L_2-Regularized Empirical Risk Minimization Guarantees Small Smooth Calibration Error

Masahiro Fujisawa*
The University of Osaka / RIKEN AIP / Lattice Lab, Toyota Motor Corporation
fujisawa@ist.osaka-u.ac.jp

Futoshi Futami*
The University of Osaka / RIKEN AIP / The University of Tokyo
futami.futoshi.es@osaka-u.ac.jp

Abstract

Calibration of predicted probabilities is critical for reliable machine learning, yet it is poorly understood how standard training procedures yield well-calibrated models. This work provides the first theoretical proof that canonical L_2-regularized empirical risk minimization directly controls the smooth calibration error (smCE) without post-hoc correction or specialized calibration-promoting regularizer. We establish finite-sample generalization bounds for smCE based on optimization error, regularization strength, and the Rademacher complexity. We then instantiate this theory for models in reproducing kernel Hilbert spaces, deriving concrete guarantees for kernel ridge and logistic regression. Our experiments confirm these specific guarantees, demonstrating that L_2-regularized ERM can provide a well-calibrated model without boosting or post-hoc recalibration. The source code to reproduce all experiments is available at https://github.com/msfuji0211/erm_calibration.

1 Introduction

Calibration—the alignment between predicted probabilities and the true label frequency—is essential for building reliable machine learning models, especially in risk-sensitive applications [8, 36]. In such settings, calibration metrics are central to assessing model reliability [23, 16, 24], motivating the need for theoretical guarantees on the calibration performance of learned models. To achieve well-calibrated models, a variety of post-hoc recalibration techniques [18, 37, 19] and regularization-based training strategies [24, 25, 9, 30] have been proposed. However, most of these methods are supported mainly by empirical evidence, and the theoretical principles underlying how to design algorithms that achieve both high accuracy and low calibration error remain poorly understood.

This work addresses this problem by providing a theoretical analysis of the relationship between empirical risk minimization (ERM) [28]—a core principle in machine learning—and calibration error. While ERM with proper losses [15, 27] is widely used for classification due to its ability to achieve high accuracy and approximate true conditional probabilities, recent studies have shown that it can still yield poorly calibrated models, particularly in deep learning [18]. This has motivated growing interest in understanding how loss functions and training algorithms influence calibration performance. A key recent development is the notion of the *post-processing gap* [2], which characterizes calibration error as the improvement obtainable through Lipschitz transformations. This perspective has attracted increasing attention [3, 13, 21].

Building on this framework, Błasiok et al. [2] showed that minimizing a proper loss with a structured regularization can yield well-calibrated models. While their analysis provides valuable theoretical insights, it has two key limitations: (1) the optimization is defined over population risk, which is impractical for implementation; and (2) the regularization is abstract and does not correspond to standard training algorithms used in practice.

To address these limitations, in this work, we provide a theoretical analysis showing that calibration error—specifically, the smooth calibration error (CE) [1]—can be effectively controlled via L_2-regularized ERM. This establishes a principled foundation for designing calibration-aware algorithms using standard ERM. Our main contributions are as follows:

*Equal contribution.

We consider binary classification under L_2-regularized ERM with a hypothesis class closed under Lipschitz transformations. In this setting, we show that the smooth CE computed on the training data–referred to as the *training smooth CE*–can be bounded in terms of the regularization coefficient and the optimization error (Theorem 3). We further establish a generalization bound that quantifies the gap between the smooth CE (defined over the population) and training smooth CE via the Rademacher complexity [28] of the hypothesis class (Theorem 4). These results reveal how regularization, optimization, and function class complexity jointly control smooth CE.

To demonstrate practical applicability, we instantiate our theory for models in reproducing kernel Hilbert spaces (RKHS) [28]. Our analysis identifies a key distinction based on the input dimensionality: the RKHS induced by the Laplace kernel in one dimension satisfies our closure assumption. This leads to guarantees with a faster convergence rate for kernel ridge and logistic regression (Corollaries 2 and 5), a setting particularly relevant for the practical task of *recalibration* [37]. In contrast, we prove that this property does not hold in higher dimensions or for other universal kernels like the Gaussian kernel (Theorem 6).

Even in these more general settings where the closure property is absent, we prove that smooth CE can still be controlled for any universal kernel, albeit with a slower convergence rate (Theorem 7). Furthermore, our bounds provide a formal characterization of the bias-variance trade-off governed by the regularization coefficient, where balancing model complexity and goodness-of-fit is critical for achieving a small smooth CE (Corolaries 3 and 6).

In conclusion, our work shifts the perspective on calibration from a post-hoc correction to an intrinsic property of well-posed, regularized learning. We demonstrate that smooth CE can be controlled via the L_2-regularized ERM, providing a principled foundation for designing models that are inherently both accurate and reliable.

2 Preliminaries

We denote random variables by uppercase letters (e.g., X) and deterministic quantities by lowercase letters (e.g., x). The Euclidean inner product and norm are denoted by $\cdot$ and $\|\cdot\|$, respectively.

2.1 Binary Classification

We assume that data points $(X, Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ are drawn from a distribution $\mathcal{D}$, where $\mathcal{X}$ is the input space and $\mathcal{Y} = \{0, 1\}$ is the label space. For a given input $X = x$, the label Y is drawn from a Bernoulli distribution $Y \sim \mathrm{Ber}(v)$ for some $v \in [0, 1]$ with the true conditional probability $f^*(x) = \mathbb{E}[Y \mid x]$. Let $\mathcal{F}$ be a hypothesis class consisting of functions $f : \mathcal{X} \to [0, 1]$. A predictor $f \in \mathcal{F}$ is trained to approximate f^* based on samples drawn from $\mathcal{D}$.

To evaluate the performance of f, we define a loss function $\ell : [0, 1] \times \mathcal{Y} \to \mathbb{R}$, denoted by $\ell(v, y)$. We often adopt a *proper loss* [14] as ℓ, where ℓ is called *proper* if, for any $v, v' \in [0, 1]$, it satisfies $\mathbb{E}_{Y\sim\mathrm{Ber}(v)}[\ell(v, Y)] \leq \mathbb{E}_{Y\sim\mathrm{Ber}(v)}[\ell(v', Y)]$. This paper focuses on the two such losses: the squared loss $\ell_{\mathrm{sq}}(v, y) \coloneqq (y - v)^2$ and the cross-entropy loss $\ell_{\mathrm{xent}}(v, y) \coloneqq -y \log v - (1 - y) \log(1 - v)$. We note that our analysis framework is general and can be extended to other proper losses mentioned by Błasiok et al. [2] (see Appendix B for details).

For ease of discussion, we introduce the *Savage representation* [14] of ℓ_{xent}. Let $\psi : \mathbb{R} \to \mathbb{R}$ be the convex function defined by $\psi(s) \coloneqq \log(1 + e^s)$ for $s \in \mathbb{R}$. The proper scoring rule induced by ψ is then given by $\ell^\psi(s, y) \coloneqq \psi(s) - sy$, known as the logistic loss. Identifying the score with the logit function $g(x) \coloneqq \log \frac{f(x)}{1-f(x)}$, the corresponding probability can be recovered via the sigmoid function $f(x) = \sigma(g(x)) = 1/(1 + \exp(-g(x)))$. his yields the equivalence $\ell^\psi(g(x), y) = \ell_{\mathrm{xent}}(f(x), y)$, a representation that facilitates our subsequent analysis.

2.2 Calibration Metrics and Loss Functions

Given a distribution $\mathcal{D}$, f is *perfectly calibrated* if $\mathbb{E}[Y \mid f(X)] = f(X)$ almost surely. While the common binning-based expected calibration error (ECE) [18] is widely used, it is known to exhibit numerical instabilities and present theoretical difficulties [25, 1] (see Section 4 for examples). We therefore focus on the following *smooth CE* [1], a metric with desirable theoretical properties such as continuity and computational efficiency [1, 22] (see Section 4 for details).

Definition 1 (smooth CE [1])**.** *Let* Lip_L *be the set of all* L*-Lipschitz functions from* $[0, 1]$ *to* $[-1, 1]$*. Then, the smooth CE is defined as*

$$\mathrm{smCE}(f, \mathcal{D}) \coloneqq \sup_{h\in\mathrm{Lip}_{L=1}} \mathbb{E}\left[h(f(X)) \cdot (Y - f(X))\right].$$

For notational simplicity, we denote the class of 1-Lipschitz functions $\mathrm{Lip}_{L=1}$ by Lip_1. Since binning ECE can be both upper- and lower-bounded by smooth CE, we focus on smooth CE (see Section 4). A key property of smooth CE is its characterization via the *post-processing gap* of f, which measures the potential improvement in squared loss through Lipschitz post-processing.

Definition 2 (Post-processing gap under ℓ_{sq} [2]). *Given a predictor $f \in \mathcal{F}$ and a distribution $\mathcal{D}$, the post-processing gap under the squared loss is defined as*

$$\mathrm{pGap}(f, \mathcal{D}) \coloneqq \mathbb{E}[\ell_{\mathrm{sq}}(f(X), Y)] - \inf_{h \in \mathrm{Lip}_{L=1}} \mathbb{E}[\ell_{\mathrm{sq}}(f(X) + h(f(X)), Y)].$$

Theorem 1 (Theorem 2.4 in Błasiok et al. [2]). *For any predictor $f \in \mathcal{F}$ and distribution $\mathcal{D}$,*

$$\mathrm{smCE}(f, \mathcal{D})^2 \leq \mathrm{pGap}(f, \mathcal{D}) \leq 2\mathrm{smCE}(f, \mathcal{D}).$$

This result shows that a predictor is well-calibrated in terms of in terms of $\mathrm{smCE}(f, \mathcal{D})$ if it cannot be improved by simple post-processing.

In classification models, we often minimize the cross-entropy loss, focusing on a logit function. Thus, it is natural to consider the following post-processing gap under ℓ^{ψ}.

Definition 3 (Dual post-processing gap under ℓ^{ψ} [2]). *Let $\mathrm{Lip}_1(\mathbb{R}, [-4, 4])$ denote the class of 1-Lipschitz functions from $\mathbb{R}$ to $[-4, 4]$. Then, the dual post-processing gap is defined as*

$$\mathrm{pGap}^{(\psi,1/4)}(g, \mathcal{D}) \coloneqq \mathbb{E}[\ell^{\psi}(g(X), Y)] - \inf_{h \in \mathrm{Lip}_1(\mathbb{R}, [-4,4])} \mathbb{E}[\ell^{\psi}(g(X) + h(g(X)), Y)].$$

Definition 4 (Dual smooth CE [2]). *Let $\mathrm{Lip}_{1/4}(\mathbb{R}, [-1, 1])$ denote the class of $1/4$-Lipschitz functions from $\mathbb{R}$ to $[-1, 1]$. Given a logit function $g : \mathcal{X} \to \mathbb{R}$ and prediction $f(x) = \sigma(g(x))$, the dual smooth calibration error is defined as*

$$\mathrm{smCE}^{(\psi,1/4)}(g, \mathcal{D}) \coloneqq \sup_{h \in \mathrm{Lip}_{1/4}(\mathbb{R}, [-1,1])} \mathbb{E}\left[h(g(X)) \cdot (Y - f(X))\right].$$

We remark that the choice of the Lipschitz constant $1/4$ and the bounded ranges $[-4, 4]$ and $[-1, 1]$ follow from the smoothness of the softplus function ψ. An analogue of Theorem 1 holds for $\mathrm{pGap}^{(\psi,1/4)}(g, \mathcal{D})$ and $\mathrm{smCE}^{(\psi,1/4)}(g, \mathcal{D})$, as stated in the following corollary.

Corollary 1 (Corollary 2.4 in Błasiok et al. [2]). *The dual post-processing gap and the dual smooth calibration error satisfy the following bounds:*

$$2\mathrm{smCE}^{(\psi,1/4)}(g, \mathcal{D})^2 \leq \mathrm{pGap}^{(\psi,1/4)}(g, \mathcal{D}) \leq 4\mathrm{smCE}^{(\psi,1/4)}(g, \mathcal{D}).$$

Finally, we remark that the original smooth CE is upper bounded by its dual counterpart:

$$\mathrm{smCE}(f, \mathcal{D}) \leq \mathrm{smCE}^{(\psi,1/4)}(g, \mathcal{D}). \tag{1}$$

Błasiok et al. [2] clarified that a certain regularized risk minimization leads to a small smooth CE:

Theorem 2 (Claim 5.1 [2]). *Assume that for any $f \in \mathcal{F}$ and $h \in \mathrm{Lip}_1$, the composition $f + h \circ f$ belongs to $\mathcal{F}$. Let $\mu : \mathcal{F} \to \mathbb{R}^+$ be a complexity measure satisfying $\mu(f + h \circ f) \leq \mu(f) + 1$ for all $f \in \mathcal{F}$ and $h \in \mathrm{Lip}_1$. For any $\lambda > 0$, the minimizer f^* of the regularized minimization problem*

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \ell_{\mathrm{sq}}(f(X_i), Y_i) + \lambda \mu(f)$$

satisfies $\mathrm{pGap}(f^*, \mathcal{D}) \leq \lambda$, and thus $\mathrm{smCE}(f^*, \mathcal{D}) \leq \sqrt{\lambda}$.

This result provides a theoretical basis for optimization-based calibration. Błasiok et al. [2] suggested that structured regularization may align with standard training methods, such as stochastic gradient descent (SGD). However, their formulation relies on *population risk minimization*, which is impractical in real settings, and uses an abstract complexity measure μ that is hard to implement. To address this, the following section investigates whether smooth CE can be controlled through standard learning algorithms using training data. We show that L_2-regularized ERM leads to calibration-aware learning.

3 Regularized Empirical Risk Minimization for Calibration

In this section, we first extend the theoretical framework of Błasiok et al. [2] to the context of ERM (Section 3.1). We then show that two representative frameworks—kernel ridge regression and kernel logistic regression—satisfy the assumptions required for our derived bounds, and we provide smooth CE bounds tailored to each of these methods (Sections 3.2 and 3.3).

