Appendix

A ADDITIONAL FORMULATION DETAILS

A.1 MAXIMUM MEAN CALIBRATION ERROR

Along with our proposed S-TLBCE and the baseline S-MCE, we also apply the S-MMCE loss function proposed in Fisch et al. (2022) for training a selective calibration system. This loss function is defined as:

$$l_{\text{S-MMCE}}(f, \hat{g}, h, x, y) = \frac{1}{n}\Big[\sum_{i,j} |y_i^c - \hat{h}(f(x_i))|^q |y_j^c - \hat{h}(f(x_j))|^q \hat{g}(x_i)\hat{g}(x_j)\phi\big(\hat{h}(f(x_i)), \hat{h}(f(x_j))\big)\Big]^{\frac{1}{q}} \tag{10}$$

where ϕ is some similarity kernel, like Laplacian. On a high level, this loss penalizes pairs of instances that have similar confidence and both are far from the true label y^c (which denotes prediction correctness 0 or 1). Further details and motivation for such an objective can be found in Fisch et al. (2022).

A.2 HIGH PROBABILITY COVERAGE GUARANTEES

Since $\hat{g}(x) \geqslant \tau$ is a random variable with a Bernoulli distribution, we can apply the Hoeffding bound (Hoeffding, 1963) to guarantee that with high probability empirical coverage $\hat{\beta}$ (the proportion of the target distribution where $\hat{g}(x) \geqslant \tau$) will be in some range.

Given a set $\mathcal{V}$ of n_u i.i.d. unlabeled examples from the target distribution, we denote empirical coverage on $\mathcal{V}$ as $\tilde{\beta}$. With probability at least $1 - \delta$, $\hat{\beta}$ will be in the range $[\tilde{\beta} - \epsilon, \tilde{\beta} + \epsilon]$ where

$$\epsilon = \sqrt{\frac{\log(\frac{2}{\delta})}{2n_u}}$$

For some critical coverage level β, τ can be decreased until $\tilde{\beta} - \epsilon \geqslant \beta$.

B ADDITIONAL EXPERIMENT DETAILS

In training we drop the denominator in l_{sel}, as the coverage loss suffices to keep $\hat{g}$ from collapsing to 0. Recalibration model code is taken from the accompanying code releases from Guo et al. (2017)[4] (Temperature Scaling) and Kumar et al. (2019)[5] (Platt Scaling, Histogram Binning, Platt Binning).

B.1 CALIBRATION MEASURES

We calculate ECE-1 and ECE-2 using the python library released by Kumar et al. (2019) [6]. ECE-q is calculated as:

$$\text{ECE-q} = \left(\frac{1}{|B|}\sum_{j=1}^{|B|}\left(\frac{\sum_{i\in B_j}\mathbf{1}\{y_i = \hat{y}_i\}}{|B_j|} - \frac{\sum_{i\in B_j}\tilde{f}(x_i)}{|B_j|}\right)^q\right)^{\frac{1}{q}} \tag{11}$$

where $B = B_1, ..., B_m$ are a set of m equal-mass prediction bins, and predictions are sorted and binned based on their maximum confidence $\tilde{f}(x)$. We set $m = 15$.

---

[4] `https://github.com/gpleiss/temperature_scaling`

[5] `https://github.com/p-lambda/verified_calibration`

[6] `https://github.com/p-lambda/verified_calibration`

B.2 BASELINES

Our selection baselines include confidence-based rejection (“Confidence”) and multiple out-of-distribution (OOD) detection methods (“Iso. Forest”, “One-class SVM”). The Confidence baseline rejects examples with the smallest $\hat{f}(x)$ (or $\hat{h}(f(x))$), while the OOD methods are measured in the embedding space of the pre-trained model. These methods are typical in selective classification for accuracy. All selection baselines are applied to the recalibrated model in order to make the strongest comparison. We make further comparisons to recalibration baselines, including temperature scaling and Platt scaling, which have been described previously. Also present are binning methods like histogram binning and Platt binning (Kumar et al., 2019). While these algorithms are non-differentiable, and thus not eligible to be used as h in selective recalibration, they are quite effective and thus important to include on their own for the sake of a thorough empirical investigation.

Next we describe how baseline methods are implemented. Our descriptions are based on creating an ordering of the test set such that at a given coverage level β, a $1-\beta$ proportion of examples from the end of the ordering are rejected.

B.2.1 CONFIDENCE-BASED REJECTION

Confidence based rejection is performed by ordering instances in a decreasing order based on $\tilde{f}(x)$, the maximum confidence the model has in any class for that example.

B.2.2 OUT OF DISTRIBUTION SCORES

The sklearn python library (Pedregosa et al., 2011) is used to produce the One-Class SVM and Isolation Forest models. Anomaly scores are oriented such that more typical datapoints are given higher scores; instances are ranked in a decreasing order.

B.3 IN-DISTRIBUTION EXPERIMENTS

Our selector g is a shallow fully-connected network (2 hidden layers with dimension 128).

B.3.1 CAMELYON17

Camelyon17 (Bandi et al., 2018) is a task where the input x is a 96x96 patch of a whole-slide image of a lymph node section from a patient with potentially metastatic breast cancer, the label y is whether the patch contains a tumor, and the domain d specifies which of 5 hospitals the patch was from. We pre-train a DenseNet-121 model on the Camelyon17 train set using the code from Koh et al. (2021)[7]. The validation set has 34,904 examples and accuracy of 91%, while the test set has 84,054 examples, and accuracy of 83%. Our selector g is trained with a learning rate of 0.0005, the coverage loss weight λ is set to 32 (following (Geifman & El-Yaniv, 2019)), and the model is trained with 1000 samples for 1000 epochs with a batch size of 100.

B.3.2 IMAGENET

ImageNet is a large scale image classification dataset. We extract the features, scores, and labels from the 50,000 ImageNet validation samples using a pre-trained ResNet34 model from the torchvision library. Our selector g is trained with a learning rate of 0.00001, the coverage loss weight λ is set to 32 (following (Geifman & El-Yaniv, 2019)), and the model is trained with 2000 samples for 1000 epochs with a batch size of 200.

