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. [3], 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

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

where we used that $L_n(f) = L_n(f_n^*) + \mathrm{err}_n(f)$ and

$$\|r \circ f\|_{\mathcal{F}}^2 - \|f\|_{\mathcal{F}}^2 = (\|r \circ f\|_{\mathcal{F}} + \|f\|_{\mathcal{F}})(\|r \circ f\|_{\mathcal{F}} - \|f\|_{\mathcal{F}}) \leq 2\|r \circ f - f\|_{\mathcal{F}} \leq 2,$$

which is a direct consequence of the assumption on the norm of $\mathcal{F}$. 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. [2];

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

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

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

From the proof of Theorem 3.3 in Mohri et al. [32], we have, with probability at least $1 - \delta$,

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

This can be shown using McDiarmid's inequality, following the same argument as in the proof of Theorem 3.3 in Mohri et al. [32] 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. [32].

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

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

Then by the property of the supremum,

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

Then, from Propositions 11.2 and 11.2 in Mohri et al. [32],

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

and

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

and

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

holds. In conclusion, we have

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

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 [35], the RKHS associated with the Laplace kernel on $\mathbb{R}$ corresponds to the Sobolev space $H^1 = W^{2,1}$ (see also Buchholz [7]). Then, by Theorem 6 in Bourdaud [5], 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 [5] 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

Proof. The corresponding Rademacher complexity is

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

The first term can be bounded by $\alpha\Lambda/\sqrt{n}$ from the standard derivation of the Rademacher complexity for the kernel functions, see Mohri et al. [32] 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

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

In conclusion, we have

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

We obtain the result. □

A.5 Proof of Theorem 6

Proof. From Proposition 7 in Rakhlin and Zhai [35], 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 [7]). According to Theorem 7 in Bourdaud [5], 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 [36], 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 [36], 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 [36] for details.

With this in mind, and following the argument in the proof of Claim 5.1 in Błasiok et al. [3], we upper bound the post-processing gap as follows. By the definition of the infimum,

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

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

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

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. [32], we have

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

Then we have

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

□

A.7 Proof of Corollary 3

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

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

The proof of this is exactly the same as that of Lemma 4.7 in Błasiok et al. [3], 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

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

where we used that $L_n(f) = L_n(f_n^*) + \mathrm{err}_n(f)$ and

$$\|r \circ g\|_{\mathcal{G}}^2 - \|g\|_{\mathcal{G}}^2 = (\|r \circ g\|_{\mathcal{G}} + \|g\|_{\mathcal{G}})(\|r \circ g\|_{\mathcal{G}} - \|g\|_{\mathcal{G}}) \leq 2\|r \circ g - g\|_{\mathcal{G}} \leq 2G^2$$

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. (4), 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 4

By the assumptions, we can apply Corollary 3 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.

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 constants Λ and α such that $\sup_{x,x' \in \mathcal{X}} k(x, x') \leq \Lambda$, $\|g\|_{\mathcal{H}} \leq \alpha$, and $|b| \leq \alpha\Lambda + 1$. Then, with probability at least $1 - \delta$ over the draw of the training dataset, it holds that*

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

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. [3], 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. [3], 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. [3].

Błasiok et al. [3] 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. [2] 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*

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

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. [2] for details. They showed that the smooth CE both upper and lower bounds the true distance to calibration:

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

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

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

is discontinuous, and Błasiok et al. [2] showed that ECE does not lower bound $\mathrm{dCE}_{\mathcal{D}}(f)$ unless continuity of the conditional expectation is assumed.

Błasiok et al. [2] 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:

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

where

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

and $w(I)$ denotes the width of interval I. Then, the following bound holds (Theorem 6.3 in Błasiok et al. [2]):

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

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).

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 [13], Fujisawa and Futami [12].

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 [8], Random Forests [6], 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. [39] [1].

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 [13], Fujisawa and Futami [12]. 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.

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

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 $n \in \{50, \dots, 2000\}$ (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: $\alpha \in \{10^{-4}, \dots, 10^{2}\}$. 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 [13, 12], 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.

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.

