Characterizing Out-of-Distribution Error via Optimal Transport (Appendix)

Anonymous Author(s)
Affiliation
Address
email

A Deferred Proofs

For readers' convenience, we review the statements of the propositions and corollaries and provide the full proofs below.

A.1 Proof of Proposition 1

Proposition 1 ($\hat{\epsilon}_{\text{AC}}$-$W_\infty$ Equivalence)**.** *Let $(\mathcal{P}(\mathbb{R}^k), W_\infty)$ be the metric space of all distributions over $\mathbb{R}^k$, where W_∞ is the Wasserstein distance with $c(x, y) = \|x - y\|_\infty$. Then, the estimated error of AC-MC is given by $\hat{\epsilon}_{AC} = W_\infty(\vec{f}_{\#}P(\vec{c}), P_{pseudo}(\vec{y}))$.*

Proof. We first show the following equality. Let $j^* = \arg\max_j \vec{f}_j(x)$

$$1 - \vec{f}_{j^*}(x) = \|\vec{y} - \vec{f}(x)\|_\infty \tag{1}$$

where $\vec{y}_j = \mathbb{1}[j = j^*]$. Let $\vec{f}_{-j^*}(x)$ denote the vector $\vec{f}(x)$ with j^*-th element removed. Since for a confidence vector $\|\vec{f}(x)\|_1 = 1$,

$$1 - \vec{f}_{j^*}(x) = \|\vec{f}_{-j^*}(x)\|_1 \geq \|\vec{f}_{-j^*}(x)\|_\infty$$

Therefore, we have obtained the desired Equality 1:

$$\|\vec{y} - \vec{f}_j(x)\|_\infty = \max\{\|\vec{f}_{-j^*}(x)\|_\infty, 1 - \vec{f}_{j^*}(x)\} = 1 - \vec{f}_{j^*}(x)$$

Next, we consider the optimal transport plan between $\vec{f}_{\#}P(\vec{c})$ and $P_{\text{pseudo}}(\vec{y})$. Namely, we show *all* confidence vectors $\vec{f}(x^{(i)})$ are coupled with their one-hot pseudo-labels $\vec{y}^{(i)}$. This can be observed by the fact that the one-hot pseudo-label is the one-hot label that achieves the lowest L-infinity cost, i.e.

$$\|\vec{f}(x^{(i)}) - \vec{y}^{(i)}\|_\infty \leq \|\vec{f}(x^{(i)}) - \vec{y}'\|_\infty, \forall \vec{y}' \in \{0, 1\}^k \cap \Delta^{k-1}$$

Suppose there exist confidence vectors that are not coupled with their one-hot pseudo-labels, then *all* individual costs are suboptimal and the total cost is suboptimal as well, contradicting the assumption that the transport plan is optimal. Therefore,

$$W_\infty(\vec{f}_{\#}P(\vec{c}), P_{\text{pseudo}}(\vec{y})) = \frac{1}{n}\sum_{i=1}^{n} \|\vec{f}(x^{(i)}) - \vec{y}^{(i)}\|_\infty \tag{2}$$

Combining Equality 1 and Equality 2, we obtain the desired relationship between AC error estimate and W_∞ distance

$$\hat{\epsilon}_{\text{AC}} = \frac{1}{n}\sum_{i=1}^{n}(1 - \max_j \vec{f}_j(x^{(i)})) = \frac{1}{n}\sum_{i=1}^{n} \|\vec{f}(x^{(i)}) - \vec{y}^{(i)}\|_\infty = W_\infty(\vec{f}_{\#}P(\vec{c}), P_{\text{pseudo}}(\vec{y}))$$

□

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

A.2 Proof of Corollary 1

Corollary 1 ($P_{\text{pseudo}}(\vec{y})$ is closest to $\vec{f}_{\#}P(\vec{c})$)**.** *Let $P'(\vec{y}) \in \mathcal{P}(\{0,1\}^k \cap \Delta^{k-1})$ be a one-hot label distribution. Then $W_\infty(\vec{f}_{\#}P(\vec{c}), P'(\vec{y})) \geq W_\infty(\vec{f}_{\#}P(\vec{c}), P_{pseudo}(\vec{y}))$.*

Proof. We first show the following equality, which establishes the relationship between AC accuracy estimate with W_∞ distance:

$$1 - \hat{\epsilon}_{\text{AC}} = W_\infty(\vec{f}_{\#}P(\vec{c}), \delta_0) \tag{3}$$