B.4 OUT-OF-DISTRIBUTION EXPERIMENTS

Our selector g is a shallow fully-connected network (1 hidden layer with dimension 64) trained with a learning rate of 0.0001, the coverage loss weight λ is set to 8, and the model is trained for 50 epochs (to avoid overfitting since this is an OOD setting) with a batch size of 256.

---

[7] `https://github.com/p-lambda/wilds`

B.4.1 RXRX1

RxRx1 (Taylor et al., 2019) is a task where the input x is a 3-channel image of cells obtained by fluorescent microscopy, the label y indicates which of the 1,139 genetic treatments (including no treatment) the cells received, and the domain d specifies the batch in which the imaging experiment was run. The validation set has 9,854 examples and accuracy of 18%, while the test set has 34,432 examples, and accuracy of 27%. 1000 samples are drawn for model training. Gaussian noise with mean 0 and standard deviation 1 is added to training examples in order to promote robustness.

B.4.2 CIFAR-100

CIFAR-100 is a well-known image classification dataset, and we perform zero-shot image classification with CLIP. We draw 2000 samples for model training, and test on 50,000 examples drawn from the 750,000 examples in CIFAR-100-c. Data augmentation in training is performed using AugMix (Hendrycks et al., 2019) with a severity level of 3 and a mixture width of 3.

C ADDITIONAL EXPERIMENT RESULTS

C.1 SELECTIVE RECALIBRATION WITH I.I.D. DATA

Here we include ECE-1 results for the experiments in Section 5.1. We note that unreported experiments showed similar results with respect to Brier Score. Although our focus in this work is ECE, for completeness those results will be included in future versions of this paper.

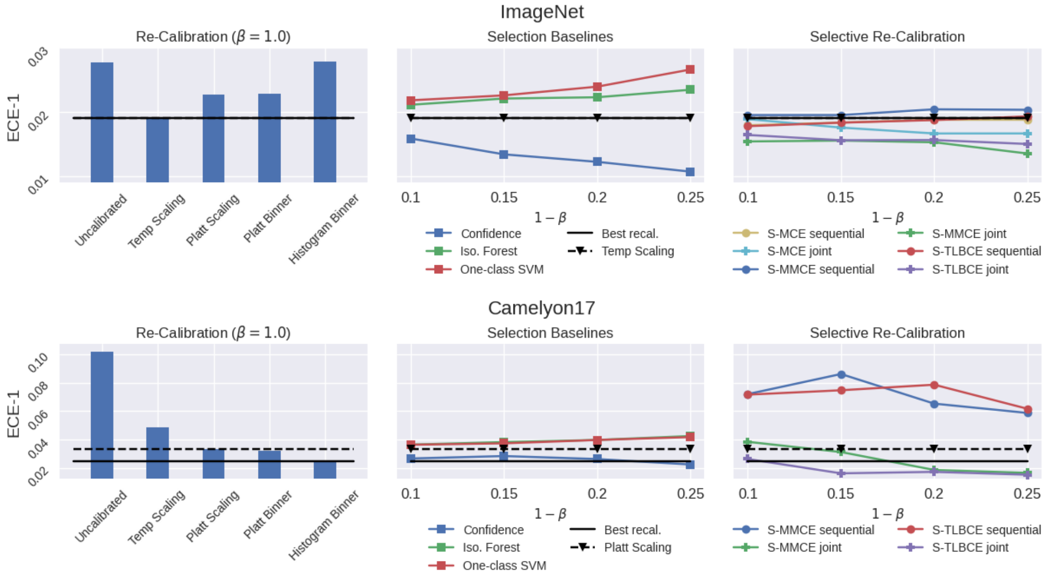

Figure 4: Selective calibration error on ImageNet and Camelyon17 for coverage level $\beta \in \{0.75, 0.8, 0.85, 0.9\}$. **Left**: Various re-calibration methods are trained using labeled validation data. **Middle**: Selection baselines including confidence-based rejection and various OOD measures. **Right**: Selective re-calibration with different loss functions.

C.2 TRADE-OFFS BETWEEN CALIBRATION ERROR AND ACCURACY

While accurate probabilistic output is the only concern in some domains and should be of at least some concern in most domains, discrete label accuracy can also be important in some circumstances. Figure 5 shows the selective accuracy and confidence histogram for the selective recalibration model trained with S-TLBCE for RxRx1 and CIFAR-100 (and applied to shifted distributions). Together, these figures illustrate that under different data and prediction distributions, selective recalibration may increase or decrease accuracy. For RxRx1, the model tends to reject examples with higher confidence, which also tend to be more accurate. Thus, while ECE@β may improve with respect to the full dataset, Accuracy@β is worse. On the other hand, for CIFAR-100-C, the model tends to reject examples with lower confidence, which also tend to be less accurate. Accordingly, both ECE@β and Accuracy@β improve with respect to the full dataset.