3.1 Smooth CE Analysis under Regularized ERM

Regularized ERM on ℓ_{sq}: Let $\mathcal{F}$ be a function class equipped with a norm $\|\cdot\|_{\mathcal{F}}$ such that $\|f\|_{\mathcal{F}} \leq 1$ for all $f \in \mathcal{F}$. Given a training set $S \coloneqq \{(X_i, Y_i)\}_{i=1}^n$ of independently and identically distributed (i.i.d.) samples from the data distribution $\mathcal{D}$, we consider the L_2-regularized ERM:

$$L_n(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{sq}}(f(X_i), Y_i) + \lambda \|f\|_{\mathcal{F}}^2. \tag{2}$$

We define the optimization error as $\mathrm{err}_n(f) \coloneqq L_n(f) - L_n(f_n^*)$, where $f_n^* \coloneqq \mathrm{argmin}_{f\in\mathcal{F}} L_n(f)$.

Analysis of the smooth CE on S: Before proceeding with our analysis, we define the empirical versions of the smooth CE [1], given a training dataset $S = \{(X_i, Y_i)\}_{i=1}^n$, as follows:

$$\mathrm{smCE}(f, S) \coloneqq \sup_{h\in\mathrm{Lip}_1} \frac{1}{n}\sum_{i=1}^{n} \left[h(f(X_i)) \cdot (Y_i - f(X_i))\right].$$

Our first contribution shows that a small optimization error $\mathrm{err}_n(f)$ and a small regularization parameter λ suffice to guarantee a small smooth CE on the training set S.

Theorem 3. *Assume that for any $f \in \mathcal{F}$ and $h \in \mathrm{Lip}_1$, the composite function $f + h \circ f$ also belongs to $\mathcal{F}$. Then, for any $f \in \mathcal{F}$, we have* $\mathrm{smCE}(f, S) \leq \sqrt{\lambda + \mathrm{err}_n(f)}$.

Proof outline. We relate the empirical post-processing gap, $\mathrm{pGap}(f, S)$, to the regularized objective L_n. By adding and subtracting the regularization term, we can upper-bound the gap using the difference in the objective function values at f and $f + h \circ f$ as follows:

$$\begin{aligned}\mathrm{pGap}(f, S) &\leq L_n(f) - \inf_{h\in\mathrm{Lip}_1} L_n(f + h \circ f) + \lambda \\ &\leq L_n(f_n^*) + \mathrm{err}_n(f) - \inf_{h\in\mathrm{Lip}_1} L_n(f + h\circ f) + \lambda.\end{aligned}$$

By applying the definition of the optimizer f_n^* and the composition assumption ($f + h \circ f \in \mathcal{F}$), we obtain

$$L_n(f_n^*) \leq \inf_{h\in\mathrm{Lip}_1} L_n(f + h \circ f), \tag{3}$$

which concludes the proof. □

Our result can be viewed as the ERM counterpart of Theorem 2, providing a bound on the training smooth CE in terms of the optimization error and the regularization coefficient. As in Theorem 2, the *composition assumption* that $f + h \circ f \in \mathcal{F}$ plays a central role in our analysis, leading to Eq. (3). In the remainder of this paper, we investigate function classes that satisfy this assumption.

Analysis of the smooth CE on the unknown D: While the left-hand side represents the smooth CE on the *training dataset*, our natural objective is to analyze the *population* smooth CE ($\mathrm{smCE}(f, \mathcal{D})$) over the unknown $\mathcal{D}$. The following theorem shows that $\mathrm{smCE}(f, \mathcal{D})$ can be upper bounded in terms of the optimization error, the regularization parameter λ, and the Rademacher complexity.

Theorem 4. *Let the empirical and expected Rademacher complexities of a function class $\mathcal{F}$ be defined by $\hat{\mathfrak{R}}_S(\mathcal{F}) \coloneqq \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right]$, and $\mathfrak{R}_{\mathcal{D},n}(\mathcal{F}) \coloneqq \mathbb{E}_{S\sim\mathcal{D}^n}\left[\hat{\mathfrak{R}}_S(\mathcal{F})\right]$, respectively, where σ_i are i.i.d. Rademacher random variables taking values in $\{\pm 1\}$. Under the assumptions of Theorem 3, and assuming that $\mathfrak{R}_{\mathcal{D},n}(\mathcal{F})$ is bounded, the following holds with probability at least $1 - \delta$ over the random draw of training data S:*

$$\mathrm{smCE}(f, \mathcal{D}) \leq \sqrt{\lambda + \mathrm{err}_n(f)} + 6\mathfrak{R}_{\mathcal{D},n}(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

This result indicates that when the function class is sufficiently regularized, and both the optimization error and regularization parameter are small, the smooth CE is guaranteed to be small. While Theorem 2 characterizes the smooth CE in terms of population risk minimization, Theorem 4 provides a clear connection between the population smooth CE and L_2-regularized ERM.

However, the composition assumption required by both the analysis of Błasiok et al. [2] and our own may be overly restrictive in practice. To examine this limitation, we focus on RKHS, one of the most widely used function classes in machine learning. In particular, focusing on the *kernel ridge regression* (KRR), we show that, except under certain conditions, the composition assumption does not generally hold in this setting (see Table 1). This suggests that the relationship between smooth CE and ERM depends critically on the structural properties of the underlying model class.

Table 1: Summary of our theoretical results in kernel ridge regression.

Kernel function	Input space	Ex. of application	$f + h \circ f \in \mathcal{F}$ (in Theorem 3)	Corresp. thm.
Laplace	$\mathcal{X} = \mathbb{R}$	Recalibration	✓	Corollary 2
	$\mathcal{X} = \mathbb{R}^d\ (d > 1)$	Binary Classification	–	Theorem 7
Gaussian	$\mathcal{X} = \mathbb{R}$	Recalibration	–	Theorem 7
	$\mathcal{X} = \mathbb{R}^d\ (d > 1)$	Binary Classification	–	Theorem 7

3.2 Controlling Smooth CE under KRR

We now describe kernel ridge regression under the squared loss ℓ_{sq}. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with associated RKHS $\mathcal{H}$ and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, used to learn the predictor f. The feature map of k is $\phi : \mathcal{X} \to \mathcal{H}$, so that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ for any $x, x' \in \mathcal{X}$. We define the hypothesis class as $\mathcal{F} = \mathcal{H} \oplus \mathbb{R} = \{f' + b \mid f' \in \mathcal{H}, b \in \mathbb{R}\}$. The predictor f is then trained by solving the regularized ERM problem in Eq. (2). The solution admits the following form by the representer theorem [28]: $f_n^*(x) = \frac{1}{n} \sum_{i=1}^n \alpha_i k(x, x_i) + b$, where $\{\alpha_i\}_{i=1}^n \subset \mathbb{R}$ and $b \in \mathbb{R}$ is a bias term.

We first focus on the Laplace kernel as a specific choice [34]. One motivation for this choice is its widespread use in calibration metrics such as the maximum mean calibration error (MMCE) [26]. Moreover, the RKHS induced by the Laplace kernel is known to contain Lipschitz continuous functions [1], making it a natural candidate for function classes that satisfy the composition assumption required by Theorem 4. In addition, the Laplace kernel is *universal* [34], meaning it has sufficient expressive power to approximate a broad class of functions. We also note that our result extends to general universal kernels, as formalized in Theorem 7 at the end of this section.

Therefore, this section aims to theoretically investigate whether the Laplace kernel satisfies the composition assumption in Theorem 4, and whether fitting KRR with the Laplace kernel leads to a small population smooth CE under the setup described above. Our analysis reveals a significant distinction between the following two cases: (i) the *univariate input space* $\mathcal{X} = \mathbb{R}$, and (ii) the *multivariate input space* $\mathcal{X} = \mathbb{R}^d$ with $d > 1$. We present the results for each case in turn. A summary of this section's findings is provided in Table 1.

When $\mathcal{X} = \mathbb{R}$ (Case (i)): Given $x, x' \in \mathcal{X}$, we define the Laplace kernel as $k(x, x') := \exp(-|x - x'|)$. To simplify the notation, we omit the bandwidth. The following theorem shows that, in the univariate case, the Laplace kernel satisfies the composition assumption required by Theorem 4.

Theorem 5. *Assume $\mathcal{X} = \mathbb{R}$ and let k be the Laplace kernel. Then, for any $f \in \mathcal{F} = \mathcal{H} \oplus \mathbb{R}$ and $h \in \mathrm{Lip}_1$, it holds that $f + h \circ f \in \mathcal{F}$.*

This result allows us to derive the following corollary, which states that minimizing the objective in Eq. (2) also ensures a small value of the population smooth CE.

Corollary 2. *Assume that $\mathcal{X} = \mathbb{R}$ and k is the Laplace kernel. Suppose there exist constants $\Lambda > 0$ and $\alpha > 0$ such that $\sup_{x,x' \in \mathcal{X}} k(x, x') \leq \Lambda$, $\|f'\|_{\mathcal{H}} \leq \alpha$ for any $f' \in \mathcal{H}$ and $|b| \leq \alpha\Lambda + 1$. Then, with probability at least $1 - \delta$ over the training dataset S, for any $f \in \mathcal{F}$, we have:*

$$\mathrm{smCE}(f, \mathcal{D}) \leq \sqrt{\lambda + \mathrm{err}_n(f)} + 6\frac{3\alpha\Lambda + 2}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

This theorem illustrates how the regularization coefficient, optimization error, and the complexity of $\mathcal{F}$ affect the smooth CE. Although Błasiok et al. [2] conjectured that those factors might influence the smooth CE, we formally provide the first rigorous relationships among them.

Corollary 2 provides a uniform bound over the entire class $\mathcal{F} = \{f' + b | f' \in \mathcal{H}, b \in \mathbb{R}\}$ and the complexity of this class seems to depend on α, not λ. However, the fact is that L_2-regularized ERM constrains the learned function f_n^* to a much smaller subset of $\mathcal{F}$ as follows:

Corollary 3. *Under the same settings as Corollary 2, the ERM solution of KRR lies in $\Omega := \{(f', b) \in \mathcal{H} \oplus \mathbb{R} | \|f'\|_{\mathcal{H}} \leq \lambda^{-1/2}, |b| \leq \Lambda/\lambda^{1/2} + 1\}$. Then, with probability at least $1 - \delta$ over the training dataset S, for any $f \in \Omega$, we have:*

$$\mathrm{smCE}(f, \mathcal{D}) \leq \sqrt{\lambda + \mathrm{err}_n(f)} + 6\frac{3\Lambda + 2}{\sqrt{n\lambda}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

This bound now explicitly captures the bias-variance trade-off: The first term ($\sqrt{\lambda}$), which relates to the training smooth CE (bias), increases with λ. The second term ($\mathcal{O}(1/\sqrt{\lambda n})$), which is the Rademacher complexity term (variance), decreases as

λ increases. This clearly demonstrates how a larger λ reduces the complexity of the solution space, thereby tightening the generalization bound, at the cost of increasing the training smooth CE. In Section 5, we evaluate this trade-off numerically.

In fact, when using the Gaussian kernel, $f + h \circ f \notin \mathcal{F}$ holds, which means that the composition assumption in Theorem 4 does not hold. This is because the RKHS induced by the Gaussian kernel does not include the Lipshitz functions [2], therefore $h \circ f \notin \mathcal{F}$.

A practical application where $\mathcal{X} = \mathbb{R}$ naturally arises is *recalibration* [18, 37, 10], which aims to adjust poorly calibrated model predictions. See Section 3.3 for the details.

When $\mathcal{X} = \mathbb{R}^d$ (Case (ii)): The following theorem demonstrates that in higher dimensions, the RKHS of the Laplace kernel is not closed under post-processing. As a result, Theorem 4 does not apply, rendering a direct application of Theorem 4 infeasible.

Theorem 6. *Let $\mathcal{X} = \mathbb{R}^d$ for $d > 1$ and k be the Laplace kernel. Then for any $f \in \mathcal{F} = \mathcal{F} \oplus \mathbb{R}$ and any $h \in \mathrm{Lip}_1$, it holds that $f + h \circ f \notin \mathcal{F}$.*

Similarly to the univariate case, the Gaussian kernel is not closed under post-processing in the multivariate case either, meaning that $f + h \circ f \notin \mathcal{F}$ still holds.

Nevertheless, by leveraging the approximation property of universal kernels in place of the assumption in Theorem 4, we can still derive a valid (albeit somewhat looser) upper bound on $\mathrm{smCE}(f_n^*, \mathcal{D})$.

Theorem 7. *Let k be a universal kernel with associated RKHS $\mathcal{H}$. Suppose there exist constant Λ such that $\sup_{x,x' \in \mathcal{X}} k(x, x') \leq \Lambda$. Then, with probability at least $1 - \delta$ over the draw of the training dataset, the ERM solution of KRR (f_n^*) satisfies*

$$\mathrm{smCE}(f_n^*, \mathcal{D}) \leq \sqrt{\lambda + 4\frac{3\Lambda + 2}{\sqrt{\lambda n}} + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}}.$$

This theorem holds only for f_n^*, whereas Corollary 2 applies to any $f \in \mathcal{F}$. In terms of the upper bound, this theorem yields a rate of $\mathcal{O}(1/n^{1/4})$, while Corollaries 2 and 3 achieve $\mathcal{O}(1/n^{1/2})$. Owing to the relaxed assumptions, however, it applies to a broader class of input spaces $\mathcal{X}$, covering both $\mathcal{X} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^d$ $(d > 1)$. Moreover, this result applies to universal kernels such as the Gaussian kernel. Finally, similar to Corollary 3, there is a bias–variance trade-off concerning λ.

3.3 Dual smooth CE and Regularized ERM under Kernel Logistic Regression

In this section, following the analytical framework developed in Sections 3.1 and 3.2, we theoretically analyze the relationship between *dual smooth CE* and regularized ERM with the *cross-entropy loss*, specifically focusing on the setting of *kernel logistic regression* (KLR).

