A $\mathcal{H}$-CONSISTENCY BOUND PROOF

A.1 CALIBRATION GAP FOR REJECTION LOSS

The following gives the expression of the calibration gap $\Delta\mathcal{C}_{\ell_2}$.

Lemma 2. *The Bayes solution r^* for the rejection loss can be expressed for all $x \in \mathcal{X}$ by $r^*(x) = \eta(x) - (1-c)$. The calibration gap for the rejection loss is given for any $r \in \mathcal{R}_{\text{all}}$ and $x \in \mathcal{X}$ by*

$$\Delta\mathcal{C}_{\ell_2}(r,x) = |\eta(x) - (1-c)|\,\mathbb{I}_{r(x)r^*(x)\leq 0}\,.$$

Proof. For any $r \in \mathcal{R}_{\text{all}}$ and $x \in \mathcal{X}$, we can write

$$\begin{aligned}
\mathcal{C}_{\ell_2}(r,x) &= \eta(x)\ell_2(r,x,+1)\\
&\quad + [1-\eta(x)]\ell_2(r,x,-1)\\
&= \eta(x)\left[\mathbb{I}_{+1=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{r(x)\leq 0}\right]\\
&\quad + [1-\eta(x)]\left[\mathbb{I}_{-1=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{r(x)\leq 0}\right]\\
&= c\,\mathbb{I}_{r(x)\leq 0} + [1-\eta(x)]\,\mathbb{I}_{r(x)>0}\,.
\end{aligned}$$

For the optimal $\mathcal{C}^*_{\ell_2}$, we would always pick the lower of c or $1-\eta(x)$, which gives: $\mathcal{C}^*_{\ell_2}(x) = \min\{c, 1-\eta(x)\}$. The corresponding Bayes solution r^* can be defined by $r^*(x) = \eta(x) - (1-c)$. Thus, the calibration gap is given by

$$\begin{aligned}
\Delta\mathcal{C}_{\ell_2}(r,x) &= c\,\mathbb{I}_{r(x)\leq 0} + [1-\eta(x)]\,\mathbb{I}_{r(x)>0}\\
&\quad - \min\{c, 1-\eta(x)\}.
\end{aligned}$$

If $r(x)$ correctly chooses the lower of the two, we have $r(x)r^*(x) > 0$ and then $\Delta\mathcal{C}_{\ell_2} = 0$. Otherwise,

$$\Delta\mathcal{C}_{\ell_2}(r,x) = \begin{cases} c - (1-\eta(x)) & \text{if } r(x) \leq 0\\ (1-\eta(x)) - c & \text{otherwise}\end{cases}.$$

Thus, for all $x \in \mathcal{X}$, we have $\Delta\mathcal{C}_{\ell_2}(r,x) = |\eta(x) - (1-c)|\,\mathbb{I}_{r(x)r^*(x)\leq 0}$. This completes the proof. □

A.2 CALIBRATION GAP FOR SURROGATE LOSS

Here, we analyze the calibration gap for the surrogate loss.

Lemma 3. *Let $I_\eta(x)$ be defined by $I_\eta(x) = \eta(x)e^{-\frac{\alpha}{2}} + (1-\eta(x))e^{\frac{\alpha}{2}}$ and define γ by $\gamma = \frac{\alpha}{\alpha+2\beta}$. Then, the calibration gap for the surrogate loss is given for any $r \in \mathcal{R}_{\text{all}}$ and $x \in \mathcal{X}$ by*

$$\Delta\mathcal{C}_{\ell_1}(r,x) = e^{\frac{\alpha}{2}r(x)}I_\eta(x) + ce^{-\beta r(x)} - \frac{1}{1-\gamma}\left(\frac{2\beta c}{\alpha}\right)^{\gamma} I_\eta(x)^{1-\gamma}.$$

Proof. By definition, the calibration function for ℓ_1 can be expressed for all $x \in \mathcal{X}$ by

$$\begin{aligned}
\mathcal{C}_{\ell_1}(r,x) &= \eta(x)\ell_1(r,x,+1)\\
&\quad + [1-\eta(x)]\ell_1(r,x,-1)\\
&= \eta(x)\left[e^{\frac{\alpha}{2}[r(x)-1]} + ce^{-\beta r(x)}\right]\\
&\quad + [1-\eta(x)]\left[e^{\frac{\alpha}{2}[r(x)+1]} + ce^{-\beta r(x)}\right]\\
&= e^{\frac{\alpha}{2}r(x)}I_\eta(x) + ce^{-\beta r(x)}.
\end{aligned}$$

Since the exponential function is convex, $\Delta\mathcal{C}_{\ell_1}(r,x)$ is a convex function of $r(x)$. Thus, for $r \in \mathcal{R}_{\text{all}}$, we obtain the minimum $r_0(x)$ by differentiating with respect to $r(x)$ and setting to 0:

$$\begin{aligned}
&\frac{\alpha}{2}e^{\frac{\alpha}{2}r(x)}I_\eta(x) - \beta ce^{-\beta r(x)} = 0\\
\Leftrightarrow\ & r_0(x) = \log\left[\left(\frac{2\beta c}{\alpha I_\eta(x)}\right)^{\frac{2}{2\beta+\alpha}}\right].
\end{aligned}$$

Plugging in this expression in $\mathcal{C}_{\ell_1}$ gives the corresponding minimal calibration $\mathcal{C}^*_{\ell_1}(x)$: $\mathcal{C}^*_{\ell_1}(x) = \left[\left(\frac{2\beta c}{\alpha}\right)^{\gamma}\right] I_\eta(x)^{1-\gamma}\left(\frac{1}{1-\gamma}\right)$. This completes the proof. □

A.3 $\mathcal{H}$-CONSISTENCY BOUND

In this section, we prove our main result. The following will provide a key tool to derive our result.