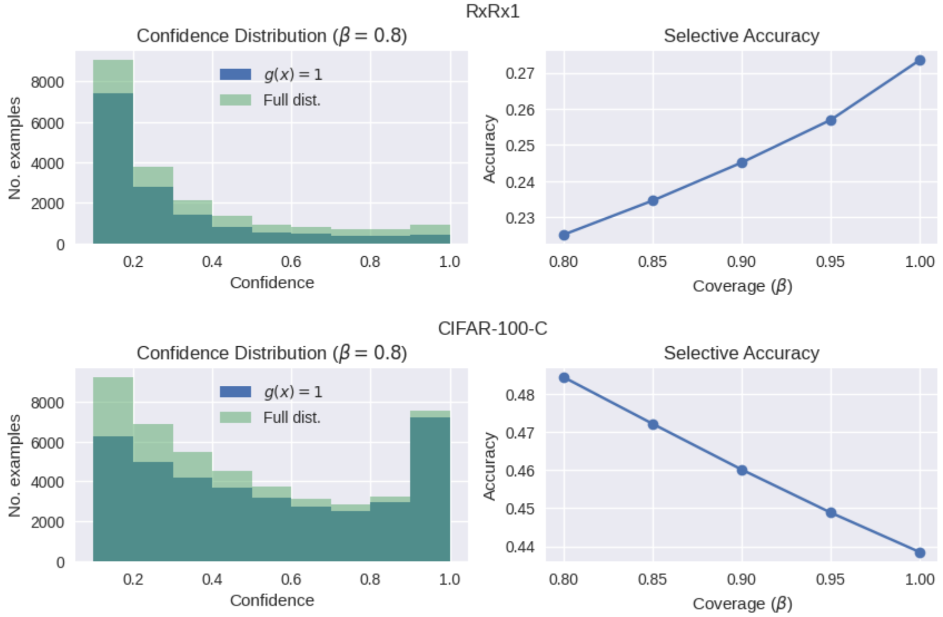

Figure 5: Left: Distribution of confidence among the full distribution and those examples accepted for prediction (i.e. where $g(x) = 1$) at coverage level $\beta = 0.8$. Right: Selective accuracy in the range $\beta = [0.8, 1.0]$.

D TECHNICAL DETAILS

D.1 DETAILS ON DATA GENERATION MODEL

Definition 2 (Formal version of definition 1)**.** *For $\theta^* \in \mathbb{R}^p$, a $(\theta^*, \sigma, \alpha, r_1, r_2)$-perturbed truncated-Gaussian model is defined as the following distribution over $(x, y) \in \mathbb{R}^p \times \{1, -1\}$:*

$$x \mid y \sim zJ_1 + (1 - z)J_2.$$

Here, J_1 and J_2 are two truncated Guassian distributions, i.e.

$$J_1 \sim \rho_1 \mathcal{N}(y \cdot \theta^*, \sigma^2 I)\mathbf{1}\{x \in \mathbb{B}(\theta^*, r_1) \cup \mathbb{B}(-\theta^*, r_1)\},$$
$$J_2 \sim \rho_2 \mathcal{N}(-y \cdot \alpha\theta^*, \sigma^2 I)\mathbf{1}\{x \in \mathbb{B}(\alpha \cdot \theta^*, r_2) \cup \mathbb{B}(-(\alpha \cdot \theta^*, r_2)\}$$

where ρ_1, ρ_2 are normalization coefficients to make J_1 and J_2 properly defined; y follows the Bernoulli distribution $\mathbb{P}(y = 1) = \mathbb{P}(y = -1) = 1/2$; and z follows a Bernoulli distribution $\mathbb{P}(z = 1) = \beta$.

For simplicity, throughout this paper, we set $\rho_1 = \rho_2$ and this is always achievable by setting r_1/r_2 appropriately. We also set $\alpha \in (0, 1/2)$.

D.2 JOINT LEARNING VERSUS SEQUENTIAL LEARNING

We also demonstrate that jointly learning a selection model g and temperature scaling parameter T can outperform sequential learning of g and T.

Let us first denote $\tilde{g} := \operatorname{argmin} \text{S-ECE}(g)$ such that $\mathbb{E}[\tilde{g}(x)] \geq \beta$ and $\tilde{T} := \operatorname{argmin} \text{R-ECE}(T)$. We denote two types of expected calibration error under sequential learning of g and T, depending on which is optimized first.

$$\text{ECE}^{R\to S} := \min_{g:\mathbb{E}[g(x)]\geq\beta} \text{S-ECE}(g,\tilde{T});$$
$$\text{ECE}^{S\to R} := \min_{T\in\mathbb{R}} \text{R-ECE}(\tilde{g},T).$$

Our second theorem shows these two types of expected calibration error for sequential learning are lower bounded, while jointly learning g, T can reach zero calibration error.

Theorem 2. *Under Assumption 3, if $\beta > 2(1-\beta)$, for any $\delta \in (0,1)$ and $\hat{\theta}$ output by $\mathscr{A}$, there exist thresholds $M \in \mathbb{N}^+$ and $\tau_2 > \tau_1 > 0$: if $\max\{r_1, r_2, \sigma\} < \tau_2$, $\tau_1 < \sigma$, and $m > M$, then there exists a positive lower bound L, with probability at least $1-\delta$ over S^{tr}*

$$\min\left\{\mathit{ECE}^{R\to S}, \mathit{ECE}^{S\to R}\right\} > L.$$

However, there exists g_0 satisfying $\mathbb{E}[g_0(x)] \geq \beta$ and T_0, such that

$$\mathit{SR\text{-}ECE}(g_0, T_0) = 0.$$

Intuition and interpretation. If we first optimize the temperature scaling model to obtain $\tilde{T}$, $\tilde{T}$ will not be equal to $\hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2)$. Then, when applying selection, there exists no g that can reach 0 calibration error since the temperature is not optimal for data in $\mathcal{A}$ or $\mathcal{B}$. On the other hand, if we first optimize the selection model and obtain $\tilde{g}$, $\tilde{g}$ will reject points in $\mathcal{A}$ instead of those in $\mathcal{B}$ because points in $\mathcal{A}$ incur higher calibration error, and thus data from both $\mathcal{A}$ and $\mathcal{B}$ will be selected. In that case, temperature scaling not will be able to push calibration error to zero because, similar to the case in the earlier R-ECE analysis, the calibration error in $\mathcal{A}$ and $\mathcal{B}$ cannot reach 0 simultaneously using a single temperature scaling model. On the other hand, the optimal jointly-learned solution yields a set of predictions with zero expected calibration error.

D.3 DETAILS ON $\hat{\theta}$

Recall that we consider the $\hat{\theta}$ that is the output of a training algorithm $\mathscr{A}(S^{tr})$ that takes the i.i.d. training data set $S^{tr} = \{(x_i^{tr}, y_i^{tr})\}_{i=1}^m$ as input. We imposed the following assumption on $\hat{\theta}$.