Regularized ERM with ℓ^ψ: Let $\mathcal{G}$ denote a class of logit functions $g : \mathcal{X} \to \mathbb{R}$, endowed with a norm $\|\cdot\|_{\mathcal{G}}$. Given a training dataset $S = (X_i, Y_i)_{i=1}^n$, we define the regularized empirical risk as

$$L_n(g) \coloneqq \frac{1}{n} \sum_{i=1}^n \ell^\psi(g(X_i), Y_i) + \lambda \|g\|_{\mathcal{G}}^2.$$

We define $g_n^* \coloneqq \mathrm{argmin}_{g \in \mathcal{G}} L_n(g)$ and $\mathrm{err}_n(g) \coloneqq L_n(g) - L_n(g_n^*)$. We now present a generalization bound for the dual smooth CE, analogous to Theorem 4 in the squared loss setting.

Corollary 4. *Assume that $\|g\|_{\mathcal{G}} \leq G$ for any g. Moreover assume that for any $g \in \mathcal{G}$ and $h \in \mathrm{Lip}_{1/4}(\mathbb{R} \times [-1, 1])$, it holds that $g + h \circ g \in \mathcal{G}$, and the Rademacher complexity $\mathfrak{R}_{\mathcal{D},n}(\mathcal{G})$ is bounded. Then with probability at least $1 - \delta$ over a draw of S, for any $g \in \mathcal{G}$, we have*

$$\mathrm{smCE}^{(\psi,1/4)}(g, \mathcal{D}) \leq \sqrt{\lambda G^2 + \frac{1}{2}\mathrm{err}_n(g)} + 6\mathfrak{R}_{\mathcal{D},n}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

Similar to Theorem 4, the dual smooth calibration error is upper bounded by the optimization error, the regularization coefficient, and the complexity of the hypothesis class. The key difference is that the bound explicitly involves the norm of the logit function g, denoted by G. Moreover, by Eq. (1), the dual smooth CE upper bounds the standard smooth CE; hence, this corollary also provides a theoretical guarantee for the smooth CE.

Kernel logistic regression (KLR): Here, we analyze how KLR controls the dual smooth CE. Following the setup in Section 3.2, we adopt the same kernel notation and consider the hypothesis class $\mathcal{G} = \mathcal{H} \oplus \mathbb{R} = \{g' + b \mid g' \in \mathcal{H}, b \in \mathbb{R}\}$. As with the squared loss setting in Section 3.2, we focus on the Laplace kernel and examine two separate cases: $\mathcal{X} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^d$. We begin with the univariate case, $\mathcal{X} = \mathbb{R}$:

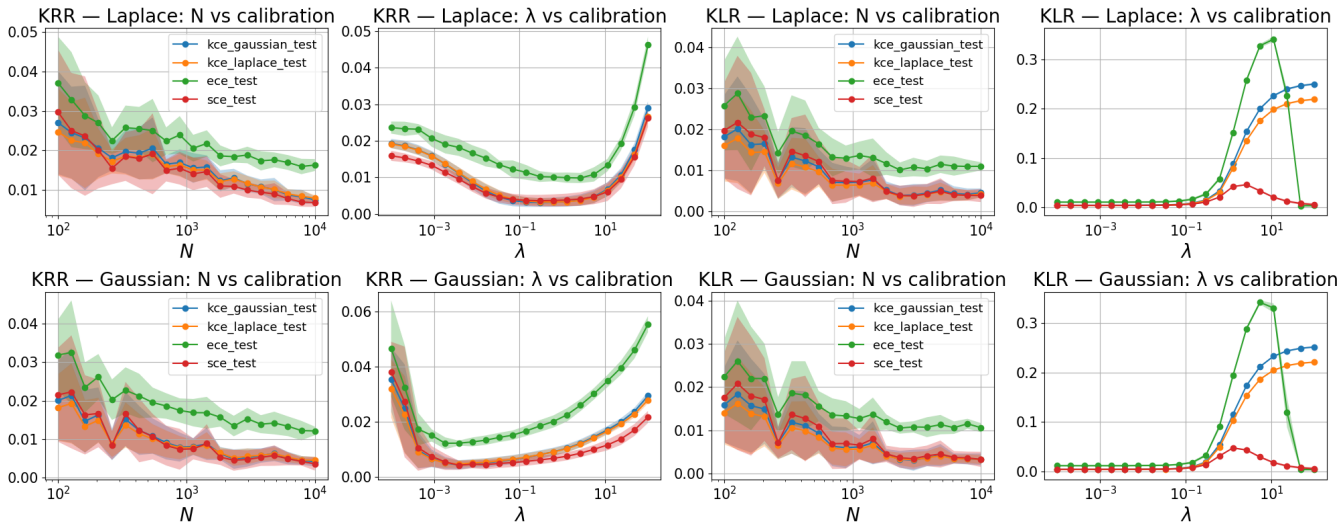

Figure 1: Experimental Results: Effect of Sample Size and Regularization on Calibration Metrics on the toy dataset.

Corollary 5. *Assume $\mathcal{X} = \mathbb{R}$ and k is the Laplace kernel. Suppose there exist constants Λ and G such that $\sup_{x,x' \in \mathcal{X}} k(x,x') \leq \Lambda$, $\|g'\|_{\mathcal{H}} \leq G$ for all $g' \in \mathcal{H}$ and and $|b| \leq G\Lambda + 1$. Then, for any $g \in \mathcal{G}$, we have*

$$\mathrm{smCE}^{(\psi,1/4)}(g, \mathcal{D}) \leq \sqrt{\lambda G + \frac{1}{2}\mathrm{err}_n(g)} + 6\frac{3G\Lambda + 2}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

This result parallels Corollary 2 for the squared loss, where the optimal solution f_n^* is available in closed form. In contrast, under the cross-entropy loss, no such closed-form solution exists for g_n^*. Nevertheless, due to the strong convexity of the objective function $L_n(g)$, the solution g_n^* can still be computed efficiently, and the optimization error $\mathrm{err}_n(g)$ remains controllable. Assuming such an ERM solution is available, we obtain the following result:

Corollary 6. *Under the same settings as Corollary 5, the ERM solution of KLR lies in $\Omega' := \{(g', b) \in \mathcal{H} \oplus \mathbb{R} | \|g'\|_{\mathcal{H}} \leq \lambda^{-1/2}, |b| \leq \Lambda/\lambda^{1/2} + 1\}$. Then, with probability at least $1 - \delta$ over the training dataset S, for any $g \in \Omega'$, we have:*

$$\mathrm{smCE}^{(\psi,1/4)}(g, \mathcal{D}) \leq \sqrt{\lambda + \mathrm{err}_n(g)} + 6\frac{3\Lambda + 2}{\sqrt{n\lambda}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

Similar to Corollary 3, this bound now explicitly captures the bias-variance trade-off concerning λ.

When $\mathcal{X} = \mathbb{R}^d$, an analogous result to Theorem 7 holds for any universal kernel. This means that even without the closure property, the dual smooth CE is controllable, albeit with the slower $\mathcal{O}(n^{-1/4})$ convergence rate. The complete statement and proof are presented in Appendix A.

Application to recalibration: As discussed in Section 3.2, the case of $\mathcal{X} = \mathbb{R}$ is particularly relevant for recalibration. In binary classification, recalibration typically involves training a parametric corrector that takes the model's predicted probability for the top label, represented as a scalar, as input and outputs a calibrated probability estimate. A classical example of this approach is Platt scaling, which fits a logistic regression model to the output scores of a base classifier [29].

We now outline the recalibration procedure. Let g be a pre-trained logit function that need not belong to an RKHS. If the predictor $f = \sigma(g)$ is miscalibrated, we apply a post-hoc correction using a function $\eta : \mathbb{R} \to [0, 1]$, trained on a separate recalibration dataset. Given g and $\mathcal{S}_{\mathrm{re}} = \{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^{n_{\mathrm{re}}}$, we form $\mathcal{S}'_{\mathrm{re}} = \{(g(\tilde{X}_i), \tilde{Y}_i)\}_{i=1}^{n_{\mathrm{re}}}$ and train η on $\mathcal{S}'_{\mathrm{re}}$. The recalibrated predictor is $\eta \circ g$, with predicted probability $\sigma(\eta(g(x)))$. Since $\{g(\tilde{X}_i)\}$ is one-dimensional and independent, assuming η is KLR, Corollary 5 applies directly. To our knowledge, this gives the first theoretical guarantee for the recalibration performance of KLR. We empirically evaluate kernel-based recalibration in Section 5.2.

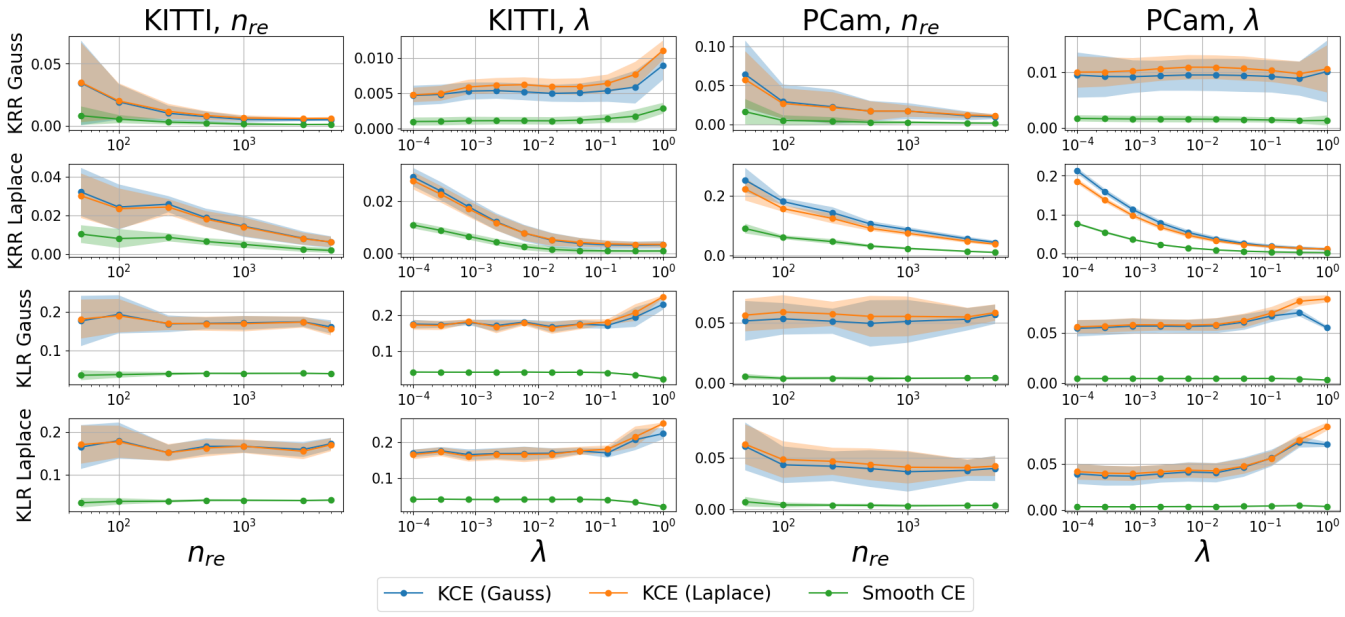

Figure 2: Effect of sample size and regularization on calibration metrics on the KITTI and PCam datasets (1-layer neural networks).

4 Related Work

Kernel-based calibration metrics: Our work connects closely with kernel-based metrics, particularly the Maximum Mean Calibration Error (MMCE) [26]. MMCE is defined as $\mathrm{kCE}(f, \mathcal{D}, \mathcal{H}_1) \coloneqq \sup_{\phi \in \mathcal{H}_1} \mathbb{E}\left[\phi(f(X)) \cdot (Y - f(X))\right]$, where the supremum is taken over functions in an RKHS, instead of the Lipschitz functions used in smooth CE. The choice of kernel is critical. For the Laplace kernel, the RKHS contains Lipschitz functions, establishing a direct inequality: $\frac{1}{3}\mathrm{smCE}(f, \mathcal{D}) \leq \mathrm{kCE}(f, \mathcal{D}, k)$. This connection reinforces our theoretical finding in Theorem 5 that the Laplace kernel's RKHS is well-suited for calibration analysis. While Kumar et al. [26] proposed algorithms that explicitly regularize with an MMCE term, our work is fundamentally different: we show that standard L_2-regularization—a much simpler and more common technique–is sufficient to control smCE.

Theoretical guarantees for calibration: Our primary contribution is providing theoretical guarantees for calibration under a standard training procedure. The most related work is that of Błasiok et al. [2], whose population-level analysis served as the starting point for our work, as detailed in Section 2.2. While they analyze the finite-sample estimation of smCE, we provide the first explicit algorithmic conditions under which the training smCE itself is guaranteed to be small (Theorem 3). Other works have focused on the generalization properties of the binning-based ECE [11], but did not identify how to modify the training algorithm to ensure a small training calibration error. Our work fills this gap by proving that canonical L_2-regularized ERM provides precisely such a guarantee.

Recalibration methods: Finally, our results provide new theoretical insights into recalibration. This task naturally reduces the input space to a single dimension, making our one-dimensional analysis (e.g., Corollary 2) directly applicable. Previous theoretical guarantees for recalibration have largely focused on non-parametric methods [20, 19]. In contrast, our work provides the first theoretical justification for using parametric models like kernel logistic regression for recalibration, explaining the empirical success of such methods in real-world settings.

5 Experiment

In this section, we conduct experiments to empirically validate our theoretical findings. We focus on two key settings that directly correspond to our main theoretical results: (i) a toy dataset where the input space is univariate ($\mathcal{X} = \mathbb{R}$), designed to test our tightest bounds (e.g., Corollaries 3 and 6), and (ii) a practical recalibration task on real-world data, which also reduces to a one-dimensional problem. Experiments on higher-dimensional classification tasks ($\mathcal{X} = \mathbb{R}^d$) are presented in Appendix D, with a comprehensive summary of all settings and implementation details in Appendix C.

5.1 Toy Data Experiments ($\mathcal{X} = \mathbb{R}$)

To numerically validate our theoretical findings, we conduct experiments on a one-dimensional synthetic dataset where the ground-truth conditional probability is analytically tractable. Our theory makes two key predictions that we test here: (i) the existence of a bias-variance trade-off governed by the regularization parameter λ (Corollaries 3 and 6), and (ii) the convergence of smCE as the sample size n increases (Theorem 7). We designed two experiments to investigate these phenomena for both KRR and KLR. For comparison, we also plot the binning-based ECE and MMCE alongside smCE in Figure 1.