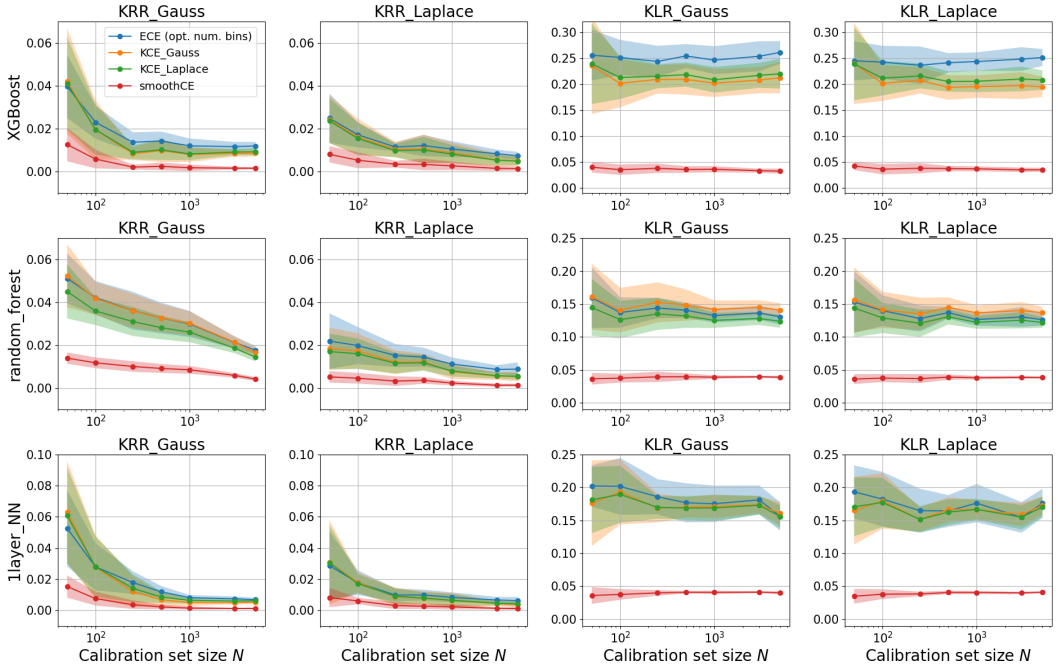

Figure 3: All Experimental Results of Recalibration: Effect of Recalibration Sample Size on Calibration Metrics on the KITTI dataset.

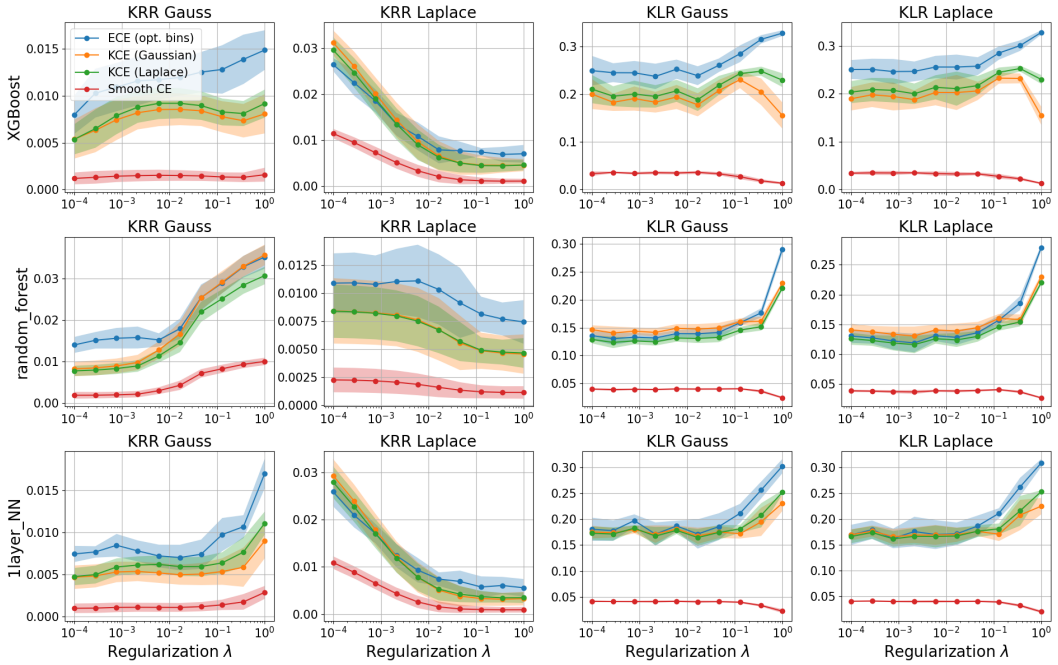

Figure 4: All Experimental Results of Recalibration: Effect of Regularization parameter λ on Calibration Metrics on the KITTI dataset.