Assumption 3. *For any given $\delta \in (0,1)$, there exists $\theta_0 \in \mathbb{R}^p$ with $\|\theta_0\| = \Theta(1)$, that with probability at least $1-\delta$*

$$\|\hat{\theta} - \theta_0\| < \phi(\delta, m),$$

and $\phi(\delta, m)$ goes to 0 as m goes to infinity. Also, there exist a threshold $M \in \mathbb{N}^+$ such that if $m > M$, $\phi(\delta, m)$ is a decreasing function of δ and n. Moreover,

$$\min\left\{\frac{\theta_0^\top\theta^*}{\|\theta_0\|^2}, \theta_0^\top\theta^*, \|\theta_0\|\right\} > 0.$$

We will prove the following lemma as a cornerstone for our future proofs.

Lemma 1. *Under Assumption 3, for any $\delta \in (0,1)$, there exists a threshold $M \in \mathbb{N}^+$, and constants $0 < I_1 < I_2$, $0 < I_3 < I_4 < \alpha I_3$, $0 < I_5 < I_6$, such that if $m > M$, with probability at least $1-\delta$ over the randomness of S^{tr},*

$$\frac{\hat{\theta}^\top\theta^*}{\|\hat{\theta}\|^2} \in [I_1, I_2], \quad \hat{\theta}^\top\theta^* \in [I_3, I_4], \quad \|\hat{\theta}\| \in [I_5, I_6].$$

Proof. Under Assumption 3, we know $m \to \infty$ leads to $\hat{\theta} \to \theta_0$. In addition, for any $\delta \in (0,1)$ there exists a threshold $M \in \mathbb{N}^+$ such that if $m > M$, $\phi(\delta, m)$ is a decreasing function of δ and m, which leads to

$$\frac{\hat{\theta}^\top\theta^*}{\|\hat{\theta}\|^2} \in [\frac{\theta_0^\top\theta^*}{\|\theta_0\|^2} - \varepsilon, \frac{\theta_0^\top\theta^*}{\|\theta_0\|^2} + \varepsilon], \quad \hat{\theta}^\top\theta^* \in [\theta_0^\top\theta^* - \varepsilon, \theta_0^\top\theta^* + \varepsilon], \quad \|\hat{\theta}\| \in [\|\theta_0\| - \varepsilon, \|\theta_0\| + \varepsilon]$$

for some small $\varepsilon > 0$ that makes the left end of the above intervals larger than 0 and $\theta_0^\top\theta^* + \varepsilon < \alpha(\theta_0^\top\theta^* - \varepsilon)$ hold for all r_1, r_2, σ, m as long as $m > M$. Then, we set I_i's accordingly to each value above. □

D.3.1 AN EXAMPLE ON TRAINING $\hat{\theta}$

In this section, we provide one example to justify Assumption 3, i.e. $\hat{\theta} = \sum_{i=1}^{m} x_i^{tr} y_i^{tr}/m$, where the training set is drawn from an unperturbed Gaussian mixture, i.e. $x^{tr}|y^{tr} \sim \mathcal{N}(y^{tr} \cdot \theta^*, \sigma^2 I)$ and y^{tr} follows a Bernoulli distribution $\mathbb{P}(y^{tr} = 1) = 1/2$. Directly following the analysis of Zhang et al. (2022), we have

$$\hat{\theta}^\top \theta^* = O_{\mathbb{P}}(\frac{1}{\sqrt{m}})\|\theta^*\| + \|\theta^*\|^2.$$

For $\|\hat{\theta}\|^2$, notice that

$$\hat{\theta} = \theta^* + \epsilon_m$$

where $\epsilon_m \sim \mathcal{N}(0, \frac{\sigma^2 I}{m})$. Then, we have

$$\|\hat{\theta}\|^2 = \|\theta^*\|^2 + 2\epsilon_m^\top \theta^* + \|\epsilon_m\|^2 = \|\theta^*\|^2 + \frac{p}{m} + O_{\mathbb{P}}(\frac{\sqrt{p}}{m}) + O_{\mathbb{P}}(\frac{1}{\sqrt{m}})\|\theta^*\|.$$

Given $p/m = O(1)$, combined with the form of classic concentration inequalities, one can verify this example satisfies Assumption 3.

D.4 BACKGROUND: ECE CALCULATION

Recall we denote $\hat{f}(x) = \max\{\hat{p}_{-1}(x), \hat{p}_1(x)\}$ and denote the predicition result $\hat{y} = \hat{C}(x)$. The definition of ECE is:

$$\text{ECE} = \mathop{\mathbb{E}}_{\hat{f}(x)} |\mathbb{P}[y = \hat{y} \mid \hat{f}(x) = p] - p|.$$

Notice that there are two cases.

- Case 1: $\hat{f}(x) = \hat{p}_1(x)$, by reparameterization, we have

$$\begin{aligned}|\mathbb{P}[y = \hat{y} \mid \hat{f}(x) = p] - p| &= \left|\mathbb{P}[y = 1 \mid \hat{f}(x) = \frac{e^{2v}}{1+e^{2v}}] - \frac{e^{2v}}{1+e^{2v}}\right| \\ &= \left|\mathbb{P}[y = 1 \mid \hat{\theta}^\top x = v] - \frac{e^{2v}}{1+e^{2v}}\right|.\end{aligned}$$

- Case 2: $\hat{f}(x) = \hat{p}_{-1}(x)$, by reparameterization, we have

$$\begin{aligned}|\mathbb{P}[y = \hat{y} \mid \hat{f}(x) = p] - p| &= \left|\mathbb{P}[y = -1 \mid \hat{f}(x) = \frac{1}{1+e^{2v}}] - \frac{1}{1+e^{2v}}\right| \\ &= \left|\mathbb{P}[y = -1 \mid \hat{\theta}^\top x = v] - \frac{1}{1+e^{2v}}\right| \\ &= \left|1 - \mathbb{P}[y = -1 \mid \hat{\theta}^\top x = v] - (1 - \frac{e^{2v}}{1+e^{2v}})\right| \\ &= \left|\mathbb{P}[y = 1 \mid \hat{\theta}^\top x = v] - \frac{e^{2v}}{1+e^{2v}}\right|.\end{aligned}$$