Effect of λ: First, we fix the sample size n and vary λ. For KRR, the results (Figure 1, second panel) show a distinct U-shaped curve. This is an empirical manifestation of the bias-variance trade-off described in Corollary 3: for small λ, the model overfits (high variance), while for large λ, it underfits (high bias). An optimal λ balances these two sources of error to achieve the lowest smCE.

For KLR (Figure 1, rightmost panel), we observe an analogous U-shaped curve. Interestingly, as λ becomes extremely large, the predictor collapses to the marginal probability $P(Y)$. While this model is uninformative, it is perfectly calibrated, causing both smCE and ECE to approach zero.

Effect of n: Next, we investigate the effect of the sample size n. For KRR, we set λ according to the rates that minimize our theoretical bounds: $\lambda = O(n^{1/2})$ for the Laplace kernel and $\lambda = O(n^{1/3})$ for the Gaussian kernel. As predicted by our theory, Figure 1 (left panel) shows that smCE consistently decreases as n grows. For KLR, obtaining a closed-form solution for the optimal λ is intractable, so we fix it at a constant value ($\lambda = 0.01$). The results (third panel) show that smCE decreases as n increases, which is consistent with the reduction of the generalization error term in our bounds.

5.2 Recalibration on Real-World Datasets ($\mathcal{X} = \mathbb{R}$)

We now validate our one-dimensional theoretical analysis in a practical setting: post-hoc recalibration as discussed in Section 3. We trained a one-layer neural network on two benchmark datasets, KITTI [12] and PCam [33]. We then used the outputs of this base model to form a one-dimensional recalibration dataset, $\{(f(X_i), Y_i)\}_{i=1}^{n_{\mathrm{re}}}$, on which we applied KRR and KLR as recalibrators. We evaluated the effect of both the recalibration sample size n_{re} and λ.

The empirical results, shown in Figure 2, support our theoretical guarantees by confirming two key predictions. We observe that as the number of recalibration samples n_{re} increases, the smooth CE consistently decreases across all settings. This validates the convergence behavior predicted by our bounds, where the generalization error term diminishes as the sample size grows. Furthermore, the results illustrate the predicted trade-off with respect to the regularization parameter λ. We see that increasing λ beyond an optimal point generally leads to a higher smooth CE, which is consistent with our theoretical bias-variance trade-off where excessive regularization increases the bias term. These findings underscore the practical relevance of our theory, demonstrating that our analysis accurately describes the behavior of kernel-based recalibration in real-world applications.

6 Conclusion and Limitation

This work establishes a foundational link between a standard machine learning principle—L_2-regularized ERM—and model calibration. We demonstrated that this canonical training procedure directly controls the smCE for binary classification. Our finite-sample bounds, instantiated for kernel methods, formally characterize the interplay between regularization, optimization, and model complexity. Our analysis, as a first step, highlights important directions for future research. A key challenge is to improve the convergence rates of our bounds for the more general function classes that do not satisfy the strict closure assumption (as in Theorem 7). Furthermore, extending our framework from the binary to the multiclass setting—leveraging recent developments in multiclass calibration metrics [17]—is a critical next step.

Acknowledgments

MF was supported by KAKENHI Grant Number 25K21286, Japan. FF was supported by JSPS KAKENHI Grant Number JP23K16948, Japan. FF was supported by JST, PRESTO Grant Number JPMJPR22C8, Japan.

References

[1] J. Błasiok, P. Gopalan, L. Hu, and P. Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, page 1727–1740, 2023.

[2] J. Błasiok, P. Gopalan, L. Hu, and P. Nakkiran. When does optimizing a proper loss yield calibration? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[3] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, Adam Tauman Kalai, and Preetum Nakkiran. Loss minimization yields multicalibration for large neural networks. *arXiv preprint arXiv:2304.09424*, 2023.

[4] Gerard Bourdaud. An introduction to composition operators in sobolev spaces. *arXiv preprint arXiv:2204.01118*, 2022.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] Simon Buchholz. Kernel interpolation in sobolev spaces is not consistent in low dimensions. In *Conference on Learning Theory*, pages 3410–3440. PMLR, 2022.

[7] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[8] A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. doi: 10.1080/01621459.1982.10477856.

[9] Victor Dheur and Souhaib Ben Taieb. A large-scale study of probabilistic calibration in neural network regression. In *International Conference on Machine Learning*, pages 7813–7836. PMLR, 2023.

[10] Masahiro Fujisawa and Futoshi Futami. PAC-Bayes analysis for recalibration in classification. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=eJzZryJfri.

[11] Futoshi Futami and Masahiro Fujisawa. Information-theoretic generalization analysis for expected calibration error. In *Advances in Neural Information Processing Systems*, volume 37, pages 84246–84297. Curran Associates, Inc., 2024.

[12] A. Geiger. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE Computer Society, 2012.

[13] Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11459–11492. PMLR, 23–29 Jul 2023.

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

[15] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[16] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007.

[17] Parikshit Gopalan, Lunjia Hu, and Guy N Rothblum. On computationally efficient multi-class calibration. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1983–2026. PMLR, 2024.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330, 2017.

[19] C. Gupta and A. Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3942–3952. PMLR, 18–24 Jul 2021.

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

[21] Dutch Hansen, Siddartha Devic, Preetum Nakkiran, and Vatsal Sharan. When is multicalibration post-processing necessary? *arXiv preprint arXiv:2406.06487*, 2024.

[22] Lunjia Hu, Arun Jambulapati, Kevin Tian, and Chutong Yang. Testing calibration in nearly-linear time. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[23] V. Kuleshov and P. S Liang. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*, volume 28, pages 3474–3482, 2015.

[24] Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *International Conference on Machine Learning*, pages 11683–11693. PMLR, 2022.

[25] A. Kumar, P. S Liang, and T. Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32: 3792–3803, 2019.

[26] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2805–2814. PMLR, 10–15 Jul 2018.

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

[28] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018. ISBN 0262039400.

[29] 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.

[30] Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7933–7946. Curran Associates, Inc., 2022.

[31] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2595–2623. PMLR, 25–28 Jun 2019.

[32] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008. ISBN 9780387772424.

[33] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pages 210–218, 2018.

[34] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[35] J. Wenger, H. Kjellström, and R. Triebel. Non-parametric calibration for classification. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 178–190, 2020.

[36] D. Widmann, F. Lindsten, and D. Zachariah. Calibration tests beyond classification. In *International Conference on Learning Representations*, 2021.

[37] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Icml*, volume 1, pages 609–616, 2001.

A Proofs

A.1 Proof of Theorem 3

First, we define the empirical post-processing gap

$$\mathrm{pGap}(f, S) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{sq}}(f(X_i), Y_i) - \inf_{h \in \mathrm{Lip}_{L=1}} \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{sq}}(f(X_i) + h(f(X_i)), Y_i)].$$

We can prove that

$$\mathrm{smCE}(f, S)^2 \leq \mathrm{pGap}(f, S) \leq 2\mathrm{smCE}(f, S).$$

The proof follows directly from Lemma 4.7 in Błasiok et al. [2], with the only modification of replacing the population expectation under $\mathcal{D}$ with the empirical expectation in the last step of their proof.

Accordingly, we upper bound the empirical post-processing gap using the L_2-regularized ERM objective evaluated at f and $f + h \circ f$. For notational simplicity, we denote ℓ_{sq} by ℓ and write $r \circ f \coloneqq f + h \circ f$.

Then we have

$$
\begin{aligned}
&\mathrm{pGap}(f, S) \\
&= \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) - \inf_{h \in \mathrm{Lip}_{L=1}} \frac{1}{n}\sum_{i=1}^{n} \ell(r \circ f(X_i), Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda\|f\|_{\mathcal{F}}^2 - \lambda\|f\|_{\mathcal{F}}^2 - \inf_{h \in \mathrm{Lip}_{L=1}} \frac{1}{n}\sum_{i=1}^{n} \ell(r \circ f(X_i), Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda\|f\|_{\mathcal{F}}^2 - \lambda\|f\|_{\mathcal{F}}^2 - \inf_{h \in \mathrm{Lip}_{L=1}} \left(\frac{1}{n}\sum_{i=1}^{n} \ell(r \circ f(X_i), Y_i) + \lambda\|r \circ f\|_{\mathcal{F}}^2 - \lambda\|r \circ f\|_{\mathcal{F}}^2 \right) \\
&= L_n(f) - \inf_{h \in \mathrm{Lip}_{L=1}} \left(\frac{1}{n}\sum_{i=1}^{n} \ell(r \circ f(X_i), Y_i) + \lambda\|r \circ f\|_{\mathcal{F}}^2 + \lambda\|f\|_{\mathcal{F}}^2 - \lambda\|r \circ f\|_{\mathcal{F}}^2 \right) \\
&\leq L_n(f) - \inf_{h \in \mathrm{Lip}_{L=1}} \left(\frac{1}{n}\sum_{i=1}^{n} \ell(r \circ f(X_i), Y_i) + \lambda\|r \circ f\|_{\mathcal{F}}^2 \right) - \inf_{h \in \mathrm{Lip}_{L=1}} \left(\lambda\|f\|_{\mathcal{F}}^2 - \lambda\|r \circ f\|_{\mathcal{F}}^2 \right) \\
&\leq L_n(f_n^*) + \mathrm{err}_n(f) - \inf_{h \in \mathrm{Lip}_{L=1}} \left(\frac{1}{n}\sum_{i=1}^{n} \ell(r \circ f(X_i), Y_i) + \lambda\|r \circ f\|_{\mathcal{F}}^2 \right) + \lambda \\
&= L_n(f_n^*) + \mathrm{err}_n(f) - \inf_{h \in \mathrm{Lip}_{L=1}} L_n(r \circ f) + \lambda \\
&\leq \mathrm{err}_n(f) + \lambda,
\end{aligned}
$$

where we used that $L_n(f) = L_n(f_n^*) + \mathrm{err}_n(f)$ and $\|r \circ f\|_{\mathcal{F}}^2 \leq 1$, which is a direct consequence of the assumption on the norm of $\mathcal{F}$ and $\|f\|_{\mathcal{F}} \leq 1$. In the final line, we also used the following relation:

$$
L_n(f_n^*) \leq \inf_{h \in \mathrm{Lip}_{L=1}} L_n(r \circ f)
$$

since f_n^* minimizes $L_n(f)$ over $\mathcal{F}$ and we assume $f + h \circ f \in \mathcal{F}$.

A.2 Proof of Theorem 4

Proof. We follow the proof technique of Theorem 9.5 in Błasiok et al. [1];

$$
\begin{aligned}
&\mathrm{smCE}(f, \mathcal{D}) - \mathrm{smCE}(f, S) \\
&= \sup_{\eta} \mathbb{E}\eta(f(X)) \cdot (Y - f(X)) - \sup_{\eta}{}' \frac{1}{n}\sum_{i} \eta'(f(X_i)) \cdot (Y_i - f(X_i)) \\
&= \sup_{\eta} \mathbb{E}\left[\frac{1}{2}(Y - (f - \eta(f)))^2 - \frac{1}{2}(Y - f)^2 - \frac{1}{2}\eta(f)^2 \right] \\
&- \sup_{\eta'} \frac{1}{n}\sum_{i} \left[\frac{1}{2}(Y_i - (f_i - \eta'(f_i)))^2 - \frac{1}{2}(Y - f_i)^2 - \frac{1}{2}\eta'(f_i)^2 \right] \\
&\leq \sup_{\eta} \Bigg(\mathbb{E}\left[\frac{1}{2}(Y - (f - \eta(f)))^2 - \frac{1}{2}(Y - f)^2 - \frac{1}{2}\eta(f)^2 \right] \\
&- \frac{1}{n}\sum_{i} \left[\frac{1}{2}(Y_i - (f_i - \eta(f_i)))^2 - \frac{1}{2}(Y - f_i)^2 - \frac{1}{2}\eta(f_i)^2 \right] \Bigg)
\end{aligned}
$$

and then we take the supremum for f,

$$
\begin{aligned}
&\sup_{f \in \mathcal{F}} \mathrm{smCE}(f, \mathcal{D}) - \mathrm{smCE}(f, S) \\
&\leq \sup_{f \in \mathcal{F}} \sup_{\eta} \Bigg(\mathbb{E}\left[\frac{1}{2}(Y - (f - \eta(f)))^2 - \frac{1}{2}(Y - f)^2 - \frac{1}{2}\eta(f)^2 \right] \\
&- \frac{1}{n}\sum_{i} \left[\frac{1}{2}(Y_i - (f_i - \eta(f_i)))^2 - \frac{1}{2}(Y - f_i)^2 - \frac{1}{2}\eta(f_i)^2 \right] \Bigg).
\end{aligned}
$$

By setting

$$\Phi(S) = \sup_{f\in\mathcal{F}} \sup_{\eta} \mathbb{E}\left[\frac{1}{2}(Y-(f-\eta(f)))^2 - \frac{1}{2}(Y-f)^2 - \frac{1}{2}\eta(f)^2\right] - \frac{1}{n}\sum_i \left[\frac{1}{2}(Y_i-(f_i-\eta(f_i)))^2 - \frac{1}{2}(Y-f_i)^2 - \frac{1}{2}\eta(f_i)^2\right],$$

From the proof of Theorem 3.3 in Mohri et al. [28], we have, with probability at least $1-\delta$,

$$\Phi(S) \le \mathbb{E}_S[\phi(S)] + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

This can be shown using McDiarmid's inequality, following the same argument as in the proof of Theorem 3.3 in Mohri et al. [28] since

$$\eta(f(X))\cdot(Y-f(X)) = \frac{1}{2}(Y_i-(f_i-\eta(f_i)))^2 - \frac{1}{2}(Y-f_i)^2 - \frac{1}{2}\eta(f_i)^2$$

and the constants for the bounded differences in McDiarmid's inequality are equal to 1, and thus the result is identical to that appearing in Theorem 3.3 in Mohri et al. [28].