Proposition 4. *Assume that there exists a convex function $\Psi\colon \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) = 0$ such that the following holds for all $r \in \mathcal{R}_{\text{all}}$ and $x \in \mathcal{X}$: $\Psi\big(|\eta(x) - (1-c)|\,\mathbb{I}_{r(x)r^*(x)\leq 0}\big) \leq \Delta\mathcal{C}_{\ell_1}(0, x)$. Let $\bar{I}_c$ be defined by $\bar{I}_c = ce^{\frac{\alpha}{2}} + (1-c)e^{-\frac{\alpha}{2}}$ and assume that $\frac{2\beta c}{\alpha} = \bar{I}_c$. Then, for any $r \in \mathcal{R}_{\text{all}}$:*

$$\Psi\big(R_{\ell_2}(r) - R^*_{\ell_2}\big) \leq R_{\ell_1}(r) - R^*_{\ell_1}. \tag{3}$$

Proof. We will show that the following holds: $\inf_{r(x)r^*(x)\leq 0} \Delta\mathcal{C}_{\ell_1}(r, x) = \Delta\mathcal{C}_{\ell_1}(0, x)$. The result then follows by Theorem 1 and Lemma 2. Since we have $\frac{2\beta c}{\alpha} = \bar{I}_c$, the following equivalence holds:

$$\begin{aligned}
r_0(x) > 0 &\Leftrightarrow \frac{2\beta c}{\alpha I_\eta(x)} > 1 \\
&\Leftrightarrow I_\eta(x) < \bar{I}_c \\
&\Leftrightarrow \eta(x) > \frac{e^{\frac{\alpha}{2}} - \bar{I}_c}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}} \\
&\Leftrightarrow \eta(x) > \frac{(1-c)e^{\frac{\alpha}{2}} - (1-c)e^{-\frac{\alpha}{2}}}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}} \\
&\Leftrightarrow r^*(x) > 0.
\end{aligned}$$

This implies $\inf_{r(x)r^*(x)\leq 0} \mathcal{C}_{\ell_1}(r, x) = \inf_{r(x)r_0(x)\leq 0} \mathcal{C}_{\ell_1}(r, x)$. Now, since $r_0(x)$ is the unique minimizer of the strictly convex function $\mathcal{C}_{\ell_1}(r, x)$ of $r(x)$, then, as a function of $r(x)$, $\mathcal{C}_{\ell_1}(r, x)$ is decreasing from $-\infty$ to $r_0(x)$ and increasing from there to $+\infty$. Thus, if $r_0(x) > 0$, the infimum of $\mathcal{C}_{\ell_1}(r, x)$ over $r(x) \leq 0$ is reached for $r(x) = 0$. Similarly, if $r_0(x) < 0$, the infimum of $\mathcal{C}_{\ell_1}(r, x)$ over $r(x) \geq 0$ is reached for $r(x) = 0$. This shows that $\inf_{r(x)r_0(x)\leq 0} \mathcal{C}_{\ell_1}(r, x) = \mathcal{C}_{\ell_1}(0, x)$, and completes the proof. □

The proof of our main result makes use of the following identity, which is a refinement of Bernoulli's inequality. The result could be of independent interest in other contexts, we give a concise proof below.

Lemma 6 (Bernoulli-type inequality)**.** *The following identity holds for all $x, r \in (0, 1)$,*

$$(1+x)^r \leq 1 + rx + \frac{r(r-1)x^2}{4}.$$

Proof. Let $f_r(x) = (1+x)^r - \left(1 + rx + \frac{r(r-1)x^2}{4}\right)$. We will show that $f_r(x) \leq 0$ for all $x, r \in (0, 1)$. We have $f'_r(x) = r(1+x)^{r-1} - \left(r + \frac{r(r-1)x}{2}\right)$, and $f'_r(0) = 0$. To see that $f'_r(1) \leq 0$, observe $r2^{r-1} - \left(r + \frac{r(r-1)}{2}\right) \leq 0 \Leftrightarrow 2^{r-1} - \frac{(r-1)}{2} \leq 1$. The left-hand side of the last inequality is a convex function of r, and equal to 1 when $r = 0$ or $r = 1$. Thus, the left-hand side is less than or equal 1 for $r \in (0, 1)$, giving $f'_r(1) \leq 0$. Since $f'_r(x)$ is a convex function of x, with $f'_r(0) \leq 0$ and $f'_r(1) \leq 0$, then $f'_r(x) \leq 0$ for all $x \in (0, 1)$, which shows f_r is decreasing. Then, since $f_r(0) = 0$, $f_r(x) \leq 0$ for all $x, r \in (0, 1)$. □

The following is our main result; it relates the surrogate excess error to that of the rejection loss.

Theorem 5. *Let $\alpha, \beta > 0$ be such that $\frac{2\beta c}{\alpha} = \bar{I}_c$, where $\bar{I}_c = ce^{\frac{\alpha}{2}} + (1-c)e^{-\frac{\alpha}{2}}$. Then, the following inequality holds for any $r \in \mathcal{R}_{\text{all}}$:*

$$R_{\ell_2}(r) - R^*_{\ell_2} \leq \frac{2}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}} \sqrt{\frac{(c + \bar{I}_c)\bar{I}_c}{c}\big(R_{\ell_1}(r) - R^*_{\ell_1}\big)}.$$

Proof. Using the expression of $\Delta\mathcal{C}_{\ell_1}$ given by Lemma 3, we can write

$$\Delta\mathcal{C}_{\ell_1}(0,x) = I_\eta(x) + c - \frac{1}{1-\gamma}\left(\frac{2\beta c}{\alpha}\right)^\gamma I_\eta(x)^{1-\gamma}$$
$$= I_\eta(x) + c - \left(\bar{I}_c + c\right)\left(\frac{I_\eta(x)}{\bar{I}_c}\right)^{1-\gamma}.$$