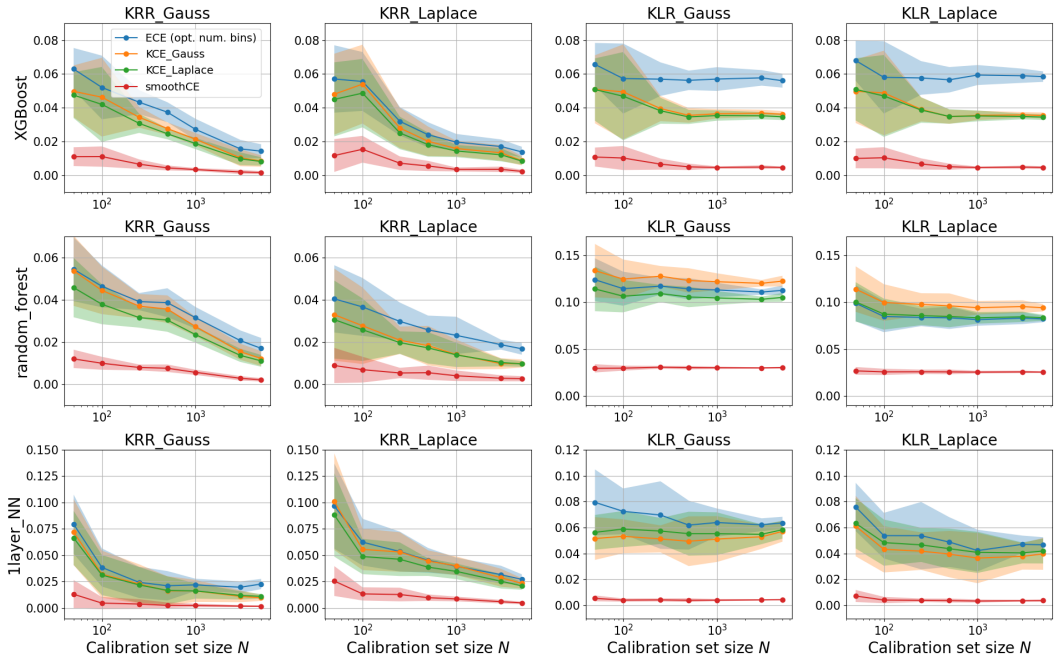

Figure 5: All Experimental Results of Recalibration: Effect of Recalibration Sample Size on Calibration Metrics on the PCam dataset.