We then set $\omega(f,\eta,Y,X) = \frac{1}{2}(Y-(f-\eta(f)))^2 - \frac{1}{2}(Y-f)^2 - \frac{1}{2}\eta(f)^2$, and by the standard symmetrization argument, we have

$$\mathbb{E}_S[\phi(S)] \le 2\mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i \omega(f,\eta,Y_i,X_i)\right].$$

Then by the property of the supremum,

$$\begin{aligned}
&\mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i \omega(f,\eta,Y_i,X_i)\right] \\
&\le \frac{1}{2}\mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i (Y_i-(f_i-\eta(f_i)))^2\right] + \frac{1}{2}\mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i (Y_i-f_i)^2\right] \\
&+ \frac{1}{2}\mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i \eta(f_i)^2\right].
\end{aligned}$$

Then, from Propositions 11.2 and 11.2 in Mohri et al. [28],

$$\frac{1}{2}\mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i (Y_i-(f_i-\eta(f_i)))^2\right] \le \mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i (f_i-\eta(f_i))\right] \le \mathfrak{R}_{\mathcal{D},n}(\mathcal{F})$$

and

$$\frac{1}{2}\mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i (Y_i-f_i)^2\right] \le \mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i f_i\right] \le \mathfrak{R}_{\mathcal{D},n}(\mathcal{F})$$

and

$$\frac{1}{2}\mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i \eta(f_i)^2\right] \le \mathbb{E}_{\sigma,S}\left[\sup_f \sup_\eta \frac{1}{n}\sum_{i=1}^n \sigma_i \eta(f_i)\right] \le \mathfrak{R}_{\mathcal{D},n}(\mathcal{F})$$

holds. In conclusion, we have

$$\mathbb{E}_S[\phi(S)] \le 6\mathfrak{R}_{\mathcal{D},n}(\mathcal{F})$$

where we used the composition assumption that $f+\eta\circ f\in\mathcal{F}$ in the last inequality. This concludes the proof. □

A.3 Proof of Theorem 5

Proof. From Proposition 7 in Rakhlin and Zhai [31], the RKHS associated with the Laplace kernel on $\mathbb{R}$ corresponds to the Sobolev space $H^1 = W^{2,1}$ (see also Buchholz [6]). Then, by Theorem 6 in Bourdaud [4], the composition of a Lipschitz function with a function in H^1 remains in H^1. To apply Theorem 6, the Lipschitz function r must satisfy $r(0)=0$. If $r(0)\neq 0$, define $\tilde{r}(x) = r(x) - r(0)$, so that $\tilde{r}\circ f \in H^1$. Since constant functions are included in the Sobolev space, we have $r\circ f(x) = \tilde{r}\circ f(x) + r(0) \in H^1$.

Similarly, Theorem 7 in Bourdaud [4] shows that the composition of $k\in H^1$ and $f\in H^1$ yields $k\circ f\in H^1$. □

A.4 Proof of Corollary 2 and 3

Proof. First assume that $|b| \leq \alpha\Lambda + 1$, which will be shown later. Then the corresponding Rademacher complexity is

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{(f,b)\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(f(x_i)+b)\right] \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\|f\|_\mathcal{H}\leq\alpha,\,|b|\leq\alpha\Lambda+1}\sum_{i=1}^{n}\sigma_i f(x_i) + \sum_{i=1}^{n}\sigma_i b\right] \\
&\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\|f\|_\mathcal{H}\leq\alpha}\sum_{i=1}^{n}\sigma_i f(x_i)\right] + \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{|b|\leq\alpha\Lambda+1}\sum_{i=1}^{n}\sigma_i b\right]. \qquad (4)
\end{aligned}$$

On the basis of the above, the first term of Eq. (4) can be bounded by $\alpha\Lambda/\sqrt{n}$ from the standard derivation of the Rademacher complexity for the kernel functions, see Mohri et al. [28] for the derivation. The second term is that the supremeum with respect to b is that if $\sum\sigma_i$ is positive, then $b = \alpha\Lambda + 1$, if $\sum\sigma_i$ is negative, then $b = -(\alpha\Lambda + 1)$.

Thus, from the Massart's lemma, where the hypothesis set is

$$\{(\alpha\Lambda + 1),\ -(\alpha\Lambda + 1)\} \subset \mathbb{R}^1$$

Thus, by setting $A \coloneqq \{(\Lambda + 1),\ -(\Lambda + 1)\} \subset \mathbb{R}$ and this results in

$$\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{|b|\leq\Lambda+1}\sum_{i=1}^{n}\sigma_i b\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{z\in A}\sum_{i=1}^{n}\sigma_i z_i\right] \leq \frac{\sqrt{2\log 2}(\Lambda+1)}{\sqrt{n}} \leq \frac{2(\alpha\Lambda+1)}{\sqrt{n}}$$

In conclusion, we have

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leq \frac{3\alpha\Lambda + 2}{\sqrt{n}}. \qquad (5)$$

This concludes the proof of Corollary 2.

Next, we show that any solution f_λ to the regularized ERM problem must have a norm bounded by λ. Consider the objective function $J(f) = L_n(f) + \lambda\|f\|_\mathcal{H}^2$. By definition, the optimal solution $f_\lambda = f'_\lambda + b$ with $f'_\lambda \in \mathcal{H}$ and $b \in \mathbb{R}$ must satisfy:

$$L_n(f_\lambda) + \lambda\|f_\lambda\|_\mathcal{H}^2 = \inf_{f\in\mathcal{H}} L_n(f) + \lambda\|f\|_\mathcal{H}^2$$

The infimum must be less than or equal to the objective value for any function, including the zero function $f = 0$. Therefore:

$$\inf_{f\in\mathcal{H}} L_n(f) + \lambda\|f\|_\mathcal{H}^2 \leq L_n(0) + \lambda\|0\|_\mathcal{H}^2 = L_n(0).$$

Since the loss is non-negative ($L_n(f_\lambda) \geq 0$), we have:

$$\lambda\|f_\lambda\|_\mathcal{H}^2 \leq L_n(f_\lambda) + \lambda\|f_\lambda\|_\mathcal{H}^2 \leq L_n(0)$$

This yields the desired relationship: $\|f_\lambda\|_\mathcal{H} \leq \sqrt{L_n(0)/\lambda}$. Thus, the norm of the optimal solution is of the order $\mathcal{O}(1/\sqrt{\lambda})$. Note that $L_n(0) \leq 1$.

Based on this, we plug this λ-dependent norm bound into the standard formula for the Rademacher complexity of a function class in an RKHS. The complexity for a class of functions $\{f + b \mid \|f\|_\mathcal{H} \leq \alpha, |b| \leq \beta\}$ is bounded by the sum of the complexities for the kernel part and the bias part. And α corresponds to $\sqrt{L_n(0)/\lambda}$. As for β, let us recall the definition of the sqaured loss

$$\begin{aligned}
L_n(f) &= \sum_i \frac{1}{n}(y_i - (\sum_j a_j k(x_i, x_j) + b))^2 + \lambda\|f\|_\mathcal{F}^2 \\
&= \sum_{i:y_i=1} \frac{1}{n}(1 - (\sum_j a_j k(x_i, x_j) + b))^2 + \lambda\|f\|_\mathcal{F}^2 + \sum_{i:y_i=0} \frac{1}{n}(0 - (\sum_j a_j k(x_i, x_j) + b))^2 + \lambda\|f\|_\mathcal{F}^2
\end{aligned}$$

From the above relation, we can see that $|b| \leq \alpha\Lambda + 1$ is the region where the optimal solution exists.

When using ERM solution, by substituting $\alpha = \sqrt{L_n(0)/\lambda}$ into the estimate of the Rademacher complexity in Eq. (5), we obtain the result. □

A.5 Proof of Theorem 6

Proof. From Proposition 7 in Rakhlin and Zhai [31], the RKHS associated with the Laplace kernel on $\mathbb{R}^d$ is the Sobolev space $H^s = W^{2,s}$ with $s = (d+1)/2$ (see also Buchholz [6]). According to Theorem 7 in Bourdaud [4], for given functions f and r, we have $r \circ f \in H^s$ if and only if $r \in H^s$. However, here we take r from a Lipschitz function class, so r is not generally in H^s, and thus the result follows.

Similarly, $k \in \mathcal{K}_1$ implies $k \in H^1$ but not $k \in H^s$. Therefore, we obtain the result. □

A.6 Proof of Theorem 7

Proof. We use the approximation theory given in Corollary 5.29 of Steinwart and Christmann [32], which states that for a continuous Nemitski loss function ℓ, the Bayes risk and the minimum achievable risk within the RKHS are equivalent. According to Definition 2.16 in Steinwart and Christmann [32], the squared loss with L_2 regularization is a Nemitski loss function. Then, Corollary 5.29 implies that

$$\inf_f L(f) = \inf_{f \in \mathcal{F}} L(f),$$

where the infimum on the left-hand side is taken over all measurable functions from $\mathcal{X} \to \mathbb{R}$. This equivalence follows from the approximation power of universal kernels; see the proof of Corollary 5.29 in Steinwart and Christmann [32] for details.

With this in mind, and following the argument in the proof of Claim 5.1 in Błasiok et al. [2], we upper bound the post-processing gap as follows. By the definition of the infimum,

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}\ell_{\mathrm{sq}}(f(X), Y) + \lambda \|f\|_{\mathcal{F}}^2.$$

Consider the solution f^* and an arbitrary $h \in \mathrm{Lip}_{L=1}$. Note that $f^* + h \circ f^*$ is a measurable function. For notational simplicity, we denote $r \circ f^* \coloneqq f^* + h \circ f^*$.

$$\begin{aligned}
&\mathbb{E}\ell_{\mathrm{sq}}(f_n^*(X), Y) - \mathbb{E}\ell_{\mathrm{sq}}(r \circ f_n^*(X), Y) \\
&= \mathbb{E}\ell_{\mathrm{sq}}(f_n^*(X), Y)) + \lambda \|f\|_{\mathcal{F}}^2 - \lambda \|f\|_{\mathcal{F}}^2 - \mathbb{E}\ell_{\mathrm{sq}}(r \circ f_n^*(X), Y) \\
&\leq L(f_n^*) - L(f^*) + L(f^*) - L(r \circ f_n^*) + \lambda \\
&\leq L(f^*) + \mathrm{err}_{\mathrm{ex}}(n) - L(r \circ f_n^*) + \lambda \\
&\leq \lambda + \mathrm{err}_{\mathrm{ex}}(n)
\end{aligned}$$

In the last inequality, we used the fact that

$$L(f^*) - L(r \circ f_n^*) = \inf_{f \in \mathcal{F}} L(f) - L(r \circ f_n^*) = \inf_f L(f) - L(r \circ f_n^*) \leq 0$$

Since infimum is taken with respect to all measurable functions, therefore $\inf_f L(f) \leq L(r \circ f_n^*)$ holds.

Next we upper bound $\mathrm{err}_{\mathrm{ex}}(n)$. This is well studied in the literature of the generalization analysis and from Proposition 4.1 in Mohri et al. [28], we have

$$L(f_n^*) - \inf_{f \in \mathcal{F}} L(f) \leq \mathrm{err}_{\mathrm{ex}}(n) \leq 2 \sup_{f \in \mathcal{F}} |L(f) - L_n(f)| \leq 2 \left(2\mathfrak{R}_{\mathcal{D},n}(\mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \right).$$

Then we have

$$\mathrm{smCE}(f_n^*, \mathcal{D}) \leq \sqrt{\lambda + 4\mathfrak{R}_{\mathcal{D},n}(\mathcal{F}) + \sqrt{2\frac{\log \frac{2}{\delta}}{n}}}.$$

□

A.7 Proof of Corollary 4

We first upper bound the training dual smooth CE. The proof is almost identical to that of Theorem 3. We first introduce the training dual smooth CE as

$$\mathrm{smCE}^{(\psi,1/4)}(g, S) \coloneqq \sup_{h \in \mathrm{Lip}_{1/4}(\mathbb{R},[-1,1])} \frac{1}{n} \sum_{i=1}^{n} \left[h(g(X_i)) \cdot (Y_i - f(X_i)) \right].$$

We also define the empirical counterpart of the dual post-processing gap as

$$\mathrm{pGap}^{(\psi,1/4)}(g,S) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(g(X_i),Y_i) - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}\frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(f(X_i)+h(g(X_i)),Y_i)]$$

We can prove that

$$2\mathrm{smCE}^{(\psi,1/4)}(g,S)^2 \leq \mathrm{pGap}^{(\psi,1/4)}(g,S) \leq 4\mathrm{smCE}^{(\psi,1/4)}(g,S). \tag{6}$$

The proof of this is exactly the same as that of Lemma 4.7 in Błasiok et al. [2], where we simply replace the expectation by $\mathcal{D}$ with that of the empirical expectation.

To simplify the notation, we express $r\circ g \coloneqq g + h\circ g$. Then we will upper bound the training

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(g(X_i),Y_i) - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}\frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(r\circ g(X_i),Y_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(g(X_i),Y_i) + \lambda\|g\|_{\mathcal{G}}^2 - \lambda\|g\|_{\mathcal{G}}^2 - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}\frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(r\circ g(X_i),Y_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(g(X_i),Y_i) + \lambda\|g\|_{\mathcal{G}}^2 - \lambda\|g\|_{\mathcal{G}}^2\\
&\quad - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}\left(\frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(r\circ g(X_i),Y_i) + \lambda\|r\circ g\|_{\mathcal{G}}^2 - \lambda\|r\circ g\|_{\mathcal{G}}^2\right)\\
&= L_n(g) - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}\left(\frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(r\circ g(X_i),Y_i) + \lambda\|r\circ g\|_{\mathcal{G}}^2 + \lambda\|g\|_{\mathcal{G}}^2 - \lambda\|r\circ g\|_{\mathcal{G}}^2\right)\\
&\leq L_n(g) - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}\left(\frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(r\circ g(X_i),Y_i) + \lambda\|r\circ g\|_{\mathcal{G}}^2\right)\\
&\quad - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}\left(\lambda\|g\|_{\mathcal{G}}^2 - \lambda\|r\circ g\|_{\mathcal{G}}^2\right)\\
&\leq L_n(g_n^*) + \mathrm{err}_n(g) - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}\left(\frac{1}{n}\sum_{i=1}^{n}\ell^{\psi}(r\circ g(X_i),Y_i) + \lambda\|r\circ g\|_{\mathcal{G}}^2\right) + \lambda G^2\\
&= L_n(g_n^*) + \mathrm{err}_n(g) - \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}L_n(r\circ g) + \lambda G^2\\
&\leq \mathrm{err}_n(g) + \lambda G^2
\end{aligned}$$

where we used that $L_n(f) = L_n(f_n^*) + \mathrm{err}_n(f)$ and $\|r\circ g\|_{\mathcal{G}}^2 - \|g\|_{\mathcal{G}}^2 \leq G$. Moreover, we used the relation

$$L_n(g_n^*) \leq \inf_{h\in\mathrm{Lip}_{L=1}(\mathbb{R},[-4,4])}L_n(r\circ g)$$

in the last line. Then, using Eq. (6), we have the upper-bound for the training dual smooth CE.