To summarize,

$$\text{ECE} = \mathop{\mathbb{E}}_{\hat{f}(x)} |\mathbb{P}[y = \hat{y} \mid \hat{f}(x) = p] - p| = \mathbb{E}_{\hat{\theta}^\top x} \left|\mathbb{P}[y = 1 \mid \hat{\theta}^\top x = v] - \frac{1}{1+e^{-2v}}\right|.$$

Temperature scaling.

$$p^T_{-1}(x) = \frac{1}{e^{2\cdot\hat\theta^\top x/T}+1}, \quad p^T_1(x) = \frac{e^{2\cdot\hat\theta^\top x/T}}{e^{2\cdot\hat\theta^\top x/T}+1}. \tag{12}$$

Thus,

$$RECE = \mathbb{E}_{\hat\theta^\top x}\left|\mathbb{P}[y=1 \mid \hat\theta^\top x = vT] - \frac{1}{1+e^{-2v}}\right|.$$

Platt scaling.

$$p^{w,b}_{-1}(x) = \frac{1}{e^{2w\cdot\hat\theta^\top x+2b}+1}, \quad p^{w,b}_1(x) = \frac{1}{e^{-2w\cdot\hat\theta^\top x-2b}+1}. \tag{13}$$

$$\mathrm{ECE}_{w,b} = \mathbb{E}_{\hat\theta^\top x}\left|\mathbb{P}[y=1 \mid w\cdot\hat\theta^\top x + b = v] - \frac{1}{1+e^{-2v}}\right|.$$

ECE calculation. The distribution of $\hat\theta^\top x$ has the following properties.

- $\hat\theta^\top x|y=1 \sim z\rho_1\mathcal{N}(\hat\theta^\top\theta^*, \sigma^2\|\hat\theta\|^2)\mathbf{1}\{\hat\theta^\top x \in \mathbb{B}(\hat\theta^\top\theta^*, r_1\|\hat\theta\|) \cup \mathbb{B}(-\hat\theta^\top\theta^*, r_1\|\hat\theta\|)\}$ $+ (1-z)\rho_2\mathcal{N}(-\alpha\cdot\hat\theta^\top\theta^*, \sigma^2\|\hat\theta\|^2))\mathbf{1}\{x \in \mathbb{B}(\alpha\hat\theta^\top\theta^*, r_2\|\hat\theta\|) \cup \mathbb{B}(-(\alpha\hat\theta^\top\theta^*, r_2\|\hat\theta\|)\}$;
- $\hat\theta^\top x|y=-1 \sim z\rho_1\mathcal{N}(-\hat\theta^\top\theta^*, \sigma^2\|\hat\theta\|^2)\mathbf{1}\{\hat\theta^\top x \in \mathbb{B}(\hat\theta^\top\theta^*, r_1\|\hat\theta\|) \cup \mathbb{B}(-\hat\theta^\top\theta^*, r_1\|\hat\theta\|)\}$ $+ (1-z)\rho_2\mathcal{N}(\alpha\cdot\hat\theta^\top\theta^*, \sigma^2\|\hat\theta\|^2))\mathbf{1}\{\hat\theta^\top x \in \mathbb{B}(\alpha\hat\theta^\top\theta^*, r_2\|\hat\theta\|) \cup \mathbb{B}(-\alpha\hat\theta^\top\theta^*, r_2\|\hat\theta\|)\}$.

Now, we are ready to calculate ECE. Specifically, given $\hat\theta$, For notation simplicity, we denote $\mathcal{A} = \mathbb{B}(\hat\theta^\top\theta^*, r_1\|\hat\theta\|) \cup \mathbb{B}(-\hat\theta^\top\theta^*, r_1\|\hat\theta\|)$ and $\mathcal{B} = \mathbb{B}(\alpha\cdot\hat\theta^\top\theta^*, r_2\|\hat\theta\|) \cup \mathbb{B}(-(\alpha\cdot\hat\theta^\top\theta^*, r_2\|\hat\theta\|)$. Meanwhile, for simplicity, we choose r_1, r_2 such that $\rho_1 = \rho_2 = \rho$. This is always manageable and there exists infinitely many choices, we only require $S_1 \cap S_2 = \emptyset$ for any $S_1 \neq S_2$, $S_1, S_2 \in \{\mathbb{B}(\hat\theta^\top\theta^*, r_1\|\hat\theta\|), \mathbb{B}(-\hat\theta^\top\theta^*, r_1\|\hat\theta\|), \mathbb{B}(\alpha\hat\theta^\top\theta^*, r_2\|\hat\theta\|), \mathbb{B}(-\alpha\hat\theta^\top\theta^*, r_2\|\hat\theta\|)\}$. In able to achieve $\rho_1 = \rho_2 = \rho$, it only depends on r_1/r_2. Apparently, there exists a threshold $\phi > 0$ such that if r_1 r_2 are both smaller than ϕ (one can choose r_1, r_2 as functions of σ with appropriate choosen σ), then $\mathcal{A} \cap \mathcal{B} = \emptyset$ can be achieved.

$$\begin{aligned}\mathbb{P}[y=1 \mid \hat\theta^\top x = v] &= \frac{\mathbb{P}(\hat\theta^\top x = v \mid y=1)}{\mathbb{P}(\hat\theta^\top x = v \mid y=1) + \mathbb{P}(\hat\theta^\top x = v \mid y=-1)} \\ &= \frac{1}{1+\exp\left(\frac{-2\hat\theta^\top\theta^*}{\sigma^2\|\hat\theta\|^2}\cdot v\right)}\mathbf{1}\{v \in \mathcal{A}\} + \frac{1}{1+\exp\left(\frac{2\alpha\hat\theta^\top\theta^*}{\sigma^2\|\hat\theta\|^2}\cdot v\right)}\mathbf{1}\{v \in \mathcal{B}\}\end{aligned}$$