Since $W_\infty(\vec{f}_{\#}P(\vec{c}), \delta_0)$ transports $\vec{f}_{\#}P(\vec{c})$ to δ_0, the optimal transport plan couples every element in $\vec{f}_{\#}P(\vec{c})$ to 0. For each $x^{(i)}$, its confidence vector $\vec{f}_{\#}P(\vec{c})$ has a transport cost $\|\vec{f}(x^{(i)}) - 0\|_\infty$. Hence,

$$1 - \hat{\epsilon}_{\text{AC}} = \frac{1}{n}\sum_{i=1}^{n} \max_j \vec{f}_j(x^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} \|\vec{f}(x^{(i)}) - 0\|_\infty = W_\infty(\vec{f}_{\#}P(\vec{c}), \delta_0)$$

With this, our inequality is simply the Triangle Inequality in $(\mathcal{P}(\mathbb{R}^k), W_\infty)$,

$$W_\infty(\vec{f}_{\#}P(\vec{c}), \delta_0) + W_\infty(\vec{f}_{\#}P(\vec{c}), P'(\vec{y})) \geq W_\infty(P'(\vec{y}), \delta_0) = 1$$

Combined with Equation 3, we obtain the desired inequality

$$W_\infty(\vec{f}_{\#}P(\vec{c}), P_{\text{pseudo}}(\vec{y})) = 1 - W_\infty(\vec{f}_{\#}P(\vec{c}), \delta_0) \leq W_\infty(\vec{f}_{\#}P(\vec{c}), P'(\vec{y}))$$

□

A.3 Proof of Proposition 2

Notations: Let $\mathcal{C}(\vec{c}) = \{\vec{c}' \in \Delta^{k-1} | \arg\max_j \vec{c}'_j = \arg\max_j \vec{c}_j\}$ be the set of confidence vectors whose one-hot pseudo-labels that match with that of $\vec{c} \in \Delta^{k-1}$. Let $\mathcal{P}(\Delta^{k-1})$ be the set of all distributions of confidence vectors and $\mathcal{P}_c[P_{\text{pseudo}}(\vec{y})] = \{P'(\vec{c}) \in \mathcal{P}(\Delta^{k-1}) | P_{\text{pseudo}}(\vec{y}) = P'(\arg\max_j \vec{c}_j = \arg\max_j \vec{y}_j)\}$ be the set of distributions of confidence vectors that share the same pseudo-label distribution $P_{\text{pseudo}}(\vec{y})$.

$\mathcal{P}_c[P_{\text{pseudo}}(\vec{y})]$ defines an equivalence class for the space of distributions of confidence vectors $(\mathcal{P}(\Delta^{k-1}), W_\infty)$ that share the same pseudo-label distribution $P_{\text{pseudo}}(\vec{y})$. Pictorially, in Figure 1, $\mathcal{P}_c[P_{\text{pseudo}}(\vec{y})]$ represents the line between δ_0 and $P_{\text{pseudo}}(\vec{y})$. On this line, every distribution of confidence vectors shares the same pseudo-label distribution $P_{\text{pseudo}}(\vec{y})$.

To prove Proposition 2, we need the following lemma, which intuitively allows us to change the metric from measuring the distance between two points to the distance between an equivalence class and a point.[1]

Lemma 1 (Change-of-metric)**.** *Let $\vec{y}, \vec{y}' \in \{0,1\}^k \cap \Delta^{k-1}$ be two one-hot labels. Then the following holds*

$$\inf_{\vec{c} \in \mathcal{C}(\vec{y})} \|\vec{c} - \vec{y}'\|_\infty = 0.5 \times \mathbb{1}[\vec{y} \neq \vec{y}']$$

Proof. If $\vec{y} = \vec{y}'$, then we know the optimal $\vec{c} = \vec{y}$

$$\inf_{\vec{c} \in \mathcal{C}(\vec{y})} \|\vec{c} - \vec{y}'\|_\infty = \|\vec{y} - \vec{y}'\|_\infty = 0$$

If $\vec{y} \neq \vec{y}'$, then we proceed by showing equality with two inequalities. First, observe $\{(0.5+\delta)\,\vec{y} + (0.5-\delta)\,\vec{y}' | \delta \in (0, 0.5]\} \subset \mathcal{C}(\vec{y})$.