Next, we study the generalization error for $\mathrm{smCE}^{(\psi,1/4)}(g,\mathcal{D})$

$$
\begin{aligned}
&\mathrm{smCE}^{(\psi,1/4)}(g,\mathcal{D}) - \mathrm{smCE}^{(\psi,1/4)}(g,S) \\
&= \sup_h \mathbb{E}h(g(X)) \cdot (Y - f(X)) - \sup_{h'} \frac{1}{n}\sum_i h'(g(X_i)) \cdot (Y_i - f(X_i)) \\
&= \sup_h \mathbb{E}\left[\frac{1}{2}(Y - (f - h(g(X)))^2 - \frac{1}{2}(Y-f)^2 - \frac{1}{2}h(g(X))^2\right] \\
&- \sup_{h'} \frac{1}{n}\sum_i \left[\frac{1}{2}(Y_i - (f_i - h'(g(X_i)))^2 - \frac{1}{2}(Y-f_i)^2 - \frac{1}{2}h'(g(X_i))^2\right] \\
&\leq \sup_\eta \Big(\mathbb{E}\left[\frac{1}{2}(Y - (f - h(g(X)))^2 - \frac{1}{2}(Y-f)^2 - \frac{1}{2}h(g(X))^2\right] \\
&- \frac{1}{n}\sum_i \left[\frac{1}{2}(Y_i - (f_i - h(g(X_i)))^2 - \frac{1}{2}(Y-f_i)^2 - \frac{1}{2}h(g(X_i))^2\right]\Big)
\end{aligned}
$$

Then we proceed the proof exactly in the same way as Appendix A.2.

A.8 Proof of Corollary 5 and 6

By the assumptions, we can apply Corollary 4 to this setting. All we need is to estimate the Rademacher complexity.

We then define the set of functions obtained by $\sigma(g)$ for any $g \in \mathcal{G}$ as $\mathcal{F}$, and σ is $1/4$ Lipshitz function,

$$
\hat{\mathfrak{R}}_S(\mathcal{F}) \leq \frac{1}{4}\hat{\mathfrak{R}}_S(\mathcal{G})
$$

and $\hat{\mathfrak{R}}_S(\mathcal{G})$ can be bounded exactly in the same way as the proof of Corollary 2 in Appendix A.4.

Next, we show that any solution g_λ to the regularized ERM problem must have a norm bounded by λ. Consider the objective function $J(g) = L_n(g) + \lambda\|g\|_{\mathcal{H}}^2$. This yields the desired relationship: $\|g_\lambda\|_{\mathcal{G}} \leq \sqrt{L_n(0)/\lambda}$.

$\ell_{\mathrm{xent}}(v,y) \coloneqq -y\log v - (1-y)\log(1-v)$.

Based on this, we plug this λ-dependent norm bound into the standard formula for the Rademacher complexity of a function class in an RKHS. The complexity for a class of functions $\{f + b \mid \|f\|_{\mathcal{H}} \leq \alpha, |b| \leq \beta\}$ is bounded by the sum of the complexities for the kernel part and the bias part. And α corresponds to $\sqrt{L_n(0)/\lambda}$. As for β, let us recall the definition of the sqaured loss

$$
L_n(g) = \sum_{i:y_i=1}\frac{1}{n}\log(1 + e^{-(g(x_i)+b)}) + \sum_{i:y_i=0}\frac{1}{n}\log(1 + e^{g(x_i)+b})) + \lambda\|g\|_{\mathcal{G}}^2
$$

From the above relation, we can see that $|b| \leq \alpha\Lambda + 1$ is the region where the optimal solution exists. Note that $L_n(0) \leq \log 2 < 1$.

Here we also present the result that corresponds to Theorem 7:

Theorem 8. *Let k be a universal kernel with associated RKHS $\mathcal{H}$. Let $\mathcal{G} = \mathcal{H} \oplus \mathbb{R} = \{g + b \mid g \in \mathcal{H}, b \in \mathbb{R}\}$. Suppose there exist constant Λ such that $\sup_{x,x' \in \mathcal{X}} k(x,x') \leq \Lambda$. Then, with probability at least $1-\delta$ over the draw of the training dataset, it holds that*

$$
\mathrm{smCE}^{(\psi,1/4)}(g_n^*, \mathcal{D}) \leq \sqrt{\lambda G^2 + \frac{3\Lambda + 2}{\sqrt{\lambda n}} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}}.
$$

The proof of this theorem is almost identical to that of Theorem 7, since the logistic loss is also a continuous Nemitski loss and thus, we can apply the same techniques.

B Additional discussion

B.1 Post processing gap and Calibration metrics

Following Błasiok et al. [2], we introduce the general proper loss function and its relation to the post-processing gap.

A proper loss function ℓ can always be represented using a convex function ϕ as follows:

$$\ell(p, y) = -\phi(p) - \nabla\phi(p) \cdot (y - p),$$

where $\phi : [0, 1] \to \mathbb{R}$ is a convex function and $\nabla\phi(p)$ denotes a subgradient at p. Following Błasiok et al. [2], we assume that ϕ is differentiable.

We define the convex conjugate of the function $\phi(p)$ as follows: for all $s \in \mathbb{R}$,

$$\psi(s) = \sup_{p\in[0,1]} \{s \cdot p - \phi(p)\}.$$

The *dual loss* $\ell^\psi : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is then defined as

$$\ell^\psi(s, y) \coloneqq \psi(s) - s \cdot y.$$

By Fenchel–Young duality, this relationship is inverted as $p = \nabla\psi(s)$, and with these definitions, the proper loss can equivalently be written as

$$\ell(p, y) = \ell^\psi(\nabla\phi(p), y) = \psi(\nabla\phi(p)) - \nabla\phi(p) \cdot y.$$

We remark that the relation $p = \nabla\psi(s)$ can be interpreted as mapping logits to predicted probabilities. For details and proofs, see Błasiok et al. [2].

Błasiok et al. [2] considered modeling the score function s by a function g, and then applying the transformation $p = \nabla\psi(s)$. Thus, they proposed to apply post-processing to g, which leads to the following definition:

Definition 5 (Dual post-processing gap)**.** *Assume that ψ is a differentiable and convex function with derivative $\nabla\psi(t) \in [0, 1]$ for all $t \in \mathbb{R}$, and that ψ is λ-smooth. Given ψ, ℓ^ψ, $g : \mathcal{X} \to \mathbb{R}$, and distribution $\mathcal{D}$, we define the dual post-processing gap as*

$$\mathrm{pGap}^{(\psi,\lambda)}(g, \mathcal{D}) \coloneqq \mathbb{E}[\ell^\psi(g(X), Y)] - \inf_{h\in\mathrm{Lip}_1(\mathbb{R},[-1/\lambda,1/\lambda])} \mathbb{E}[\ell^\psi(g(X) + h(g(X)), Y)].$$

When considering the cross-entropy loss, the dual post-processing gap corresponds to improving the logit function.

Definition 6 (Dual smooth calibration)**.** *Consider the same setting as in the definition of the dual post-processing gap. Given ψ and g, define $f(\cdot) = \nabla\psi(g(\cdot))$. The dual calibration error of g is defined as*

$$\mathrm{smCE}^{(\psi,\lambda)}(g, \mathcal{D}) \coloneqq \sup_{h\in\mathrm{Lip}_{L=\lambda}(\mathbb{R},[-1,1])} \mathbb{E}[\eta(g(X)) \cdot (Y - f(X))].$$

Then, similarly to the relationship between the smooth ECE and the post-processing gap, the following holds: if ψ is a λ-smooth function, then

$$\frac{1}{2}\mathrm{smCE}^{(\psi,\lambda)}(g, \mathcal{D})^2 \le \frac{\lambda}{2}\mathrm{pGap}^{(\psi,\lambda)}(g, \mathcal{D})^2 \le \mathrm{smCE}^{(\psi,\lambda)}(g, \mathcal{D}),$$

and

$$\mathrm{smCE}(f, \mathcal{D}) \le \mathrm{smCE}^{(\psi,\lambda)}(g, \mathcal{D})$$

also holds. Thus, by studying the dual post-processing gap, we can obtain bounds on the smooth calibration error. By considering L_2-regularized objective function $\mathbb{E}[\ell^\psi(g(X), Y)] + \|g\|_G^2$ and its empirical counterpart, we can develop the theory for the general dual smooth CE and ERM in a similar way to the case of the squared and cross-entropy loss.

B.2 Relationships different calibration metrics

Błasiok et al. [1] introduced the ground truth distance for calibration, defined as follows:

Definition 7 (True distance to calibration)**.** *We define the true distance of a predictor f from calibration as*

$$\mathrm{dCE}_\mathcal{D}(f) \coloneqq \inf_{g\in\mathrm{cal}(\mathcal{D})} \mathbb{E}_\mathcal{D}|f(x) - g(x)|,$$

where $\mathrm{cal}(\mathcal{D})$ denotes the set of predictors that are perfectly calibrated with respect to $\mathcal{D}$.

This provides an ideal notion for measuring calibration; see Błasiok et al. [1] for details. They showed that the smooth CE both upper and lower bounds the true distance to calibration:

$$\mathrm{smCE}(f, \mathcal{D}) \le \mathrm{dCE}_\mathcal{D}(f) \le 4\sqrt{2\mathrm{smCE}(f, \mathcal{D})}.$$

On the other hand, the commonly used ECE, defined as

$$\mathrm{ECE}_{\mathcal{D}}(f) \coloneqq \mathbb{E}_{\mathcal{D}}\left[\left|\mathbb{E}_{\mathcal{D}}[y|f(x)] - f(x)\right|\right],$$

is discontinuous, and Błasiok et al. [1] showed that ECE does not lower bound $\mathrm{dCE}_{\mathcal{D}}(f)$ unless continuity of the conditional expectation is assumed.

Błasiok et al. [1] also established the relationship between $\mathrm{dCE}_{\mathcal{D}}(f)$ and the binned ECE. Given a partition $\mathcal{I} = \{I_1, \ldots, I_m\}$ of $[0, 1]$ into intervals, the binned ECE is defined as

$$\mathrm{binnedECE}_{\mathcal{D}}(f, \mathcal{I}) = \sum_{j \in [m]} \left|\mathbb{E}[(f - y)\mathbb{1}(f \in I_j)]\right|.$$

They showed that by adding the bin widths and minimizing over the choice of partition, we obtain the following definition:

$$\mathrm{intCE}(f) \coloneqq \min_{\mathcal{I}} \left(\mathrm{binnedECE}_{\mathcal{D}}(f, \mathcal{I}) + w(\mathcal{I})\right),$$

where

$$w(\mathcal{I}) \coloneqq \sum_{j \in [m]} \left|\mathbb{E}w(I_j)\mathbb{1}(f \in I_j)\right|,$$

and $w(I)$ denotes the width of interval I. Then, the following bound holds (Theorem 6.3 in Błasiok et al. [1]):

$$\mathrm{dCE}_{\mathcal{D}}(f) \leq \mathrm{intCE}(f) \leq 4\sqrt{\mathrm{dCE}_{\mathcal{D}}(f)}.$$

As we have seen, bounding the smooth CE leads to a bound on $\mathrm{dCE}_{\mathcal{D}}(f)$, which in turn bounds $\mathrm{intCE}(f)$, which corresponds to the binned ECE, which is optimized with respect to the partition.

C Details of experimental settings

In this section, we summarize the detail information of our numerical experiments in Section 5. Our experiments were conducted on NVIDIA GPUs with 32GB memory (NVIDIA DGX-1 with Tesla V100 and DGX-2). The source code to reproduce all experiments is available at https://github.com/msfuji0211/erm_calibration.

Table 2: Datasets used in our experiments

Dataset	**Classes**	**Train data** (n_{tr})	**Recalibration data** (n_{re})	**Test data** (n_{te})
KITTI	2	16000	1000	8000
PCam	2	22768	1000	9000

C.1 Toy data experiments ($\mathcal{X} = \mathbb{R}$)

To investigate the behavior of different kernel-based methods under controlled conditions, we first conduct experiments on synthetic two-dimensional binary classification tasks. These toy experiments serve to isolate and visualize model behavior with respect to classification performance and calibration quality, without the confounding factors present in real-world datasets.

Data generation. We generate synthetic data using a simple but structured stochastic process. For each of n samples, a binary label $y \in \{0, 1\}$ is drawn independently from a Bernoulli(0.5) distribution. Given the label, the input feature $x \in \mathbb{R}^2$ is sampled from a Gaussian distribution centered at $\mu_1 = [-1, -1]^T$ for $y = 1$, and at $\mu_0 = [1, 1]^T$ for $y = 0$, with identity covariance $\Sigma = I_2$ in both cases. That is,

$$x \mid y = 1 \sim \mathcal{N}([-1, -1]^T, I), \quad x \mid y = 0 \sim \mathcal{N}([1, 1]^T, I).$$

This construction induces a smooth but nonlinear Bayes decision boundary, suitable for evaluating kernel methods.

Models and kernels. We evaluate two models:

- **KRR**: Kernel Ridge Regression with theoretically motivated $\lambda_n = n^{-1/2}$ for Gaussian kernels and $\lambda_n = n^{-1/3}$ for Laplace kernels.
- **KLR**: Kernel Logistic Regression optimized via gradient descent.

Each model is evaluated using two kernels: the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$ and the Laplace kernel $k(x, x') = \exp(-\|x - x'\|/\sigma)$. For each kernel, the bandwidth σ is selected using the median heuristic on the training data.