We can express this formula in terms of $u(x) = \eta(x) - (1-c)$, using $I_\eta(x) = J_u(x) + \bar{I}_c$, with $J_u(x) = \left[e^{-\frac{\alpha}{2}} - e^{\frac{\alpha}{2}}\right]u(x)$:

$$\begin{aligned}
&\Delta\mathcal{C}_{\ell_1}(0,x)\\
&= J_u(x) + \bar{I}_c + c - \left(\bar{I}_c + c\right)\left[1 + \frac{J_u(x)}{\bar{I}_c}\right]^{1-\gamma}\\
&\geq \frac{\bar{I}_c}{c+\bar{I}_c}\frac{c}{c+\bar{I}_c}\frac{c+\bar{I}_c}{4}\left[\frac{J_u(x)}{\bar{I}_c}\right]^2\\
&= \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[\frac{J_u(x)}{\bar{I}_c}\right]^2.
\end{aligned}$$

where we used Lemma 6. The function $\Psi(u)$ defined by this expression verifies the condition of Proposition 4 and therefore we have $\Psi\left(R_{\ell_2}(h) - R^*_{\ell_2}\right) \leq R_{\ell_1}(h) - R^*_{\ell_1}$. An explicit upper-bound on $R_{\ell_2}(h) - R^*_{\ell_2}$ can be written in terms of Ψ^{-1}: $R_{\ell_2}(h) - R^*_{\ell_2} \leq \Psi^{-1}\left(R_{\ell_1}(h) - R^*_{\ell_1}\right)$. To derive the expression of Ψ^{-1}, we write $z = \Psi(u)$, that is:

$$4\frac{c+\bar{I}_c}{c\bar{I}_c}z = \left[\frac{u(x)}{\bar{I}_c}\right]^2\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]^2$$
$$\Leftrightarrow |u| = \frac{2}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}}\sqrt{\frac{(c+\bar{I}_c)\bar{I}_c}{c}z}.$$

Thus, we have, for all $r \in \mathcal{R}_{\text{all}}$, $R_{\ell_2}(r) - R^*_{\ell_2} \leq \frac{2}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}}\sqrt{\frac{(c+\bar{I}_c)\bar{I}_c}{c}\left(R_{\ell_1}(r) - R^*_{\ell_1}\right)}$. □

B EXPERIMENTAL DETAILS

B.1 DECONTEXTUALIZATION

In this section, we report the detailed results for our experiments on the decontextualization task. Table 2 presents the mean and standard deviation of the precision and coverage of various baselines over 4 cross-validation splits and Table 1 in Section 7.3 provides detailed results of the surrogate loss.

Table 2: Precision vs. Coverage for various baselines on decontextualization, with theoretical limit.

	MAXPROB		CROSS-ENTROPY		THEORETICAL LIMIT	
Target precision	precision	coverage	precision	coverage	precision	coverage
0.90	0.899 ± 0.002	0.907 ± 0.017	0.903 ± 0.016	0.968 ± 0.045	0.90	0.989 ± 0.001
0.92	0.924 ± 0.001	0.672 ± 0.052	0.930 ± 0.021	0.771 ± 0.146	0.92	0.967 ± 0.001
0.93	0.934 ± 0.025	0.552 ± 0.069	0.939 ± 0.015	0.677 ± 0.102	0.93	0.957 ± 0.001
0.94	0.938 ± 0.022	0.467 ± 0.035	0.949 ± 0.012	0.644 ± 0.103	0.94	0.950 ± 0.001
0.95	0.942 ± 0.023	0.405 ± 0.030	0.965 ± 0.015	0.509 ± 0.143	0.95	0.936 ± 0.001
0.96	0.959 ± 0.022	0.321 ± 0.041	0.976 ± 0.006	0.364 ± 0.096	0.96	0.927 ± 0.001
0.97	0.972 ± 0.018	0.225 ± 0.012	0.980 ± 0.008	0.330 ± 0.086	0.97	0.917 ± 0.001
0.98	0.972 ± 0.018	0.198 ± 0.017	0.981 ± 0.013	0.298 ± 0.069	0.98	0.908 ± 0.001
0.99	0.983 ± 0.013	0.168 ± 0.015	0.986 ± 0.015	0.150 ± 0.059	0.99	0.898 ± 0.001

B.2 IMAGE CLASSIFICATION

In this section, we provide details of our experiments on Fashion-MNIST, a fashion image dataset, and KMNIST, a cursive Japanese letter dataset. Both are perfectly balanced between their 10 classes. In both cases, we use a 5-layer fully-connected neural network to train a predictor with half of the training data. The remaining half is reserved for the rejector. Training the rejector is a binary classification task: for pairs (x, y) occuring in the usual dataset, we construct another dataset $((x, f_p(x)), \mathbb{I}_{f(x)=y})$, where f is the predictor and $f_p(x)$ is the probability that f assigns to its prediction on x. In our experiments, we observe that it is important to append $f_p(x)$ as a feature to x. Note that constructing this binary classification dataset does not require manual annotation. For Fashion-MNIST, our predictor is trained to 85.3% accuracy on its test set, and for KMNIST, our predictor is trained to 79.1% accuracy on its test set. While it is possible to improve the performance of these predictors, this is not our focus. We are focused on a rejection task given some fixed predictor.

Next, we detail the methods for rejection.

Maxprob. Similar to the decontextualization experiment, we fit thresholds on the scores assigned by the predictor. Since this method is deterministic (and the error bars here are over rejector training runs), there are no error bars to report.

Cross-entropy loss. We train another 5-layer neural network on the constructed binary classification dataset using the cross-entropy loss. Similar to the decontextualization experiment, thresholds are fitted on the scores of this neural netowrk.

Rejection loss. We train a second 5-layer neural network on the constructed binary classification dataset using our proposed surrogate rejection loss. For Fashion-MNIST, c is varied in $\{0.05, 0.1, 0.2, 0.3, 0.5\}$. For KMNIST, c is varied in $\{0.025, 0.05, 0.1, 0.15\}$. Each point on the plot represents a model trained with a different value of c. We set α in the surrogate rejection loss function to 3.5.

Cost-sensitive loss. We train a third 5-layer neural network on the constructed binary classification dataset using the cross-entropy loss, but with the positive class reweighted by $c/(1-c)$. For Fashion-MNIST, c is varied in $\{0.05, 0.1, 0.2, 0.3, 0.5\}$. For KMNIST, c is varied in $\{0.03, 0.05, 0.1, 0.2\}$. Each point on the plot represents a model trained with a different value of c.

For all methods, we use the Adam optimizer (Kingma and Ba, 2014), and tune the learning rate in $[1e-4, 1e-7]$ and number of epochs in $[20, 100]$.