D.5 PROOF OF THEOREM 1

D.5.1 TEMPERATURE SCALING ONLY

A simple reparameterization leads to:

$$\text{R-ECE} = \mathbb{E}_{v=\hat\theta^\top x}\left|\frac{\mathbf{1}\{v\in\mathcal{A}\}}{1+\exp\left(\frac{-2\hat\theta^\top\theta^*}{\sigma^2\|\hat\theta\|^2}\cdot v\right)} + \frac{\mathbf{1}\{v\in\mathcal{B}\}}{1+\exp\left(\frac{2\alpha\hat\theta^\top\theta^*}{\sigma^2\|\hat\theta\|^2}\cdot v\right)} - \frac{1}{e^{-2v/T}+1}\right|$$

The lower bound contains two parts. We choose the threshold ϕ mentioned previously small enough such that $I_3 > \max\{r_1, r_2\}$. This can be achieved because I_3 is independent of r_1, r_2.

Part I. When $v \in \mathbb{B}(\hat\theta^\top\theta^*, r_1\|\hat\theta\|) \subset \mathcal{A}$, and we know that $\mathcal{A} \cap \mathcal{B} = \emptyset$. Let us choose a threshold $\min\{I_1, I_3\}/\sigma^2 > \tau > 0$. Then for ***any*** $T > 0$, it must fall into one of the following three cases.

Case 1: T^{-1} and $\hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2)$ are far: $T^{-1} - \hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2) > \tau$, recall $v = \hat{\theta}^\top x$, then

$$\begin{aligned}
&\mathbb{E}_{x\in\mathbb{B}(\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)} - \frac{1}{1+e^{-2v/T}}\right| \\
\geq& \left[\frac{1}{1+\exp\left(-2T^{-1}(\hat{\theta}^\top\theta^* - r_1)\right)} - \frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}(\hat{\theta}^\top\theta^* + r_1)\right)}\right]\mathbb{P}(x\in\mathbb{B}(\theta^*, r_1))) \\
\geq& \frac{\beta}{2}\left[\frac{1}{1+\exp\left(-2(\hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2)+\tau)(\hat{\theta}^\top\theta^* - r_1)\right)} - \frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}(\hat{\theta}^\top\theta^* + r_1)\right)}\right] \\
\geq& \frac{\beta}{2}\left[\frac{1}{1+\exp\left(-2(\hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2)+\tau)(\hat{\theta}^\top\theta^* - r_1)\right)} - \frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}(\hat{\theta}^\top\theta^* + r_1)\right)}\right] \\
\geq& \frac{\beta}{2}\min_{c\in[I_1,I_2], d\in[I_3,I_4]}\left[\frac{1}{1+\exp\left(-2(c/\sigma^2+\tau)(d-r_1)\right)} - \frac{1}{1+\exp\left(-2c/\sigma^2(d+r_1)\right)}\right] := \beta_1
\end{aligned}$$

Case 2: T^{-1} and $\hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2)$ are far: $\hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2) - T^{-1} > \tau$,

$$\begin{aligned}
&\mathbb{E}_{x\in\mathbb{B}(\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)} - \frac{1}{1+e^{-2v/T}}\right| \\
\geq& \left[\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}(\hat{\theta}^\top\theta^* - r_1)\right)} - \frac{1}{1+\exp\left(-2T^{-1}(\hat{\theta}^\top\theta^* + r_1)\right)}\right]\cdot\frac{\beta}{2} \\
\geq& \left[\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}(\hat{\theta}^\top\theta^* - r_1)\right)} - \frac{1}{1+\exp\left(-2(\hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2)-\tau)(\hat{\theta}^\top\theta^* + r_1)\right)}\right]\cdot\frac{\beta}{2} \\
\geq& \frac{\beta}{2}\min_{c\in[I_1,I_2], d\in[I_3,I_4]}\left[\frac{1}{1+\exp\left(-2c/\sigma^2(d-r_1)\right)} - \frac{1}{1+\exp\left(-2(c/\sigma^2-\tau)(d+r_1)\right)}\right] := \beta_2
\end{aligned}$$

Case 3: When T^{-1} and $\hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2)$ are close: $|T^{-1} - \hat{\theta}^\top\theta^*/(\sigma^2\|\hat{\theta}\|^2)| \leq \tau$, then when $v \in \mathbb{B}(-\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|) \subset \mathcal{B}$. For small enough τ satisfying $\tau \leq 0.2(1-\alpha)I_1/\sigma^2$

$$\begin{aligned}
&\mathbb{E}_{v=\hat{\theta}^\top x\in\mathbb{B}(-\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)} - \frac{1}{1+e^{-2v/T}}\right| \\
\geq& \min_{a\in[\frac{-\alpha\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}, \frac{\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}+\tau]}\ \min_{v\in\mathbb{B}(-\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)}\frac{2v\exp(2va)}{(1+\exp(2av))^2}\left(\frac{2(1-\alpha)\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2} - \tau\right)\frac{1-\beta}{2} \\
\geq& \min_{a\in[-\frac{\alpha I_2}{\sigma^2}, \frac{I_2}{\sigma^2}+\tau]}\ \min_{v\in[\alpha I_3 - r_2 I_6, \alpha I_4 + r_2 I_6]}\frac{2v\exp(2va)}{(1+\exp(2av))^2}\left(1.8(1-\alpha)\frac{I_1}{\sigma^2}\right)\frac{1-\beta}{2} := \beta_3
\end{aligned}$$

Part III. Combining together, we have

$$\text{R-ECE} \geq \min\{\beta_1, \beta_2, \beta_3\}.$$

Finally, we take $r_1 \leq \min\{0.1, \tau\sigma^2/I_1\}I_3$, which makes sure $\beta_i > 0$ for all $i = 1, 2, 3$.