Metrics. We assess both accuracy and calibration using the following metrics:

- **Kernel Calibration Error (KCE)**: Evaluated with both Gaussian and Laplace kernels, with σ determined by a heuristic on the predicted confidence vector.
- **Smooth Calibration Error (SCE)**: A continuous variant of calibration error designed for better sample efficiency.
- **Expected Calibration Error (ECE)**: Classical binning-based calibration metric with the number of bins set to $\lfloor n^{1/3} \rfloor$ for n data points, following Futami and Fujisawa [11], Fujisawa and Futami [10].

Experimental protocol. We evaluate performance as a function of training set size, with n_{train} logarithmically spaced from 100 to 10,000. For each setting, experiments are repeated with 10 different random seeds for robustness. We also evaluate sensitivity to the regularization parameter λ by fixing $n_{\text{train}} = 10{,}000$ and varying λ over a logarithmic grid from 10^{-4} to 10^2.

Implementation. All methods are implemented using PyTorch. Gradient descent for KLR is run for up to 1,000 iterations with a step size of 0.01 and stopping tolerance of 10^{-6}. Results are reported on both training and test sets. Each experiment logs the metrics above and saves results in a CSV format for post-hoc statistical analysis.

C.2 Recalibration experiments ($\mathcal{X} = \mathbb{R}$)

We provide the details of the datasets along with the number of training, recalibration, and test samples in Table 2. For the models, we used XGBoost [7], Random Forests [5], and a one-layer neural network (NN) for the KITTI and PCam experiments. All experiments—including the training of XGBoost, Random Forests, and the one-layer NN—were conducted using code adapted from Wenger et al. [35] [2].

Performance evaluation: We evaluated predictive accuracy and binned ECE, using $B = \lfloor n_{\text{re}}^{1/3} \rfloor$, in accordance with the theoretical insights from Futami and Fujisawa [11], Fujisawa and Futami [10]. Additionally, we included two other calibration metrics: KCE and SCE. To train the recalibration functions, we performed 10-fold cross-validation and reported the mean and standard deviation of both performance metrics.

C.3 Real dataset experiments (Binary classification benchmarks; $\mathcal{X} = \mathbb{R}^d$)

We perform binary classification experiments using real-world tabular datasets to evaluate calibration and generalization performance across various kernel methods and sample sizes. Two separate protocols are employed:

(A) Sample size variation experiment. This setting aims to evaluate how calibration performance evolves with increasing sample size. We consider the following methods: (i) KRR and (ii) KLR, each with either an RBF or Laplace kernel. For scalability, we use random Fourier features (RFF) for KLR. The Laplace kernel is approximated via a variant of RFF using samples from a Cauchy distribution. The corresponding feature mapping is implemented in our `LaplaceSampler` class.

For each dataset, we split the data into train/test with an 80/20 ratio while maintaining class balance. For training, we apply stratified subsampling of size $n \in \{50, \ldots, 2000\}$ (log-spaced, with 10 candidates). Each experiment is repeated with 5 different random seeds. The regularization hyperparameters are fixed as follows: $\alpha = 0.1$ for KLR and $\alpha = n^{-1/2}$ (RBF) or $\alpha = n^{-1/3}$ (Laplace) for KRR, based on empirical performance.

Bandwidth parameters for both kernels are selected via the median heuristic: for the RBF kernel, $\gamma = \frac{1}{2\sigma^2}$; for the Laplace kernel, $\gamma = \frac{1}{\sigma}$, where σ is the median pairwise Euclidean distance among training samples.

(B) Regularization parameter variation. To assess sensitivity to the regularization hyperparameter, we fix the training set size at $n = 2000$ and vary α over a logarithmic grid: $\alpha \in \{10^{-4}, \ldots, 10^2\}$. The same model families are considered as in (A), using fixed kernel parameters ($\gamma = 0.1$ for all models) to isolate the impact of α.

Evaluation metrics. We report three calibration metrics: ECE with optimal bins [11, 10], smoothed ECE, and MMCE. For KRR, probabilities are obtained by clipping regression outputs to $[10^{-6}, 1 - 10^{-6}]$ for stability.

Datasets. We use six binary classification datasets from OpenML: `kr-vs-kp`, `spambase`, `sick`, `churn`, and `Satellite`. Features are standardized after applying appropriate imputation and one-hot encoding using scikit-learn pipelines. All preprocessing steps are fit only on the training set to avoid data leakage.

Reproducibility. All experiments are implemented in Python using scikit-learn and CVXPY. Stratified sampling ensures class balance in subsamples. The full experimental code and data generation scripts will be made available upon publication.

[2] https://github.com/JonathanWenger/pycalib

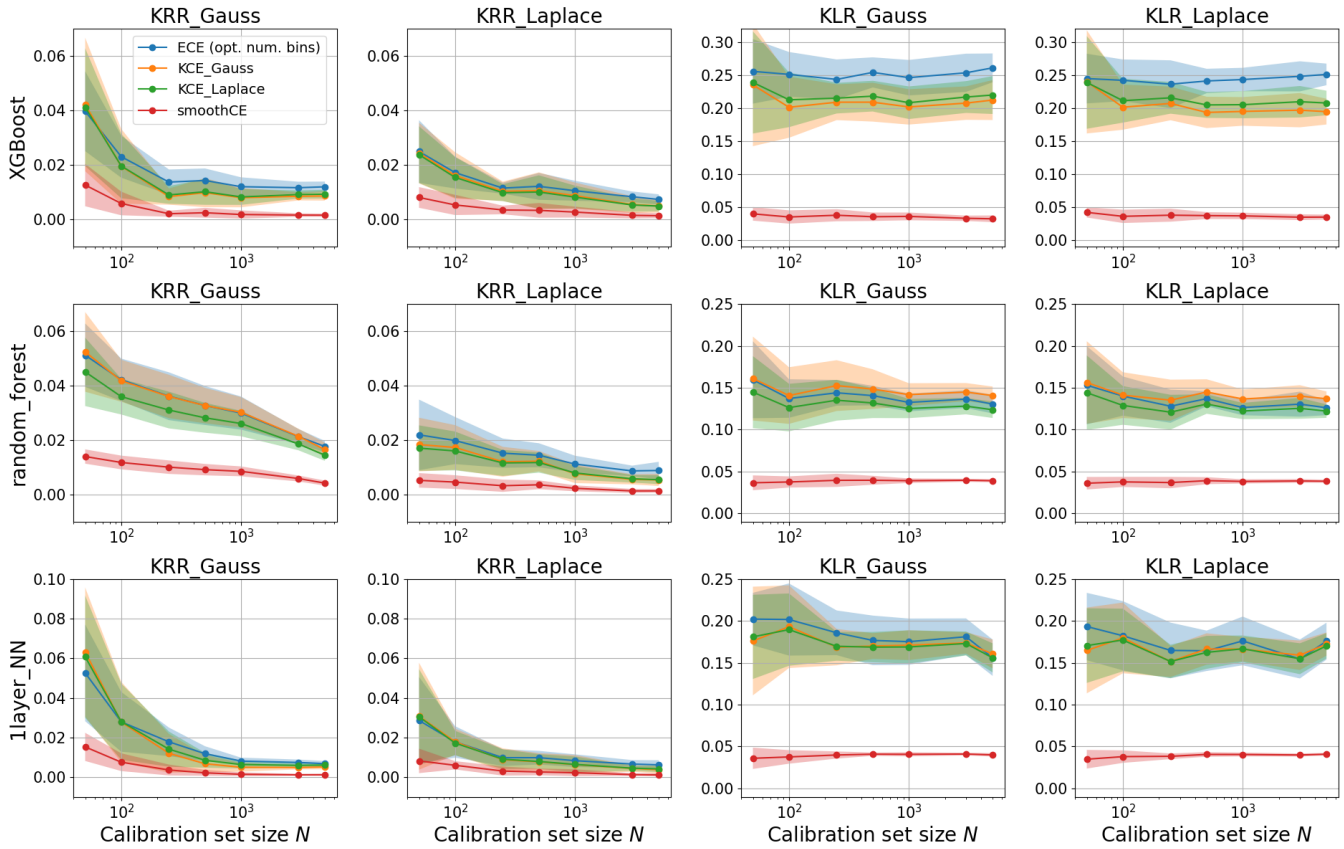

Figure 3: All Experimental Results of Recalibration: Effect of Recalibration Sample Size on Calibration Metrics on the KITTI dataset.

D Additional experimental results

In this section, we present additional experimental results. Figures 3–5 show the complete results of our recalibration experiments described in Section 5.2. Consistent with our theoretical analysis, we observe that increasing the regularization parameter λ leads to higher smooth CE for both Laplace and Gaussian kernels, reflecting the expected effect of stronger regularization. Conversely, increasing the recalibration sample size n_{re} consistently lowers the smooth CE, demonstrating the anticipated convergence behavior. These findings highlight the practical applicability of our theory to real-world recalibration scenarios.

We further show our experimental results on some real-world datasets explained in Appendix C.3 in Figures 7 and 8. Similarly, we observe that setting the regularization parameter λ too small or too large results in unstable smooth CE values for both Laplace and Gaussian kernels. In contrast, increasing the recalibration sample size n_{re} consistently reduces the smooth CE in most cases, exhibiting convergence behavior aligned with our theoretical results. These findings further support the reliability of our theory, demonstrating its applicability to real-world binary classification tasks.

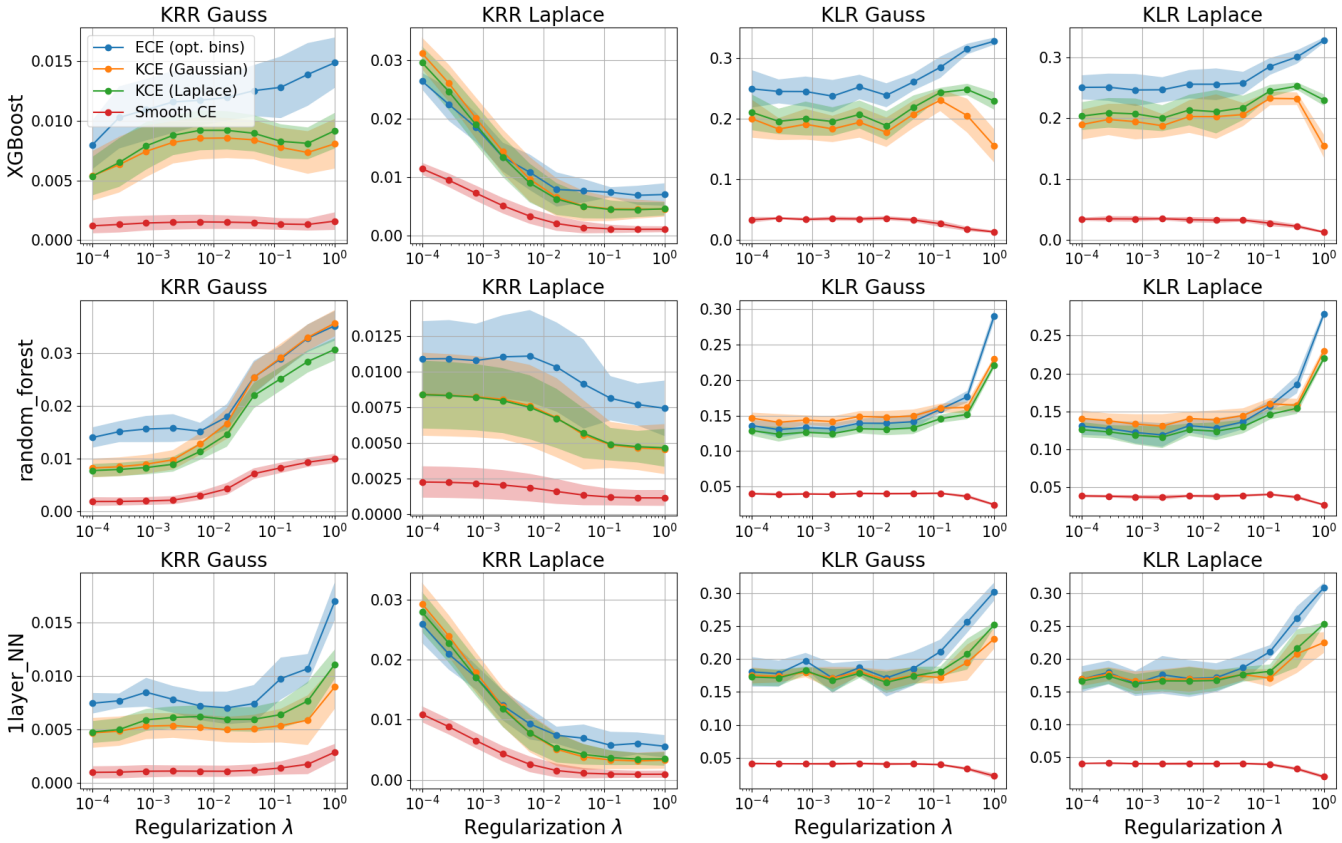

Figure 4: All Experimental Results of Recalibration: Effect of Regularization parameter λ on Calibration Metrics on the KITTI dataset.