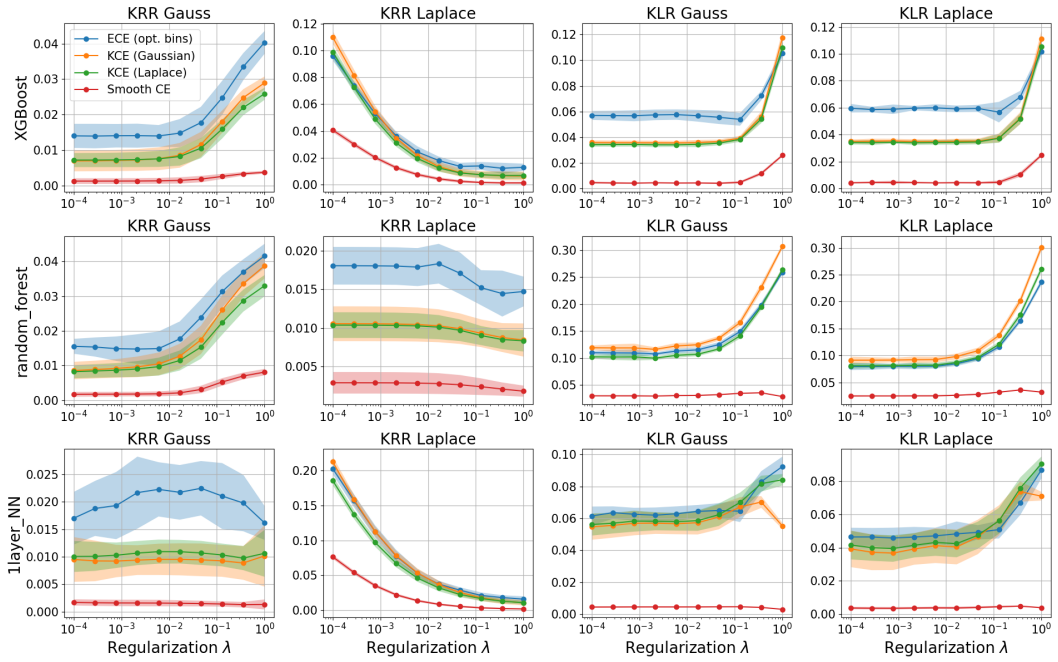

Figure 6: All Experimental Results of Recalibration: Effect of Regularization parameter λ on Calibration Metrics on the PCam dataset.

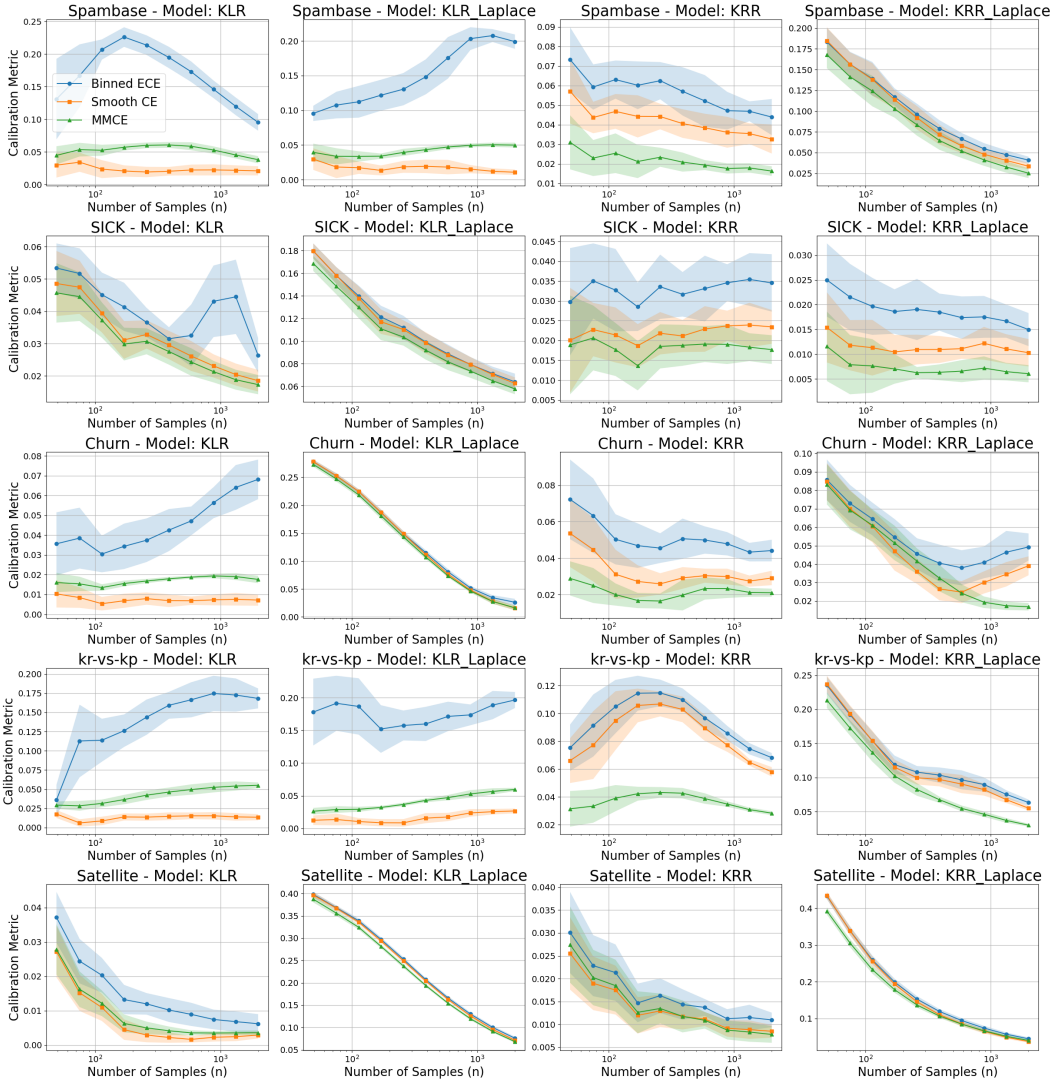

Figure 7: Effect of Sample Size on Calibration Metrics on the real world datasets.