$$\inf_{\vec{c} \in \mathcal{C}(\vec{y})} \|\vec{c} - \vec{y}'\|_\infty \leq \inf_{\delta \in (0, 0.5]} \|(0.5+\delta)\,\vec{y} + (0.5-\delta)\,\vec{y}' - \vec{y}'\|_\infty = \inf_{\delta \in (0, 0.5]} (0.5+\delta) = 0.5$$

If $\|\vec{c} - \vec{y}'\|_\infty < 0.5$, $\arg\max_j \vec{c}_j = \arg\max_j \vec{y}'_j \neq \arg\max_j \vec{y}_j$, i.e. $\vec{c} \notin \mathcal{C}(\vec{y})$. Therefore, $\inf_{\vec{c} \in \mathcal{C}(\vec{y})} \|\vec{c} - \vec{y}'\|_\infty \geq 0.5$, which further implies $\inf_{\vec{c} \in \mathcal{C}(\vec{y})} \|\vec{c} - \vec{y}'\|_\infty = 0.5$. □

[1] This is closely related to the Hausdorff distance between sets in a metric space.

We are now in a position to prove Proposition 2, which follows from the somewhat surprising fact that the left-hand side of the inequality is simply the distance between $\vec{f}_{\#}P_T(\vec{c})$ and $P_T(\vec{y})$ with a change-of-metric to the metric defined above.

Proposition 2 (Calibration independent lower bound of COT)**.** *Under the assumption that $P_T(\vec{y}) = P_S(\vec{y})$, we always have $\hat{\epsilon}_{COT} \geq 0.5W_\infty(P_{pseudo}(\vec{y}), P_T(\vec{y}))$.*

Proof. Since $P_T(\vec{y}) = P_S(\vec{y})$,

$$\begin{aligned}
\hat{\epsilon}_{\text{COT}} &= W_\infty(\vec{f}_{\#}P_T(\vec{c}), P_T(\vec{y})) \\
&= \inf_{\pi(\vec{c},\vec{y})\in\Pi(\vec{f}_{\#}P_T(\vec{c}),P_T(\vec{y}))} \int \|\vec{c}-\vec{y}\|_\infty d\pi(\vec{c},\vec{y}) \\
&\geq \inf_{\pi(\vec{c},\vec{y})\in\Pi(\vec{f}_{\#}P_T(\vec{c}),P_T(\vec{y}))} \int \inf_{\vec{c}'\in\mathcal{C}(\vec{c})} \|\vec{c}'-\vec{y}\|_\infty d\pi(\vec{c},\vec{y}) \qquad (4) \\
&= \inf_{\pi(\vec{y}',\vec{y})\in\Pi(P_{\text{pseudo}}(\vec{y}'),P_T(\vec{y}))} \int \inf_{\vec{c}'\in\mathcal{C}(\vec{y}')} \|\vec{c}'-\vec{y}\|_\infty d\pi(\vec{y}',\vec{y}) \qquad (5)
\end{aligned}$$

Equation 5 follows from the observation that $\mathcal{C}(\vec{c}) = \mathcal{C}(\vec{y}')$ for a confidence vector $\vec{c}$ and its corresponding one-hot pseudo-label $\vec{y}'$. Furthermore, since our new metric $\inf_{\vec{c}'\in\mathcal{C}(\vec{y}')} \|\vec{c}'-\vec{y}\|_\infty$ is only defined up to the equivalence class, replacing each $\vec{c}' \in \mathcal{C}(\vec{c})$ with its pseudo-label $\vec{y}'$ does not change the distance.

Plugging in Lemma 1,

$$\begin{aligned}
\hat{\epsilon}_{\text{COT}} &\geq \inf_{\pi(\vec{y}',\vec{y})\in\Pi(P_{\text{pseudo}}(\vec{y}'),P_T(\vec{y}))} \int 0.5 \times \mathbb{1}[\vec{y}' \neq \vec{y}] d\pi(\vec{y}',\vec{y}) \\
&= 0.5 \inf_{\pi(\vec{y}',\vec{y})\in\Pi(P_{\text{pseudo}}(\vec{y}'),P_T(\vec{y}))} \int \|\vec{y}'-\vec{y}\|_\infty d\pi(\vec{y}',\vec{y}) \\
&= 0.5W_\infty(P_{\text{pseudo}}(\vec{y}), P_T(\vec{y}))
\end{aligned}$$

□

A.4 Tightness of Proposition 2

While Inequality 4 seems loose, our bound is, in fact, tight if no further assumptions on the calibration status of the classifier $\vec{f}$ are made. We need the following lemma that establishes the relationship between pseudo-label shift and the total variation distance between target label distribution and pseudo-label distribution. Note this equivalence only makes sense in the context of measuring W_∞ distance between two one-hot label distributions, but not under other contexts presented in the paper.