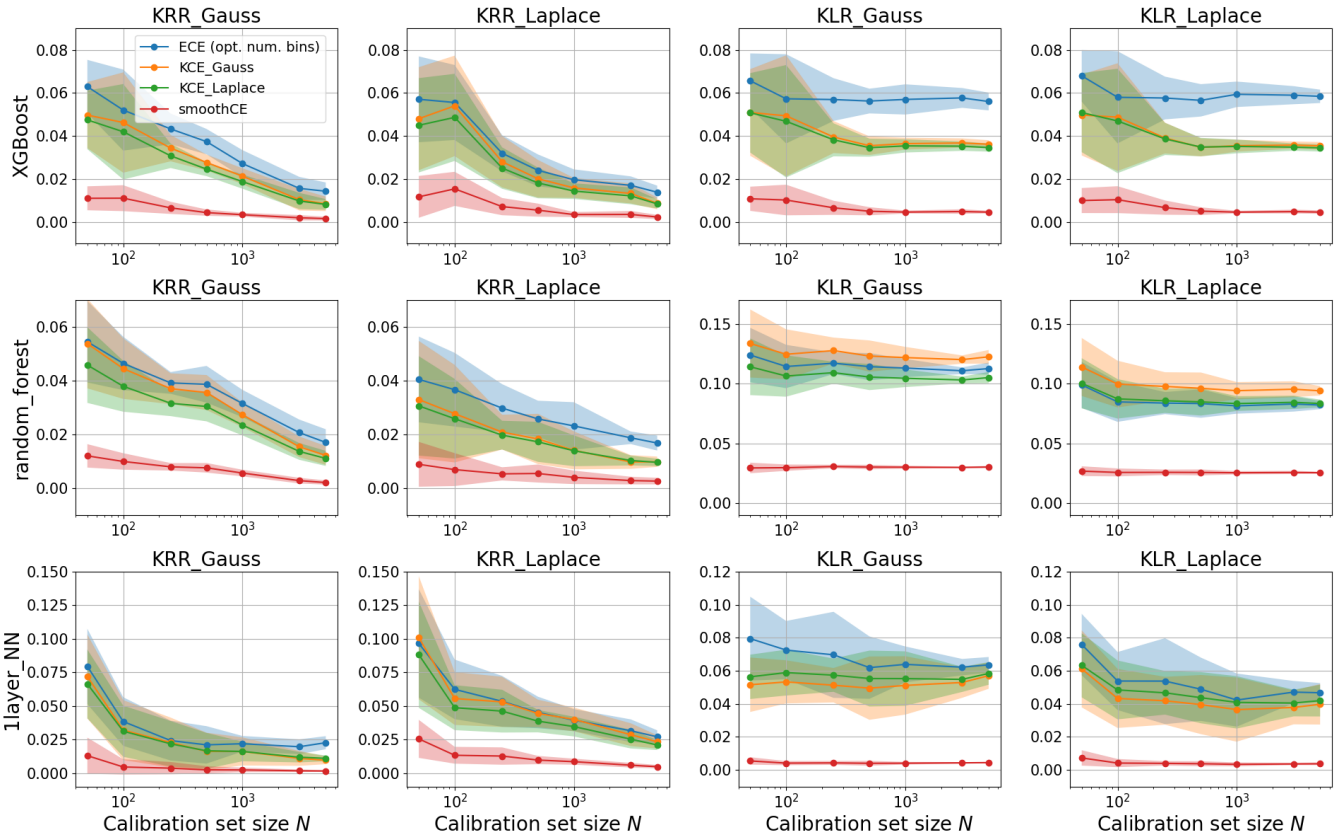

Figure 5: All Experimental Results of Recalibration: Effect of Recalibration Sample Size on Calibration Metrics on the PCam dataset.

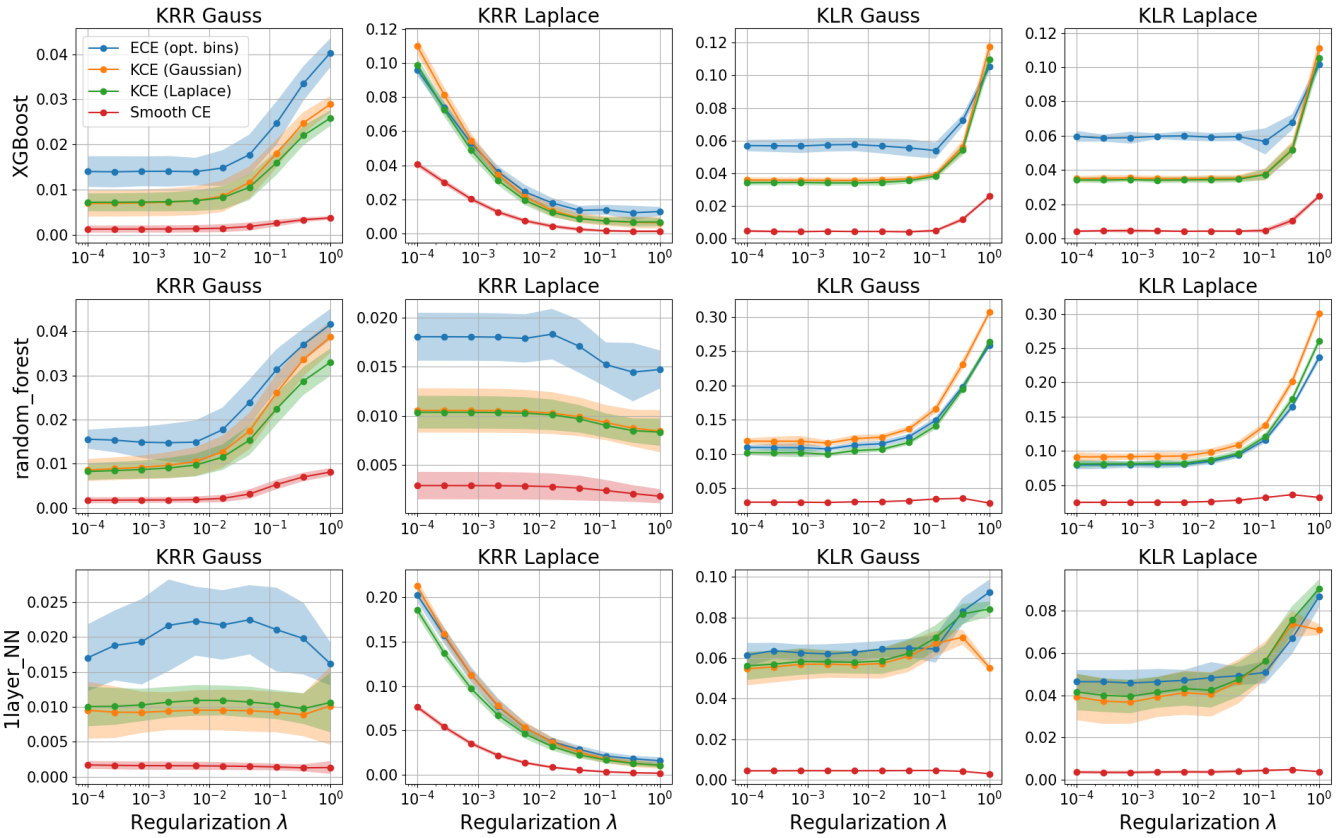

Figure 6: All Experimental Results of Recalibration: Effect of Regularization parameter λ on Calibration Metrics on the PCam dataset.

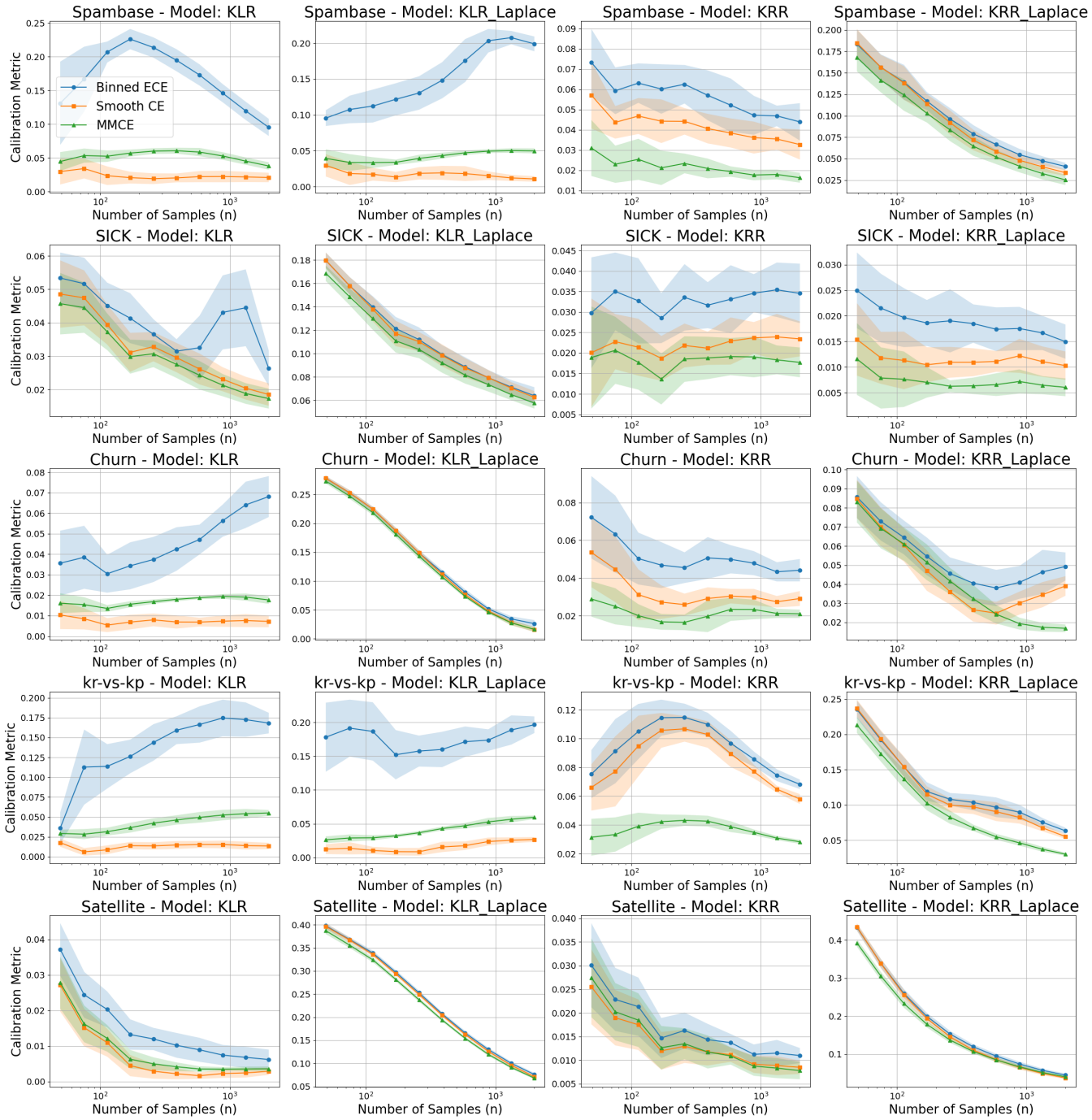

Figure 7: Effect of Sample Size on Calibration Metrics on the real world datasets.

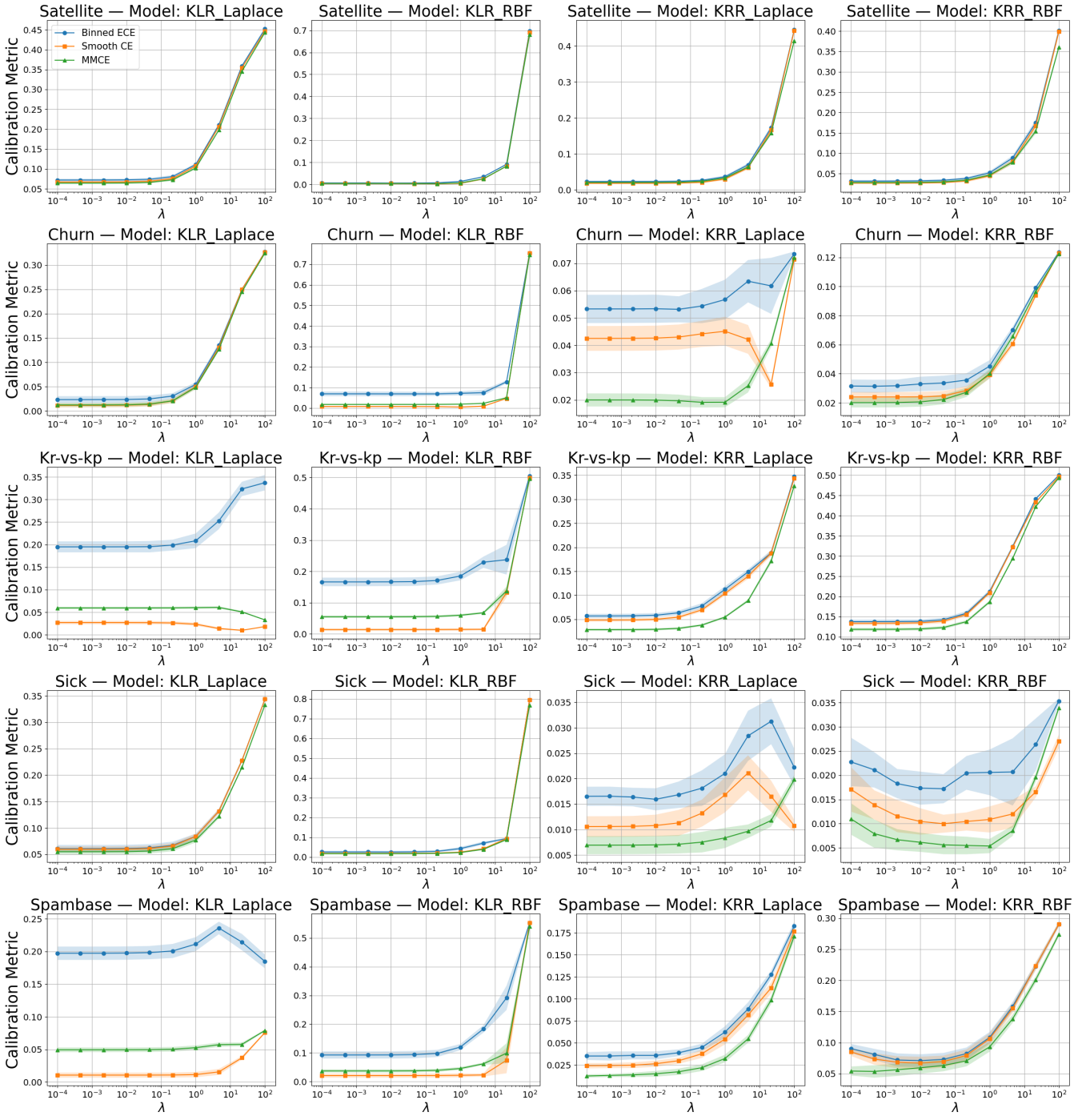

Figure 8: Effect of Sample Size on Calibration Metrics on the real world datasets.