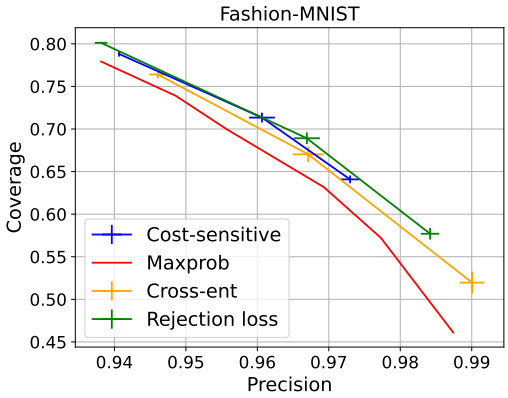

Figure 6: Precision vs. Coverage on Fashion-MNIST. Standard deviations for both precision and coverage are from 4 different training runs.

The precision vs. coverage graph for KMNIST is reported in Figure 5 in Section 7.4 and the precision vs. coverage graph for Fashion-MNIST is reported in Figure 6. We do not plot the theoretical limit since no method is near it in this setting. In both cases, we generally observe that the rejection loss lies above the baselines. It is likely that the predictor in this setting is much better calibrated than a large language model, and thus Maxprob is a much stronger baseline with not as much room for improvement as on decontextualization. We also note that it may also be possible to improve the performance of our method by tuning α.

C Comparison with cost-sensitive classification

It is worth pointing out that minimizing the induced rejection loss is equivalent to minimizing a cost-sensitive classification loss (Elkan, 2001; Steinwart, 2007; Scott, 2012; Charoenphakdee et al., 2021), since by using the decomposition $\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)>0} = (1-c)\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)>0}$ and $c\,\mathbb{I}_{r(x)\leq 0} = c\,\mathbb{I}_{a=+1}\,\mathbb{I}_{r(x)\leq 0} + c\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)\leq 0}$, the loss (2) can be rewritten as

$$\begin{aligned}\ell(r,x,a) &= \mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{r(x)\leq 0}\\ &= (1-c)\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{a=+1}\,\mathbb{I}_{r(x)\leq 0} + c\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)\leq 0}\\ &= (1-c)\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{a=+1}\,\mathbb{I}_{r(x)\leq 0} + c\,\mathbb{I}_{a=-1},\end{aligned}$$

where in the last step we use the fact that $c\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)\leq 0} + c\,\mathbb{I}_{a=-1}\,\mathbb{I}_{r(x)>0} = c\,\mathbb{I}_{a=-1}$. In light of this expression, since the last term $c\,\mathbb{I}_{a=-1}$ does not depend on r, if $x \mapsto \phi(-x)$ is a convex function upper-bounding $\mathbb{I}_{x\leq 0}$, then, ℓ_ϕ defined as follows for any $r \in \mathcal{R}$ and $(x,a) \in \mathcal{X}\times\{-1,+1\}$, is a natural surrogate loss for ℓ:

$$\ell_\phi(r,x,a) = (1-c)\,\mathbb{I}_{a=-1}\,\phi(r(x)) + c\,\mathbb{I}_{a=+1}\,\phi(-r(x)).$$

We will refer to ℓ_ϕ as cost-sensitive surrogate losses for the induced rejection loss. However, this cost-sensitive approach suffers from several issues: (i) There is a lack of any $\mathcal{H}$-consistency bound guarantees for cost-sensitive surrogate losses with respect to the induced rejection loss. Conversely, our theoretical analysis can potentially extend to an $\mathcal{H}$-consistent surrogate loss function for cost-sensitive classification. This would provide a theoretically justified algorithm for that context. Our novel contribution lies in introducing a loss function for the induced rejection loss backed by strong $\mathcal{H}$-consistency bounds; (ii) It has been shown in (Cao et al., 2022) that the cost-sensitive approach (Charoenphakdee et al., 2021) can not produce the state-of-the-art performance in the learning with rejection framework, which motivates us to propose a new theoretically guaranteed surrogate loss in our rejection scenario; (iii) As shown in (Charoenphakdee et al., 2021), the cost-sensitive approach equivalently solves n one-versus-all binary classification problems, where n is the number of classes. Therefore, when the size of the sub-sample containing some of the classes is relatively small, the one-versus-all binary classification problem may face challenges due to insufficient data or increased risk of overfitting. This issue stands out for the decontextualization task, where the samples corresponding to $a = -1$ are much fewer than those corresponding to $a = +1$; (iv) Our empirical results on the benchmark datasets show that the cost-sensitive approach is inferior to our proposed surrogate loss function, which substantiate the effectiveness of our approach.

D $\mathcal{H}$-CONSISTENCY BOUNDS BEYOND $\mathcal{H}_{\text{all}}$ AND PROOF

Here, we will show that our surrogate losses benefit from $\mathcal{R}$-consistency bounds with the hypothesis set $\mathcal{R}$ extending beyond the family of all measurable functions $\mathcal{R}_{\text{all}}$. Without loss of generality, we consider $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_p \leq 1\}$. Let $p, q \in [1, +\infty]$ be conjugate numbers such that $\frac{1}{p} + \frac{1}{q} = 1$. We will consider *bounded* hypothesis sets $\mathcal{R}$, that is, there exists a function $\overline{r}: \mathcal{X} \to \mathbb{R}_+$ such that for all $r \in \mathcal{R}$ and $x \in \mathcal{X}$, $|r(x)| \leq \overline{r}(x)$, and all values in $[-\overline{r}(x), \overline{r}(x)]$ can be reached. As shown by Awasthi et al. (2022), for the family of linear models $\mathcal{R}_{\text{lin}} = \{x \mapsto w \cdot x + b \mid \|w\|_q \leq W, |b| \leq B\}$ and one-hidden-layer ReLU networks $\mathcal{R}_{\text{NN}} = \{x \mapsto \sum_{j=1}^n u_j (w_j \cdot x + b_j)_+ \mid \|u\|_1 \leq \Lambda, \|w_j\|_q \leq W, |b_j| \leq B\}$, where $(\cdot)_+ = \max(\cdot, 0)$, we have $\overline{r}(x) = W\|x\|_p + B$ and $\overline{r}(x) = \Lambda W \|x\|_p + \Lambda B$ respectively.