Lemma 2 (Pseudo-label shift is total variation)**.**

$$W_\infty(P_{pseudo}(\vec{y}), P_T(\vec{y})) = \|P_{pseudo}(\vec{y}) - P_T(\vec{y})\|_{TV}$$

Proof. For two $\vec{y}, \vec{y}' \in \{0,1\}^k \cap \Delta^{k-1}$, the transport cost $c(\vec{y},\vec{y}') = \mathbb{1}[\vec{y} \neq \vec{y}']$. Then, the standard result on optimal transport [10] gives the desired equality. □

Corollary 2.

$$W_\infty(P_{pseudo}(\vec{y}), P_T(\vec{y})) = 1 - \sum_{\vec{y}\in\{0,1\}^k\cap\Delta^{k-1}} \min\{P_{pseudo}(\vec{y}), P_T(\vec{y})\}$$

Proof.

$$
\begin{aligned}
W_\infty(P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})) &= \|P_{\text{pseudo}}(\vec{y}) - P_T(\vec{y})\|_{\text{TV}} \\
&= \frac{1}{2} \sum_{\vec{y} \in \{0,1\}^k \cap \Delta^{k-1}} |P_{\text{pseudo}}(\vec{y}) - P_T(\vec{y})| \\
&= \frac{1}{2} \sum_{\vec{y} \in \{0,1\}^k \cap \Delta^{k-1}} \max\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\} - \min\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\} \\
&= \frac{1}{2} \sum_{\vec{y} \in \{0,1\}^k \cap \Delta^{k-1}} \max\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\} + \min\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\} \\
&- \sum_{\vec{y} \in \{0,1\}^k \cap \Delta^{k-1}} \min\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\} \\
&= \frac{1}{2} \sum_{\vec{y} \in \{0,1\}^k \cap \Delta^{k-1}} P_{\text{pseudo}}(\vec{y}) + P_T(\vec{y}) - \sum_{\vec{y} \in \{0,1\}^k \cap \Delta^{k-1}} \min\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\} \\
&= 1 - \sum_{\vec{y} \in \{0,1\}^k \cap \Delta^{k-1}} \min\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\}
\end{aligned}
$$

□

Finally, we show Proposition 2 is tight by constructing a sequence of distributions of confidence vectors, the limit of which is exactly $0.5W_\infty(P_{\text{pseudo}}(\vec{y}), P_T(\vec{y}))$ away from $P_T(\vec{y})$.

Lemma 3 (Proposition 2 is tight)**.**

$$
\inf_{P(\vec{c}) \in \mathcal{P}_c[P_{pseudo}(\vec{y})]} W_\infty(P(\vec{c}), P_T(\vec{y})) = 0.5W_\infty(P_{pseudo}(\vec{y}), P_T(\vec{y}))
$$

Proof. First, we construct the following family of distributions $\{P_\delta(\vec{c}) | \delta \in (0, 0.5]\}$, where $P_\delta(\vec{c})$ is the following mixture distribution

$$
P_\delta(\vec{c}) = \gamma P_\cap(\vec{y}) + (1 - \gamma) P_\times(\vec{t})
$$

where $P_\cap(\vec{y}) = \gamma^{-1} \min\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\}$, $\gamma = \sum_{\vec{y} \in \{0,1\}^k \cap \Delta^{k-1}} \min\{P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})\}$, $P_\times(\vec{t})$ is a distribution supported on $\Delta^{k-1} \cap \{0.5 + \delta, 0.5 - \delta, 0\}^k$ (i.e. one element in $\vec{t}$ is $0.5 + \delta$, another is $0.5 - \delta$, and the rest are 0). Additionally, $P_\times(\vec{t_i} = 0.5 + \delta) = P_{\text{pseudo}}(\vec{y_i} = 1)$ and $P_\times(\vec{t_i} = 0.5 - \delta) = P_T(\vec{y_i} = 1)$. It is easy to check that $P_\delta(\vec{c}) \in \mathcal{P}_c[P_{\text{pseudo}}(\vec{y})]$.