D.5.2 SELECTIVE CALIBRATION ONLY

We hope $\mathbb{E}[g(x) = 1] \geq \beta$. Let us first define $\mathcal{G} = \{\hat{\theta}^\top x : g(x) = 1\}$. Then, for ***any*** g, we have

$$\begin{aligned}
&\mathbb{P}[y=1 \mid \hat{\theta}^\top x = v, v \in \mathcal{G}] \\
&= \frac{\mathbb{P}(\hat{\theta}^\top x = v, v \in \mathcal{G} \mid y=1)}{\mathbb{P}(\hat{\theta}^\top x = v, v \in \mathcal{G} \mid y=1) + \mathbb{P}(\hat{\theta}^\top x = v, v \in \mathcal{G} \mid y=-1)} \\
&= \left(\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2} \cdot v\right)} \mathbf{1}\{v \in \mathcal{A}\} + \frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2} \cdot v\right)} \mathbf{1}\{v \in \mathcal{B}\} \right) \mathbf{1}\{v \in \mathcal{G}\} \\
&= \frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2} \cdot v\right)} \mathbf{1}\{v \in \mathcal{A} \cap \mathcal{G}\} + \frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2} \cdot v\right)} \mathbf{1}\{v \in \mathcal{B} \cap \mathcal{G}\}
\end{aligned}$$

Then, by choosing small enough σ, such that $I_1/\sigma^2 > 1$, the corresponding ECE is:

$$\begin{aligned}
\text{ECE}_S = \mathbb{E}_{v=\hat{\theta}^\top x \mid \hat{\theta}^\top x \in \mathcal{G}} \Bigg| & \frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2} \cdot v\right)} \mathbf{1}\{v \in \mathcal{A} \cap \mathcal{G}\} \\
& + \frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2} \cdot v\right)} \mathbf{1}\{v \in \mathcal{B} \cap \mathcal{G}\} - \frac{1}{e^{-2v}+1} \Bigg| \\
\geq & \ \mathbb{E}_{v=\hat{\theta}^\top x \mid \hat{\theta}^\top x \in \mathcal{G}} \left| \frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2} \cdot v\right)} \mathbf{1}\{v \in \mathcal{A} \cap \mathcal{G}\} - \frac{1}{e^{-2v}+1} \right| \\
& + \mathbb{E}_{v=\hat{\theta}^\top x \mid \hat{\theta}^\top x \in \mathcal{G}} \left| \frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2} \cdot v\right)} \mathbf{1}\{v \in \mathcal{B} \cap \mathcal{G}\} - \frac{1}{e^{-2v}+1} \right| \\
\geq & \ \lambda_1 \mathbb{P}(v \in \mathcal{A} \cap \mathcal{G} | v \in \mathcal{G}) + \lambda_2 \mathbb{P}(v \in \mathcal{B} \cap \mathcal{G} | v \in \mathcal{G}).
\end{aligned}$$

Since $\mathbb{P}(v \in \mathcal{A} \cap \mathcal{G} | v \in \mathcal{G}) + \mathbb{P}(v \in \mathcal{B} \cap \mathcal{G} | v \in \mathcal{G}) = 1$, it is not hard to verify that

$$\text{S-ECE} \geq \min\{\lambda_1, \lambda_2\}$$

where

$$\lambda_1 = \min_{a \in [1, \frac{I_2}{\sigma^2}]} \min_{v \in \mathcal{A}} \frac{2v \exp(2va)}{(1+\exp(2av))^2} \left| \frac{I_1}{\sigma^2} - 1 \right|$$

$$\lambda_2 = \min_{a \in [-\frac{\alpha I_2}{\sigma^2}, 1]} \min_{v \in \mathcal{B}} \frac{2v \exp(2va)}{(1+\exp(2av))^2} \left| \frac{\alpha I_1}{\sigma^2} - 1 \right|.$$

D.5.3 SELECTIVE RE-CALIBRATION

We choose $\mathcal{G} = \mathcal{B}$ and set $T^{-1} = \frac{\hat{\theta}^\top \theta^*}{\sigma^2 \|\hat{\theta}\|^2}$, then SR-ECE $= 0$. Thus, there exists appropriate choice of g and T such that

$$\text{SR-ECE}(g, T) = 0.$$

D.6 PROOF OF THEOREM 2

Usually, β is much larger than $1-\beta$, for example, $\beta = 90\%$. In this section, we impose the following assumption.

Assumption 4. *The selector g will retain most of the probabilty mass in the sense that*

$$\beta > 2(1-\beta).$$

Let us denote $\xi = \beta/2 - (1-\beta)$ and ξ is a positive constant. First, we have the following claim.

Claim 5. *Under Assumption 4, if we further have*

$$\min_{v\in\mathbb{B}(\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)}-\frac{1}{e^{-2v}+1}\right|$$
$$>\max_{v\in\mathbb{B}(-\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)}-\frac{1}{e^{-2v}+1}\right|,$$

then for $g_1 = \operatorname{argmin}_{g:\mathbb{E}[g(x)=1]\geq\beta}$ *S-ECE, we have that* $\mathbb{E}_{x\in\mathbb{B}(-\alpha\theta^*, r_2)}[g_1(x)=1] = \mathbb{P}(x\in\mathbb{B}(-\alpha\theta^*, r_2))$.

Proof. The proof is straightforward. We denote $O=\{x : x\in\mathbb{B}(-\alpha\theta^*, r_2),\ g(x)=0\}$. We will prove that $\mathbb{P}(x\in O)=0$.