D.1 MAIN RESULT

In this section, we present our main result on $\mathcal{R}$-consistency bounds with bounded hypothesis sets $\mathcal{R}$ (Theorem 7), including $\mathcal{R}_{\text{lin}}$ and $\mathcal{R}_{\text{NN}}$ considered in (Awasthi et al., 2022) as special cases (Corollary 8). The proofs are presented in Appendix D.4.

Theorem 7. *Assume that $\mathcal{R}$ is bounded with function $\overline{r}: \mathcal{X} \to \mathbb{R}$. Let $\alpha, \beta > 0$ be such that $\frac{2\beta c}{\alpha} = \bar{I}_c$, where $\bar{I}_c = ce^{\frac{\alpha}{2}} + (1-c)e^{-\frac{\alpha}{2}}$. Then, the following inequality holds for any $r \in \mathcal{R}$:*

$$R_{\ell_2}(r) - R^*_{\ell_2,\mathcal{R}} + \mathcal{M}_{\ell_2,\mathcal{R}} \leq \Gamma\Big(R_{\ell_1}(r) - R^*_{\ell_1,\mathcal{R}} + \mathcal{M}_{\ell_1,\mathcal{R}}\Big), \tag{4}$$

where $\Gamma(z) = \begin{cases} \frac{2}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}} \sqrt{\frac{(c+\bar{I}_c)\bar{I}_c}{c}} z & 0 \leq z \leq \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]^2 \\ \frac{4(c+\bar{I}_c)}{c\left[1-e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]} z & \text{otherwise.} \end{cases}$

Corollary 8. *Let $\mathcal{R} = \mathcal{R}_{\text{lin}}$ or $\mathcal{R}_{\text{NN}}$. Let $\alpha, \beta > 0$ be such that $\frac{2\beta c}{\alpha} = \bar{I}_c$, where $\bar{I}_c = ce^{\frac{\alpha}{2}} + (1-c)e^{-\frac{\alpha}{2}}$. Then, the following inequality holds for any $r \in \mathcal{R}$:*

$$R_{\ell_2}(r) - R^*_{\ell_2,\mathcal{R}} + \mathcal{M}_{\ell_2,\mathcal{R}} \leq \Gamma\Big(R_{\ell_1}(r) - R^*_{\ell_1,\mathcal{R}} + \mathcal{M}_{\ell_1,\mathcal{R}}\Big), \tag{5}$$

where $\Gamma(z) = \begin{cases} \frac{2}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}} \sqrt{\frac{(c+\bar{I}_c)\bar{I}_c}{c}} z & 0 \leq z \leq \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)B}{2c}}\right]^2 \\ \frac{4(c+\bar{I}_c)}{c\left[1-e^{-\frac{\alpha(\bar{I}_c+c)B}{2c}}\right]\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]} z & \text{otherwise} \end{cases}$ *and B is replaced by ΛB for $\mathcal{R} = \mathcal{R}_{\text{NN}}$.*

D.2 CALIBRATION GAP FOR REJECTION LOSS

We first extend Lemma 2 to any hypothesis set $\mathcal{R}$ that is *regular for rejection*.

Definition 9. *We say that a hypothesis set $\mathcal{R}$ is* regular for rejection *if for all $x \in \mathcal{X}$, there exist $r_+, r_- \in \mathcal{R}$ such that $r_+(x) > 0$ and $r_+(x) \leq 0$.*

It is clear that all bounded hypothesis sets including $\mathcal{R}_{\text{lin}}$ and $\mathcal{R}_{\text{NN}}$ are regular for rejection. The following gives the expression of the calibration gap $\Delta\mathcal{C}_{\ell_2}$ for all hypothesis sets $\mathcal{R}$ that are regular for rejection. The proof is nearly identical to Lemma 2.

Lemma 10. *Assume that $\mathcal{R}$ is regular for rejection. The best-in-class solution r^* for the rejection loss can be expressed for all $x \in \mathcal{X}$ by $r^*(x) = \eta(x) - (1-c)$. The calibration gap for the rejection loss is given for any $r \in \mathcal{R}$ and $x \in \mathcal{X}$ by*

$$\Delta\mathcal{C}_{\ell_2}(r, x) = |\eta(x) - (1-c)|\, \mathbb{I}_{r(x)r^*(x)\leq 0}\,.$$

Proof. For any $r \in \mathcal{R}$ and $x \in \mathcal{X}$, we can write

$$\begin{aligned}\mathcal{C}_{\ell_2}(r,x) &= \eta(x)\ell_2(r,x,+1) \\ &\quad + [1-\eta(x)]\ell_2(r,x,-1) \\ &= \eta(x)\left[\mathbb{I}_{+1=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{r(x)\leq 0}\right] \\ &\quad + [1-\eta(x)]\left[\mathbb{I}_{-1=-1}\,\mathbb{I}_{r(x)>0} + c\,\mathbb{I}_{r(x)\leq 0}\right] \\ &= c\,\mathbb{I}_{r(x)\leq 0} + [1-\eta(x)]\,\mathbb{I}_{r(x)>0}\,.\end{aligned}$$

For the optimal $\mathcal{C}^*_{\ell_2,\mathcal{R}}$, since $\mathcal{R}$ is regular, we would always pick the lower of c or $1-\eta(x)$, which gives: $\mathcal{C}^*_{\ell_2,\mathcal{R}}(x) = \min\{c, 1-\eta(x)\}$. The corresponding best-in-class solution r^* can be defined by $r^*(x) = \eta(x) - (1-c)$. Thus, the calibration gap is given by

$$\begin{aligned}\Delta\mathcal{C}_{\ell_2}(r,x) &= c\,\mathbb{I}_{r(x)\leq 0} + [1-\eta(x)]\,\mathbb{I}_{r(x)>0} \\ &\quad - \min\{c, 1-\eta(x)\}.\end{aligned}$$

If $r(x)$ correctly chooses the lower of the two, we have $r(x)r^*(x) > 0$ and then $\Delta\mathcal{C}_{\ell_2} = 0$. Otherwise,

$$\Delta\mathcal{C}_{\ell_2}(r,x) = \begin{cases} c - (1-\eta(x)) & \text{if } r(x) \leq 0 \\ (1-\eta(x)) - c & \text{otherwise} \end{cases}.$$

Thus, for all $x \in \mathcal{X}$, we have $\Delta\mathcal{C}_{\ell_2}(r,x) = |\eta(x) - (1-c)|\,\mathbb{I}_{r(x)r^*(x)\leq 0}$. This completes the proof. □

D.3 Calibration gap for surrogate loss

Next, we extend Lemma 3 to bounded hypothesis sets $\mathcal{R}$. The following gives the expression of the calibration gap for the surrogate loss. The proof directly extends that of Lemma 3.