Next, we construct an explicit transport plan $\pi(\vec{c}, \vec{y}) \in \Pi(P_\delta(\vec{c}), P_T(\vec{y}))$. We construct it via the factorization $\pi(\vec{c}, \vec{y}) = P_\delta(\vec{c})\pi(\vec{y}|\vec{c})$, where

$$
\pi(\vec{y}|\vec{c}) = \begin{cases} 1 & \text{if } \vec{c} = \vec{y} \text{ or } \langle \vec{c}, \vec{y} \rangle = 0.5 - \delta \\ 0 & \text{otherwise} \end{cases}
$$

The cost of this transport plan is therefore

$$
\int \|\vec{c} - \vec{y}\|_\infty d\pi(\vec{c}, \vec{y}) = (0.5 + \delta)(1 - \gamma) = (0.5 + \delta) W_\infty(P_{\text{pseudo}}(\vec{y}), P_T(\vec{y}))
$$

where the last equality follows from Lemma 2. Taking infimum,

$$
\inf_{\delta \in (0, 0.5]} \int \|\vec{c} - \vec{y}\|_\infty d\pi(\vec{c}, \vec{y}) = 0.5W_\infty(P_{\text{pseudo}}(\vec{y}), P_T(\vec{y}))
$$

Combining everything so far, we obtain the desired result:

$$
\begin{aligned}
0.5W_\infty(P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})) &= \inf_{\delta \in (0,0.5]} \int \|\vec{c} - \vec{y}\|_\infty d\pi(\vec{c}, \vec{y}) \\
&\geq \inf_{\delta \in (0,0.5]} W_\infty(P_\delta(\vec{c}), P_T(\vec{y})) \quad (6) \\
&\geq \inf_{P(\vec{c}) \in \mathcal{P}_c[P_{\text{pseudo}}(\vec{y})]} W_\infty(P(\vec{c}), P_T(\vec{y})) \quad (7) \\
&\geq 0.5W_\infty(P_{\text{pseudo}}(\vec{y}), P_T(\vec{y})) \quad (8)
\end{aligned}
$$

Inequality 6 follows from the fact that the optimal transport plan cannot have a greater cost than our explicit plan π. Inequality 7 is due to the fact that the family of distribution we are considering is a subset of $\mathcal{P}_c[P_{\text{pseudo}}(\vec{y})]$. Inequality 8 is an application of the lower bound 4 which holds for all $P(\vec{c}) \in \mathcal{P}_c[P_{\text{pseudo}}(\vec{y})]$. □

B Extended Results

B.1 Results with Standard Deviation

We show the full experimental results with standard deviation in Table 1.

B.2 Qualitative Results

We show the qualitative results (scatter plots) in Fig 1

B.3 Correlation Analysis

ProjNorm [12] leverages pseudo labels on the target domain to retrain a copy of the reference model trained on the source domain. The authors show that the difference between the two models' parameters has a strong linear correlation to the true target error. Following the paper's experimental setup, we conducted the correlation analysis on CIFAR10 and CIFAR100 using three architectures, ResNet18, ResNet50, and VGG11. We note that ProjNorm in fact implicitly leverages the assumption that $P_T(y) = P_S(y)$ as this condition holds for both CIFAR10 and CIFAR100. As Fig. 18 of their paper [12] shows, ProjNorm tends to overestimate when label shift exists.

B.4 Mild Label Shift

We motivate our methods under the assumption of no label shift. In Proposition 2, we showed that the worst-case underestimate of COT is half of the pseudo-label shift. Under mild label shifts, the guarantee for such worst-case underestimation becomes weaker. This can be observed from the following corollary of Proposition 2:

Corollary 3 (Calibration independent lower bound of COT under *mild* label shift)**.**

$$\hat{\epsilon}_{COT} \geq 0.5W_\infty(P_{pseudo}(\vec{y}), P_T(\vec{y})) - W_\infty(P_S(\vec{y}), P_T(\vec{y}))$$

Proof. By Triangle Inequality in $(\mathcal{P}(\mathbb{R}^k), W_\infty)$,

$$W_\infty(\vec{f}_\# P_T(\vec{c}), P_S(\vec{y})) + W_\infty(P_S(\vec{y}), P_T(\vec{y})) \geq W_\infty(\vec{f}_\# P_T(\vec{c}), P_T(\vec{y}))$$

Combined with Proposition 2, we obtain the desired result. □

As the label shift increases, we have a weaker guarantee of the worst-case underestimation error of COT as long as $W_\infty(P_S(\vec{y}), P_T(\vec{y})) \leq 0.5W_\infty(P_{\text{pseudo}}(\vec{y}), P_T(\vec{y}))$. However, we perform additional controlled experiments which suggest our methods remain to be the most performant despite the theoretical guarantee is not as strong as the case without label shift.

