x]) + \delta_2([x])\big). \tag{24}$$

Proof. A distribution over $Y^{(2)} = \{0, \frac{1}{2}, 1\}$ is a vector $(q_0, q_{1/2}, q_1)$. The predictor's output corresponds to $(q_0, q_{1/2}, q_1)$ and the true snapshot distribution to $(p_0, p_{1/2}, p_1)$, with

$$p_1 = \mu_2, \quad p_{1/2} = 2(\mu_1 - \mu_2), \quad p_0 = 1 - 2\mu_1 + \mu_2,$$

where $\mu_j = \mathbb{E}_{x\sim[x]}[f^*(x)^j]$. For the predictor's output, let $m([x])$ and $v([x]) = m([x])^2 + \sigma^2([x])$ denote the shared values of m and v on the partition element $[x]$. The predicted snapshot distribution is

$$q_1 = v([x]), \quad q_{1/2} = 2\big(m([x]) - v([x])\big), \quad q_0 = 1 - 2\,m([x]) + v([x]).$$

Write $\Delta_j = p_j - q_j$ for $j \in \{0, \frac{1}{2}, 1\}$, so that $\Delta_0 + \Delta_{1/2} + \Delta_1 = 0$. We now compute $W_1([x])$. Each atom of $Y^{(2)}$ is a distribution over $\{0, 1\}$: $0 \leftrightarrow (1, 0)$, $\frac{1}{2} \leftrightarrow (\frac{1}{2}, \frac{1}{2})$, $1 \leftrightarrow (0, 1)$. The ground metric is the ℓ_1 distance inherited from ΔY:

$$d(0, \tfrac{1}{2}) = 1, \qquad d(\tfrac{1}{2}, 1) = 1, \qquad d(0, 1) = 2.$$

Since $d(0, 1) = d(0, \frac{1}{2}) + d(\frac{1}{2}, 1)$, it is never cheaper to transport mass directly between the outer atoms than to route it through $\frac{1}{2}$. The optimal plan is therefore determined by two edge flows: let ϕ_1

be the net signed flow from atom 0 to atom $\frac{1}{2}$, and ϕ_2 from atom $\frac{1}{2}$ to atom 1. Mass balance at each atom gives $\phi_1 = \Delta_0$ and $\phi_2 = \Delta_0 + \Delta_{1/2} = -\Delta_1$, so

$$W_1([x]) \;=\; |\phi_1| \cdot d(0, \tfrac{1}{2}) \;+\; |\phi_2| \cdot d(\tfrac{1}{2}, 1) \;=\; |\Delta_0| + |\Delta_1|.$$

Meanwhile, the moment errors are

$$\delta_1([x]) = |\mu_1 - m| = \tfrac{1}{2}|\Delta_{1/2} + 2\Delta_1| = \tfrac{1}{2}|\Delta_0 - \Delta_1|, \qquad \delta_2([x]) = |\mu_2 - v| = |\Delta_1|.$$

Upper bound. By the triangle inequality, $|\Delta_0| \le |\Delta_0 - \Delta_1| + |\Delta_1| = 2\delta_1 + \delta_2$, so

$$W_1 = |\Delta_0| + |\Delta_1| \le 2\delta_1 + 2\delta_2.$$

This is the same as

$$\frac{3}{2} W_1 \le 3\delta_1 + 3\delta_2$$

Lower bound. We have $\delta_2 = |\Delta_1| \le W_1$ and $\delta_1 = \frac{1}{2}|\Delta_0 - \Delta_1| \le \frac{1}{2}(|\Delta_0| + |\Delta_1|) = \frac{1}{2} W_1$, so $\delta_1 + \delta_2 \le \frac{3}{2} W_1$. □

We can now state the three main properties.

Property 1 (Equivalence at zero error)**.** $\mathrm{CE}_2 = 0$ *if and only if, for D_X-almost every equivalence class $[x]$, $W_1([x]) = 0$.*

Proof. By (24), $W_1([x]) = 0$ iff $\delta_1([x]) + \delta_2([x]) = 0$. Since $\mathrm{CE}_2 = \mathbb{E}_{[x]}[\delta_1([x]) + \delta_2([x])]$, the expectation vanishes iff the integrand vanishes D_X-a.e. □

Property 2 (Ahdritz approximate calibration implies bounded CE_2)**.** *If g is ε-second-order calibrated in the sense of* (22) *(worst-case), then*

$$\mathrm{CE}_2 \;\le\; \tfrac{3}{2}\,\varepsilon.$$

Proof. For every $[x]$, the lower bound in (24) gives $\delta_1([x]) + \delta_2([x]) \le \frac{3}{2}\varepsilon$. Averaging over D_X yields the claim. □