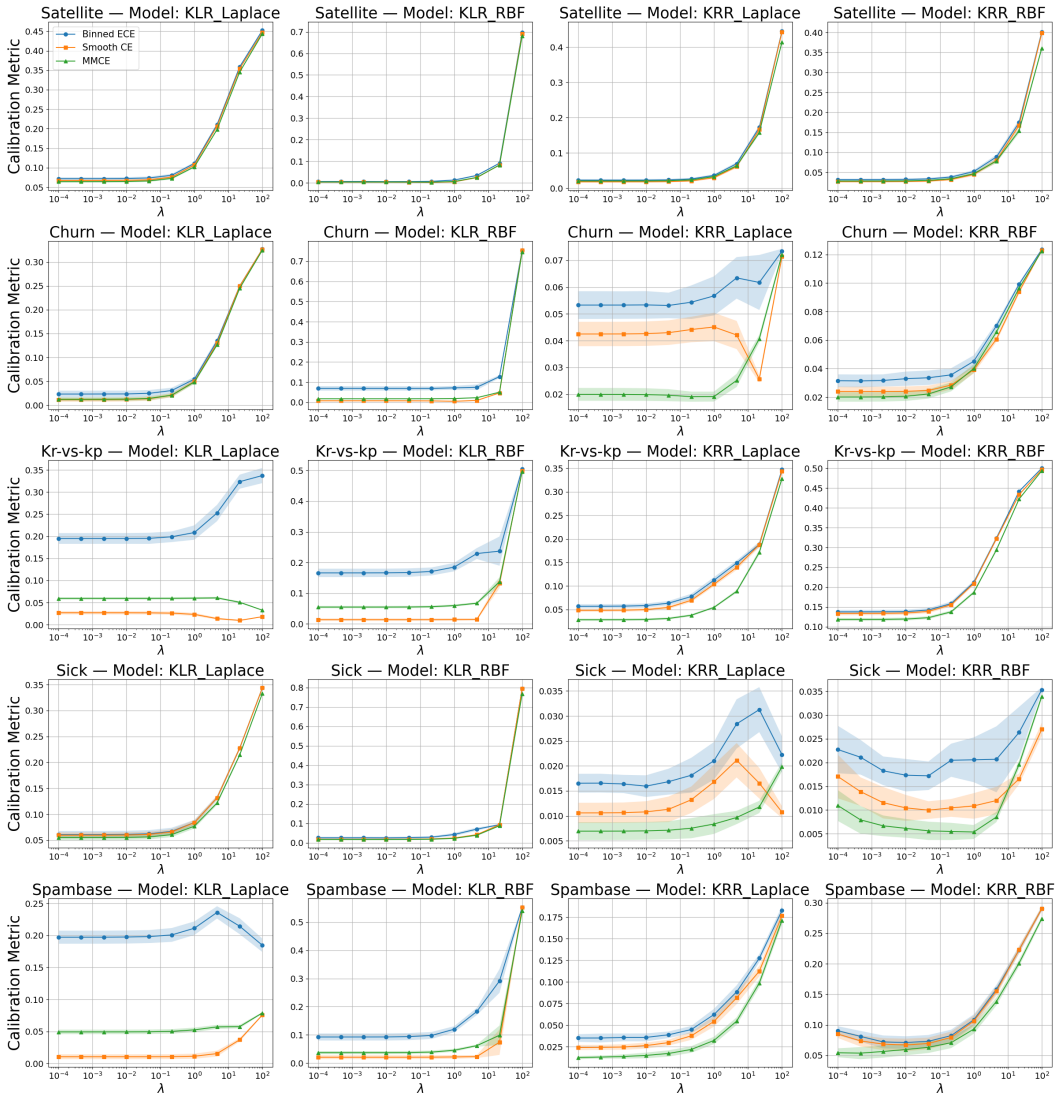

Figure 8: Effect of Sample Size on Calibration Metrics on the real world datasets.