Lemma 11. *Assume that $\mathcal{R}$ is bounded with function $\overline{r}\colon \mathcal{X} \to \mathbb{R}$. Let $I_\eta(x) = \eta(x)e^{-\frac{\alpha}{2}} + (1-\eta(x))e^{\frac{\alpha}{2}}$, $r_0(x) = \log\left[\left(\frac{2\beta c}{\alpha I_\eta(x)}\right)^{\frac{2}{2\beta+\alpha}}\right]$ and $\gamma = \frac{\alpha}{\alpha+2\beta}$. Then, the calibration gap for the surrogate loss is given for any $r \in \mathcal{R}$ and $x \in \mathcal{X}$ by*

$$\Delta\mathcal{C}_{\ell_1}(r,x) = \begin{cases} e^{\frac{\alpha}{2}r(x)}I_\eta(x) + ce^{-\beta r(x)} - \frac{1}{1-\gamma}\left(\frac{2\beta c}{\alpha}\right)^{\gamma} I_\eta(x)^{1-\gamma} & -\overline{r}(x) \leq r_0(x) \leq \overline{r}(x) \\ e^{\frac{\alpha}{2}r(x)}I_\eta(x) + ce^{-\beta r(x)} - e^{\frac{\alpha}{2}\overline{r}(x)}I_\eta(x) - ce^{-\beta\overline{r}(x)} & r_0(x) > \overline{r}(x) \\ e^{\frac{\alpha}{2}r(x)}I_\eta(x) + ce^{-\beta r(x)} - e^{-\frac{\alpha}{2}\overline{r}(x)}I_\eta(x) - ce^{\beta\overline{r}(x)} & r_0(x) < -\overline{r}(x). \end{cases}$$

Proof. By definition, the calibration function for ℓ_1 can be expressed for all $x \in \mathcal{X}$ by

$$\begin{aligned}\mathcal{C}_{\ell_1}(r,x) &= \eta(x)\ell_1(r,x,+1) + [1-\eta(x)]\ell_1(r,x,-1) \\ &= \eta(x)\left[e^{\frac{\alpha}{2}[r(x)-1]} + ce^{-\beta r(x)}\right] + [1-\eta(x)]\left[e^{\frac{\alpha}{2}[r(x)+1]} + ce^{-\beta r(x)}\right] \\ &= e^{\frac{\alpha}{2}r(x)}I_\eta(x) + ce^{-\beta r(x)}.\end{aligned}$$

Since the exponential function is convex, $\Delta\mathcal{C}_{\ell_1}(r,x)$ is a convex function of $r(x)$. Thus, for $r \in \mathcal{R}$, we obtain the minimum $r_0(x)$ by differentiating with respect to $r(x)$ and setting to 0:

$$\begin{aligned}&\frac{\alpha}{2}e^{\frac{\alpha}{2}r(x)}I_\eta(x) - \beta ce^{-\beta r(x)} = 0 \\ \Leftrightarrow\ & r_0(x) = \log\left[\left(\frac{2\beta c}{\alpha I_\eta(x)}\right)^{\frac{2}{2\beta+\alpha}}\right].\end{aligned}$$

Note that for all $x \in \mathcal{X}$, $\{r(x)\colon r \in \mathcal{R}\} = [-\overline{r}(x), \overline{r}(x)]$. If $r_0(x)$ is within this range, plugging in $r_0(x)$ in $\mathcal{C}_{\ell_1}$ gives the corresponding minimal calibration gap $\mathcal{C}^*_{\ell_1}(x)$: $\mathcal{C}^*_{\ell_1}(x) = \left[\left(\frac{2\beta c}{\alpha}\right)^{\gamma}\right] I_\eta(x)^{1-\gamma}\left(\frac{1}{1-\gamma}\right)$. Otherwise, the corresponding minimal calibration gap is achieved at

either $r(x) = \overline{r}(x)$ or $r(x) = -\overline{r}(x)$. Plugging in these expressions give the corresponding minimal calibration gap $\mathcal{C}^*_{\ell_1}(x)$:

$$\mathcal{C}^*_{\ell_1}(x) = \begin{cases} \left[\left(\frac{2\beta c}{\alpha}\right)^{\gamma}\right] I_\eta(x)^{1-\gamma}\left(\frac{1}{1-\gamma}\right) & -\overline{r}(x) \leq r_0(x) \leq \overline{r}(x) \\ e^{\frac{\alpha}{2}\overline{r}(x)} I_\eta(x) + ce^{-\beta\overline{r}(x)} & r_0(x) > \overline{r}(x) \\ e^{-\frac{\alpha}{2}\overline{r}(x)} I_\eta(x) + ce^{\beta\overline{r}(x)} & r_0(x) < -\overline{r}(x). \end{cases}$$

This completes the proof. □

D.4 $\mathcal{H}$-CONSISTENCY BOUND

In this section, we prove our main result. The following result extends Proposition 4 to any hypothesis set $\mathcal{R}$ that is *regular for rejection* and will provide a key tool to derive our result. The proof is nearly identical to Proposition 4.