If not, let us denote $\mathcal{P}=\mathbb{P}(x\in O)>0$. Since we know $\beta>2(1-\beta)$, which means even if we "throw away" all the probability mass $1-\beta$ by only setting points in $\mathbb{B}(\theta^*, r_1)$ with g value equals to 0, there will still be other remaining probability mass retained in $\mathbb{B}(\theta^*, r_1)$ with g value equals to 1. Then, there exists g_2 such that $g_2(x)=1$ for all $x\in\mathbb{B}(-\alpha\theta^*, r_2)$ and leads to $\mathbb{P}(x\in\mathbb{B}(-\alpha\theta^*, r_2), g_2(x)=1)=\mathbb{P}(x\in\mathbb{B}(-\alpha\theta^*, r_2), g_1(x)=1)+\xi$ (enabled by the fact $\beta>2(1-\beta)$) and $\mathbb{P}(g_1(x)=1)=\mathbb{P}(g_2(x)=1)$ for $x\in\mathbb{B}(\theta^*, r_1)\cup\mathbb{B}(-\alpha\theta^*, r_2)$. Since

$$\min_{v\in\mathbb{B}(\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)}-\frac{1}{e^{-2v}+1}\right|$$
$$>\max_{v\in\mathbb{B}(-\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)}-\frac{1}{e^{-2v}+1}\right|,$$

which means "throwing away" points in $\mathbb{B}(\alpha\theta^*, r_2)$ can more effectively lower the calibration error and we must have

$$\text{S-ECE}(g_2)<\text{S-ECE}(g_1).$$

□

Next, we state how to set the parameters such that the condition in Claim 5 holds. As long as we choose σ, r_1, r_2 small enough, such that

$$\frac{1}{1+\exp(-2I_1/\sigma^2(I_4+r_1I_6))}-\frac{1}{1+\exp(-2I_1(I_3-r_1I_6))}$$
$$<\frac{1}{1+\exp(-2(I_4+r_2I_6))}-\frac{1}{1+\exp(2\alpha I_2/\sigma^2(I_4+r_2I_6))}$$

then,

$$\min_{v\in\mathbb{B}(\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)}-\frac{1}{e^{-2v}+1}\right|$$
$$>\max_{v\in\mathbb{B}(-\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)}\left|\frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)}-\frac{1}{e^{-2v}+1}\right|,$$

Then, following similar derivation in Section D.5.1, we can prove with suitably chosen parameters r_1, r_2, σ, $\text{ECE}^{S\to T}>0$.

Lastly, let us further prove $\text{ECE}^{T\to S}>0$. We choose r_1 and r_2 small enough such that $v>0$ for all $v\in\mathbb{B}(\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)\cup\mathbb{B}(\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)$ and $v<0$ for all $v\in\mathbb{B}(-\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)\cup\mathbb{B}(-\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)$.

For $1/T \in [\frac{-\alpha\hat{\theta}^T\theta^*}{\sigma^2\|\hat{\theta}\|}, \frac{\hat{\theta}^T\theta^*}{\sigma^2\|\hat{\theta}\|}]$, we can calculate the derivative for R-ECE as the following:

$$\begin{aligned}\text{R-ECE} = & \mathbb{E}_{v\in\mathbb{B}(\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)}\left[\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)} - \frac{1}{e^{-2v/T}+1}\right] \\ & + \mathbb{E}_{v\in\mathbb{B}(-\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)}\left[-\frac{1}{1+\exp\left(\frac{-2\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)} + \frac{1}{e^{-2v/T}+1}\right] \\ & + \mathbb{E}_{v\in\mathbb{B}(\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)}\left[-\frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)} + \frac{1}{e^{-2v/T}+1}\right] \\ & + \mathbb{E}_{v\in\mathbb{B}(-\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)}\left[\frac{1}{1+\exp\left(\frac{2\alpha\hat{\theta}^\top\theta^*}{\sigma^2\|\hat{\theta}\|^2}\cdot v\right)} - \frac{1}{e^{-2v/T}+1}\right].\end{aligned}$$

Next, we take a derivative over $x = 1/T$ for $x \in [\frac{-\alpha\hat{\theta}^T\theta^*}{\sigma^2\|\hat{\theta}\|}, \frac{\hat{\theta}^T\theta^*}{\sigma^2\|\hat{\theta}\|}]$, which leads to

$$\frac{d\text{R-ECE}}{dx} = -2\mathbb{E}_{v\in\mathbb{B}(\hat{\theta}^\top\theta^*, r_1\|\hat{\theta}\|)}\left[\frac{2ve^{2vx}}{(e^{2vx}+1)^2}\right] + 2\mathbb{E}_{v\in\mathbb{B}(\alpha\hat{\theta}^\top\theta^*, r_2\|\hat{\theta}\|)}\left[\frac{2ve^{2vx}}{(e^{2vx}+1)^2}\right]$$

Consider the two values

$$\frac{2ve^{2vx}}{(e^{2vx}+1)^2}, \quad \frac{2\alpha ve^{2\alpha vx}}{(e^{2\alpha vx}+1)^2},$$

the ratio

$$\frac{2\alpha ve^{2\alpha vx}}{(e^{2\alpha vx}+1)^2}/[\frac{2ve^{2vx}}{(e^{2vx}+1)^2}] \to_{v\to 0} 2\alpha.$$

That means if we take suitably small r_1, r_2 and let $\sigma \in [c_1, c_2]$ with appropriately chosen c_1, c_2

$$\frac{d\text{R-ECE}}{dx}\Big|_{x=\frac{\hat{\theta}^T\theta^*}{\sigma^2\|\hat{\theta}\|}} < 0.$$

Thus, we know the best choice of $1/T$ should not be equal to $\frac{\hat{\theta}^T\theta^*}{\sigma^2\|\hat{\theta}\|}$. Then, notice $\beta > 2(1-\beta)$, which means the probability mass in $\mathbb{B}(\hat{\theta}^T\theta^*, r_1\|\hat{\theta}\|)$ cannot be all be "thrown away"; following similar derivation in Section D.5.1, we can prove with suitably chosen parameters r_1, r_2, σ, ECE$^{T\to S} > 0$.