Property 3 (Bounded CE_2 implies relaxed Ahdritz calibration)**.** *If $\mathrm{CE}_2 \le \varepsilon$, then for every $\delta > 0$,*

$$\Pr_{[x]\sim D_X}\big[W_1([x]) > 2\varepsilon/\delta\big] \;\le\; \delta.$$

In particular, g satisfies the relaxed variant (23) *with parameters (ε', δ) for $\varepsilon' = 2\varepsilon/\delta$.*

However, $\mathrm{CE}_2 \le \varepsilon$ does not *in general imply a uniform (worst-case) bound on $W_1([x])$ over all partition elements, since CE_2 is an average quantity.*

Proof. The upper bound in (24) gives $W_1([x]) \le 2\,(\delta_1([x]) + \delta_2([x]))$, so

$$\mathbb{E}_{[x]}[W_1([x])] \;\le\; 2\,\mathrm{CE}_2 \;\le\; 2\varepsilon.$$

Markov's inequality yields the stated tail bound. □

Remark 3. *If one adopts the relaxed (high-probability) variant of Ahdritz et al.'s definition, then Properties 2 and 3 together show that the two notions are quantitatively comparable: writing $\overline{W}_1 = \mathbb{E}_{[x]}[W_1([x])]$ for the average Wasserstein error,*

$$\tfrac{1}{2}\,\overline{W}_1 \;\le\; \mathrm{CE}_2 \;\le\; \tfrac{3}{2}\,\overline{W}_1.$$

G Matching Lower Bound on Rate

In this section, we prove Proposition 2.

Proof.

G.1 The two-point family

Take $\mathcal{X}$ a singleton (suppress X) and use the boundary constant predictor

$$S(x) \;\equiv\; (m_0,\, \sigma_0^2) \;:=\; (1,\, 0).$$

Since $m_0 = 1$ forces $m_0(1 - m_0) = 0$, the only valid value is $\sigma_0^2 = 0$; the score is therefore the deterministic point $(1, 0)$, which lies in the feasible region. The sech perturbation produces $\widetilde{S} = (\widetilde{m}, \widetilde{\sigma^2})$, where $\widetilde{m}$ has density $k_h(\cdot \mid 1)$ on $[0, 1]$ and $\widetilde{\sigma^2}$ has density $k_h(\cdot \mid 0)$ on $[0, 1/4]$, independently. The law of $\widetilde{S}$ is the *same* under both P_0 and P_1.

For $b \in \{0, 1\}$, let P_b be the law in which $f^* \equiv p_b$, with

$$p_0 \;:=\; \tfrac{1}{16}, \qquad p_1 \;:=\; \tfrac{1}{16} + \varepsilon,$$

for an $\varepsilon \in (0, 1/16]$ to be chosen later. So $Y_i^{(j)} \overset{\text{iid}}{\sim} \text{Bernoulli}(p_b)$, independent of $\widetilde{S}$.

G.2 The calibration error has a clean closed form

Since S is constant, $\widetilde{S}$ is independent of the labels and the calibration functions are themselves constants:

$$\eta_1(\widetilde{S}) = p_b, \qquad \eta_2(\widetilde{S}) = p_b^2.$$

(Both are trivially analytic, satisfying the structural assumptions of the main paper.) Substituting:

$$\text{CE}_2^{\text{pert}}(P_b) \;=\; \underbrace{\mathbb{E}\big[|p_b - \widetilde{m}|\big]}_{=:\,\phi(p_b)} + \underbrace{\mathbb{E}\big[|p_b^2 - W|\big]}_{=:\,\psi(p_b)}, \qquad W := \widetilde{m}^{\,2} + \widetilde{\sigma^2}. \tag{25}$$

G.3 Key lemma: stochastic dominance, uniform in h

Lemma 7. *For $m_0 = 1$ and every $h > 0$,*

$$F_{\widetilde{m}}(p) \;\le\; p \qquad \textit{for all } p \in [0, 1].$$

Equivalently, $\widetilde{m}$ stochastically dominates $U[0, 1]$.

Proof. Let $G(x) := \int_0^x \text{sech}(v)\, dv$. We have $G(0) = 0$, $G'(x) = \text{sech}(x) > 0$, and $G''(x) = -\text{sech}(x)\tanh(x) \le 0$ for $x \ge 0$, so G is concave on $[0, \infty)$.