Proposition 12. *Assume that $\mathcal{R}$ is regular for rejection. Assume that there exists a convex function $\Psi\colon \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) = 0$ such that the following holds for all $r \in \mathcal{R}$ and $x \in \mathcal{X}$: $\Psi\big(|\eta(x) - (1-c)|\,\mathbb{I}_{r(x)r^*(x)\leq 0}\big) \leq \Delta\mathcal{C}_{\ell_1}(0, x)$. Let $\bar{I}_c$ be defined by $\bar{I}_c = ce^{\frac{\alpha}{2}} + (1-c)e^{-\frac{\alpha}{2}}$ and assume that $\frac{2\beta c}{\alpha} = \bar{I}_c$. Then, for any $r \in \mathcal{R}$:*

$$\Psi\big(R_{\ell_2}(r) - R^*_{\ell_2,\mathcal{R}} + \mathcal{M}_{\ell_2,\mathcal{R}}\big) \leq R_{\ell_1}(r) - R^*_{\ell_1,\mathcal{R}} + \mathcal{M}_{\ell_1,\mathcal{R}}. \tag{6}$$

Proof. We will show that the following holds: $\inf_{r(x)r^*(x)\leq 0} \Delta\mathcal{C}_{\ell_1}(r, x) = \Delta\mathcal{C}_{\ell_1}(0, x)$. The result then follows by Theorem 1 and Lemma 10. Since we have $\frac{2\beta c}{\alpha} = \bar{I}_c$, the following equivalence holds:

$$\begin{aligned} r_0(x) > 0 &\Leftrightarrow \frac{2\beta c}{\alpha I_\eta(x)} > 1 \\ &\Leftrightarrow I_\eta(x) < \bar{I}_c \\ &\Leftrightarrow \eta(x) > \tfrac{e^{\frac{\alpha}{2}} - \bar{I}_c}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}} \\ &\Leftrightarrow \eta(x) > \tfrac{(1-c)e^{\frac{\alpha}{2}} - (1-c)e^{-\frac{\alpha}{2}}}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}} \\ &\Leftrightarrow r^*(x) > 0. \end{aligned}$$

This implies $\inf_{r(x)r^*(x)\leq 0} \mathcal{C}_{\ell_1}(r, x) = \inf_{r(x)r_0(x)\leq 0} \mathcal{C}_{\ell_1}(r, x)$. Now, since $r_0(x)$ is the unique minimizer of the strictly convex function $\mathcal{C}_{\ell_1}(r, x)$ of $r(x)$, then, as a function of $r(x)$, $\mathcal{C}_{\ell_1}(r, x)$ is decreasing from $-\infty$ to $r_0(x)$ and increasing from there to $+\infty$. Thus, if $r_0(x) > 0$, the infimum of $\mathcal{C}_{\ell_1}(r, x)$ over $r(x) \leq 0$ is reached for $r(x) = 0$. Similarly, if $r_0(x) < 0$, the infimum of $\mathcal{C}_{\ell_1}(r, x)$ over $r(x) \geq 0$ is reached for $r(x) = 0$. This shows that $\inf_{r(x)r_0(x)\leq 0} \mathcal{C}_{\ell_1}(r, x) = \mathcal{C}_{\ell_1}(0, x)$, and completes the proof. □

The following is our main result; it relates the surrogate estimation error to that of the rejection loss.

Theorem 7. *Assume that $\mathcal{R}$ is bounded with function $\overline{r}\colon \mathcal{X} \to \mathbb{R}$. Let $\alpha, \beta > 0$ be such that $\frac{2\beta c}{\alpha} = \bar{I}_c$, where $\bar{I}_c = ce^{\frac{\alpha}{2}} + (1-c)e^{-\frac{\alpha}{2}}$. Then, the following inequality holds for any $r \in \mathcal{R}$:*

$$R_{\ell_2}(r) - R^*_{\ell_2,\mathcal{R}} + \mathcal{M}_{\ell_2,\mathcal{R}} \leq \Gamma\big(R_{\ell_1}(r) - R^*_{\ell_1,\mathcal{R}} + \mathcal{M}_{\ell_1,\mathcal{R}}\big), \tag{4}$$

where $\Gamma(z) = \begin{cases} \frac{2}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}}\sqrt{\frac{(c+\bar{I}_c)\bar{I}_c}{c}}\, z & 0 \leq z \leq \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]^2 \\ \frac{4(c+\bar{I}_c)}{c\left[1-e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]}\, z & \text{otherwise.} \end{cases}$

Proof. Using the expression of $\Delta\mathcal{C}_{\ell_1}$ given by Lemma 11, we can write

$$\Delta\mathcal{C}_{\ell_1}(0,x) = \begin{cases} I_\eta(x) + c - \frac{1}{1-\gamma}\left(\frac{2\beta c}{\alpha}\right)^\gamma I_\eta(x)^{1-\gamma} & -\overline{r}(x) \le r_0(x) \le \overline{r}(x) \\ I_\eta(x) + c - e^{\frac{\alpha}{2}\overline{r}(x)} I_\eta(x) - ce^{-\beta\overline{r}(x)} & r_0(x) > \overline{r}(x) \\ I_\eta(x) + c - e^{-\frac{\alpha}{2}\overline{r}(x)} I_\eta(x) - ce^{\beta\overline{r}(x)} & r_0(x) < -\overline{r}(x). \end{cases}$$

$$= \begin{cases} I_\eta(x) + c - \left(\bar{I}_c + c\right)\left(\frac{I_\eta(x)}{\bar{I}_c}\right)^{1-\gamma} & -\overline{r}(x) \le r_0(x) \le \overline{r}(x) \\ I_\eta(x) + c - e^{\frac{\alpha}{2}\overline{r}(x)} I_\eta(x) - ce^{-\frac{\alpha\bar{I}_c\overline{r}(x)}{2c}} & r_0(x) > \overline{r}(x) \\ I_\eta(x) + c - e^{-\frac{\alpha}{2}\overline{r}(x)} I_\eta(x) - ce^{\frac{\alpha\bar{I}_c\overline{r}(x)}{2c}} & r_0(x) < -\overline{r}(x). \end{cases}$$