Substituting $v = (1 - t)/h$ in the integral defining $F_{\widetilde{m}}$:

$$F_{\widetilde{m}}(p) \;=\; \frac{\int_0^p \text{sech}\big((t-1)/h\big)\, dt}{\int_0^1 \text{sech}\big((u-1)/h\big)\, du} \;=\; 1 - \frac{G\big((1-p)/h\big)}{G(1/h)}.$$

By concavity of G on $[0, \infty)$ and $G(0) = 0$, for any $\theta \in [0, 1]$ and $x \ge 0$,

$$G(\theta x) \;=\; G\big(\theta x + (1-\theta)\cdot 0\big) \;\ge\; \theta\, G(x) + (1-\theta)\, G(0) \;=\; \theta\, G(x).$$

Setting $\theta = 1 - p$ and $x = 1/h$: $G\big((1-p)/h\big) \ge (1-p)\, G(1/h)$, so $F_{\widetilde{m}}(p) \le 1 - (1-p) = p$. □

G.4 The gap

We bound ϕ and ψ separately by integrating their derivatives.

First moment. The function $\phi(p) = \mathbb{E}|p - \widetilde{m}|$ is convex and differentiable everywhere on $(0, 1)$ (since $\widetilde{m}$ has a continuous density), with

$$\phi'(p) \;=\; 2F_{\widetilde{m}}(p) - 1 \;\le\; 2p - 1$$

by Lemma 7. Integrating from p_0 to $p_1 = p_0 + \varepsilon$,

$$\begin{aligned}\phi(p_1) - \phi(p_0) \;&\le\; \int_{p_0}^{p_1} (2p - 1)\, dp \;=\; (p_1^2 - p_0^2) - (p_1 - p_0) \\ &=\; \varepsilon\,(p_0 + p_1 - 1) \;\le\; \varepsilon\,\big(\tfrac{1}{16} + \tfrac{1}{8} - 1\big) \;=\; -\tfrac{13}{16}\,\varepsilon,\end{aligned}$$

using $p_1 \le 1/8$.

Second moment. Differentiation under the expectation (justified by the Lipschitz continuity in p of $|p^2 - W|$ and the continuous density of W) gives

$$\psi'(p) \;=\; 2p\,\big(2F_W(p^2) - 1\big), \qquad |\psi'(p)| \le 2p.$$

Hence

$$\big|\psi(p_1) - \psi(p_0)\big| \;\le\; \int_{p_0}^{p_1} 2p\,dp \;=\; p_1^2 - p_0^2 \;\le\; 2p_1\,\varepsilon \;\le\; \tfrac{\varepsilon}{4}.$$

Combining. From (25),

$$\mathrm{CE}_2^{\mathrm{pert}}(P_0) - \mathrm{CE}_2^{\mathrm{pert}}(P_1) \;=\; \big(\phi(p_0) - \phi(p_1)\big) + \big(\psi(p_0) - \psi(p_1)\big) \;\ge\; \tfrac{13}{16}\varepsilon - \tfrac{1}{4}\varepsilon \;=\; \tfrac{9}{16}\varepsilon \;\ge\; \tfrac{\varepsilon}{2}. \tag{26}$$

G.5 KL bound

The perturbed scores carry no information about b (their law is identical under P_0, P_1 and they are independent of the labels), so each observation under P_b is informationally equivalent to two i.i.d. $\mathrm{Bernoulli}(p_b)$ draws. With both $p_0 = 1/16$ and $p_1 \le 1/8$, $p_1(1-p_1) \ge (1/16)(15/16) \ge 1/18$, so

$$\mathrm{KL}\big(\mathrm{Bernoulli}(p_0) \,\|\, \mathrm{Bernoulli}(p_1)\big) \;\le\; \frac{(p_1 - p_0)^2}{p_1(1-p_1)} \;\le\; 18\,\varepsilon^2.$$

Tensorizing,

$$\mathrm{KL}\big(P_0^{\otimes n} \,\|\, P_1^{\otimes n}\big) \;\le\; 36\,n\,\varepsilon^2. \tag{27}$$

G.6 Le Cam

Le Cam's two-point lemma [Polyanskiy and Wu, 2025, Chapter 31]: for any real-valued estimator $\widehat{\theta}$,

$$\max_{b\in\{0,1\}} \mathbb{E}_{P_b^{\otimes n}}\big[|\widehat{\theta} - \theta(P_b)|\big] \;\ge\; \tfrac{1}{2}\,|\theta(P_1) - \theta(P_0)| \cdot \big(1 - \mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})\big).$$

By Pinsker and (27), $\mathrm{TV} \le \sqrt{\mathrm{KL}\,/2} \le \sqrt{18\,n\,\varepsilon^2} = 3\,\varepsilon\,\sqrt{2n}$.

Choosing ε. Set

$$\varepsilon \;:=\; \frac{1}{16\sqrt{n}}.$$

Then $\mathrm{TV} \le 1/2$, hence $1 - \mathrm{TV} \ge 1/2$. For $n \ge 1$ we have $\varepsilon \le 1/16$, so the gap bound (26) applies. Combining (26) with Le Cam:

$$\max_{b\in\{0,1\}} \mathbb{E}_{P_b^{\otimes n}}\big[|\widehat{\mathrm{CE}}_2 - \mathrm{CE}_2^{\mathrm{pert}}(P_b)|\big] \;\ge\; \tfrac{1}{2}\cdot\tfrac{\varepsilon}{2}\cdot\tfrac{1}{2} \;=\; \frac{\varepsilon}{8} \;=\; \frac{1}{128\sqrt{n}}.$$

Setting $c := 1/128$, an absolute constant, independent of h, proves Proposition 2. □

H Compute Resources

The expected compute requirements are modest: all experiments can be run on a standard laptop.

I Open Questions

We think the following questions would be useful avenues for further work.

1. Extension to multiclass classification (we handle the binary case).
2. Determine whether the $\log^{3/2} n$ factor in the upper bound is necessary. It arises from the structure of our proof, but is not present in the lower bound.
3. Study the statistical and practical tradeoffs among perturbation, partitioning, and explicit structural assumptions on the calibration functions.

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: The abstract and introduction accurately summarize the paper's main contributions: a bucket-free definition of second-order calibration, a near-$\tilde{O}(1/\sqrt{n})$ estimator based on analyticity of the sech perturbation kernel, and a second-order analogue of Platt scaling. These claims are matched by the formal statements and empirical results in the main text.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The paper discusses the main assumptions and practical limitations of the approach, including the reliance on 2-snapshots, the use of sech perturbation, the dependence on the bandwidth choice, and the current focus on binary classification.

   Guidelines:

   - The answer [N/A] 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 theoretical results are stated together with their assumptions, and complete proofs are provided in the appendix. The main theoretical guarantees appear in Lemma 1, Lemma 2, Proposition 1, and Proposition 3.

   Guidelines:

   - The answer [N/A] 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: The paper provides full details for the synthetic and real-data experiments, including model architecture, datasets, snapshot construction, baselines, evaluation metrics, hyperparameters, and perturbation settings. The paper describes the setup sufficiently for reproducing the reported results.

   Guidelines:

   - The answer [N/A] 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**

   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: The code is included in the supplementary material. The paper also describes the datasets and experimental setup in enough detail to support faithful reproduction.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training and evaluation protocol, including datasets, train/test splits, ensemble sizes, perturbation bandwidths, polynomial degrees, optimizers, and other hyperparameters used in the experiments.

   Guidelines:

   - The answer [N/A] 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: The experimental results include confidence intervals and repeated-trial statistics for the key quantitative comparisons.

   Guidelines:

- The answer [N/A] 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: The appendix describes the compute requirements for the experiments.

   Guidelines:

   - The answer [N/A] 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: To the best of our knowledge, the research described in the paper conforms to the NeurIPS Code of Ethics. The work is primarily methodological and empirical, uses standard public benchmarks and existing annotations, and does not involve new human-subject interventions.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: This paper is primarily a theoretical and methodological contribution on second-order calibration. It does not introduce a specific deployed system or application domain, so broader societal impacts are indirect rather than central to the paper's main contribution.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not introduce or release a high-risk model or scraped dataset. Its main contribution is a calibration method and associated theoretical analysis.

Guidelines:

- The answer [N/A] 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: [Yes]

Justification: The paper cites the datasets and models used in the experiments.

Guidelines:

- The answer [N/A] 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: [No]

    Justification: The paper introduces a new model/methodological contribution, but it does not release a corresponding software or model asset alongside the submission. The method and experimental setup are described in the paper, but no separately distributed asset is provided.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not report new crowdsourcing experiments or new human-subject data collection by the authors. It uses an existing public dataset with human annotations.

    Guidelines:

    - The answer [N/A] 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: [N/A]

Justification: The paper does not involve new human-subjects research conducted by the authors. It relies on existing public annotated data and does not introduce new participant interaction or data collection.

Guidelines:

- The answer [N/A] 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.

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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important, original, or non-standard component of the core methodology in this research.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.