Without loss of generality, we consider $r^*(x) = \eta(x) - (1-c) \ge 0$. Then $r_0(x) \ge 0$. As with the proof of Theorem 5, we can express $\Delta\mathcal{C}_{\ell_1}(0,x)$ in terms of $u(x) = \eta(x) - (1-c)$, using $I_\eta(x) = J_u(x) + \bar{I}_c$, with $J_u(x) = \left[e^{-\frac{\alpha}{2}} - e^{\frac{\alpha}{2}}\right]u(x)$. Note that the condition $r_0(x) \le \overline{r}(x)$ can be expressed as

$$\log\left[\left(\frac{2\beta c}{\alpha I_\eta(x)}\right)^{\frac{2}{2\beta+\alpha}}\right] \le \overline{r}(x) \iff u(x) \le \frac{\bar{I}_c\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)\overline{r}(x)}{2c}}\right]}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}}.$$

When $0 \le r_0(x) \le \overline{r}(x)$, we have

$$\Delta\mathcal{C}_{\ell_1}(0,x) \ge \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[\frac{J_u(x)}{\bar{I}_c}\right]^2 = \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[\frac{u(x)}{\bar{I}_c}\right]^2\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]^2.$$

When $r_0(x) > \overline{r}(x)$, we have

$$\Delta\mathcal{C}_{\ell_1}(0,x) \ge \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\frac{\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)\overline{r}(x)}{2c}}\right]}{\bar{I}_c}\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]u(x).$$

Therefore, the function $\Psi(u)$ defined by

$$\Psi(u) = \begin{cases} \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[\frac{u(x)}{\bar{I}_c}\right]^2\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]^2 & 0 \le u(x) \le \frac{\bar{I}_c\left[1-e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]}{e^{\frac{\alpha}{2}}-e^{-\frac{\alpha}{2}}} \\ \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\frac{\left[1-e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]}{\bar{I}_c}\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]u(x) & \text{otherwise} \end{cases}$$

verifies the condition of Proposition 12 and therefore we have $\Psi\left(R_{\ell_2}(h) - R^*_{\ell_2}\right) \le R_{\ell_1}(h) - R^*_{\ell_1}$. An explicit upper-bound on $R_{\ell_2}(h) - R^*_{\ell_2}$ can be written in terms of Ψ^{-1}: $R_{\ell_2}(h) - R^*_{\ell_2} \le \Psi^{-1}\left(R_{\ell_1}(h) - R^*_{\ell_1}\right)$. To derive the expression of Ψ^{-1}, we write $z = \Psi(u)$, that is: when $0 \le z \le \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]^2$,

$$4\frac{c+\bar{I}_c}{c\bar{I}_c}z = \left[\frac{u(x)}{\bar{I}_c}\right]^2\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]^2 \iff |u| = \frac{2}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}}\sqrt{\frac{(c+\bar{I}_c)\bar{I}_c}{c}}z.$$

Otherwise,

$$z = \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\frac{\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]}{\bar{I}_c}\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]u(x) \iff u = \frac{4(c+\bar{I}_c)}{c\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]}z$$

Thus, we have, for all $r \in \mathcal{R}$, $R_{\ell_2}(r) - R^*_{\ell_2} \le \Gamma\left(R_{\ell_1}(r) - R^*_{\ell_1}\right)$, where

$$\Gamma(z) = \begin{cases} \frac{2}{e^{\frac{\alpha}{2}}-e^{-\frac{\alpha}{2}}}\sqrt{\frac{(c+\bar{I}_c)\bar{I}_c}{c}}z & 0 \le z \le \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]^2 \\ \frac{4(c+\bar{I}_c)}{c\left[1-e^{-\frac{\alpha(\bar{I}_c+c)\inf_{x\in\mathcal{X}}\overline{r}(x)}{2c}}\right]\left[e^{\frac{\alpha}{2}}-e^{-\frac{\alpha}{2}}\right]}z & \text{otherwise.} \end{cases}$$

□

Theorem 7 implies the following corollary for $\mathcal{R} = \mathcal{R}_{\text{lin}}$ and $\mathcal{R}_{\text{NN}}$.

Corollary 8. *Let $\mathcal{R} = \mathcal{R}_{\text{lin}}$ or $\mathcal{R}_{\text{NN}}$. Let $\alpha, \beta > 0$ be such that $\frac{2\beta c}{\alpha} = \bar{I}_c$, where $\bar{I}_c = ce^{\frac{\alpha}{2}} + (1-c)e^{-\frac{\alpha}{2}}$. Then, the following inequality holds for any $r \in \mathcal{R}$:*

$$R_{\ell_2}(r) - R^{\star}_{\ell_2,\mathcal{R}} + \mathcal{M}_{\ell_2,\mathcal{R}} \leq \Gamma\Big(R_{\ell_1}(r) - R^{\star}_{\ell_1,\mathcal{R}} + \mathcal{M}_{\ell_1,\mathcal{R}}\Big), \tag{5}$$

where $\Gamma(z) = \begin{cases} \frac{2}{e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}} \sqrt{\frac{(c+\bar{I}_c)\bar{I}_c}{c}} z & 0 \leq z \leq \frac{1}{4}\frac{c\bar{I}_c}{c+\bar{I}_c}\left[1 - e^{-\frac{\alpha(\bar{I}_c+c)B}{2c}}\right]^2 \\ \frac{4(c+\bar{I}_c)}{c\left[1-e^{-\frac{\alpha(\bar{I}_c+c)B}{2c}}\right]\left[e^{\frac{\alpha}{2}} - e^{-\frac{\alpha}{2}}\right]} z & \text{otherwise} \end{cases}$ *and B is replaced by ΛB for $\mathcal{R} = \mathcal{R}_{\text{NN}}$.*

Proof. Using the fact that $\inf_{x\in\mathcal{X}} \overline{r}(x) = \inf_{x\in\mathcal{X}}\big(W\|x\|_p + B = B\big)$ for $\mathcal{R} = \mathcal{R}_{\text{lin}}$ and $\inf_{x\in\mathcal{X}} \overline{r}(x) = \inf_{x\in\mathcal{X}}\big(\Lambda W\|x\|_p + \Lambda B\big) = \Lambda B$ for $\mathcal{R} = \mathcal{R}_{\text{NN}}$, by Theorem 7, we complete the proof. □