To simulate mild label shift for datasets with $P_S(\vec{y}) = P_T(\vec{y})$, we first calculate the original target marginal and then sample the shifted target marginal from a Dirichlet distribution as in [1] with a parameter $\alpha = 50$. The parameter α controls the severity of the label shift, and a smaller α means a larger label shift. Concretely, let the shifted target marginal be $P_{\tilde{T}}(\vec{y})$. Then $P_{\tilde{T}}(\vec{y}) \sim \text{Dir}(\beta)$

Table 1: Mean Absolute Error (MAE) between the estimated error and ground truth error to compare different methods. The "shift" column denotes the nature of distribution shifts for each dataset. For vision datasets, we reported results for ResNet18 and ResNet50; for language datasets, we reported results for DistilBERT-base-uncased. The results are averaged over 3 random seeds. We highlight the best-performing method. The number in the parentheses denotes the standard deviation.

		Baselines						Ours	
Dataset	**Shift**	**AC**	**DoC**	**IM**	**GDE**	**ATC-MC**	**ATC-NE**	**COT**	**COTT**
CIFAR10	Natural	5.97 (0.10)	5.38 (0.08)	5.87 (0.09)	5.9 (0.15)	3.38 (0.14)	**3.15** (0.28)	5.41 (0.09)	3.33 (0.13)
	Synthetic	9.1 (0.25)	8.53 (0.28)	9.26 (0.35)	8.84 (0.11)	4.2 (0.38)	3.37 (0.30)	2.17 (0.09)	**1.7** (0.26)
CIFAR100	Synthetic	10.83 (0.08)	8.76 (0.22)	12.07 (0.37)	11.36 (0.25)	6.8 (0.39)	6.63 (0.43)	**2.09** (0.27)	2.59 (0.01)
ImageNet	Natural	8.5 (0.39)	7.43 (0.41)	8.62 (0.47)	5.62 (0.33)	3.57 (0.46)	2.6 (0.66)	3.88 (0.04)	**2.41** (0.11)
	Synthetic	10.34 (0.83)	9.28 (0.86)	12.87 (0.77)	6.54 (0.37)	1.59 (0.08)	3.41 (0.53)	3.24 (0.28)	**1.42** (0.29)
Entity13	Same	19.63 (2.17)	19.2 (2.51)	17.5 (0.90)	15.37 (1.06)	8.09 (0.49)	7.23 (0.49)	8.47 (0.66)	**2.61** (0.31)
	Novel	29.61 (2.61)	29.18 (2.95)	27.22 (1.08)	24.48 (0.61)	14.54 (0.95)	9.49 (0.70)	15.9 (0.80)	**5.46** (0.75)
Entity30	Same	16.97 (0.35)	16.21 (0.36)	13.56 (2.53)	13.98 (0.26)	8.19 (1.07)	9.08 (0.42)	5.9 (0.29)	**2.46** (0.65)
	Novel	27.57 (0.06)	26.81 (0.61)	23.96 (2.79)	23.4 (0.1)	13.46 (2.55)	8.57 (2.2)	15.11 (0.38)	**5.94** (1.17)
Living17	Same	14.84 (3.36)	14.67 (3.30)	11.22 (2.06)	9.94 (0.48)	4.88 (0.42)	5.43 (1.06)	6.25 (1.91)	**2.94** (1.21)
	Novel	29.61 (3.76)	29.18 (3.71)	27.22 (3.45)	24.48 (0.74)	14.54 (2.87)	9.49 (3.25)	15.9 (2.04)	**5.53** (1.93)
Nonliving26	Same	19.25 (2.45)	18.43 (3.13)	16.6 (0.96)	12.77 (0.85)	11.18 (2.77)	9.69 (0.70)	7.06 (1.17)	**3.34** (0.90)
	Novel	31.37 (2.99)	30.54 (3.65)	28.79 (1.47)	23.37 (0.61)	19.93 (4.02)	16.56 (1.28)	17.8 (1.53)	**10.46** (3.08)
Camelyon17-WILDS	Natural	9.44 (0.50)	9.44 (0.49)	10.24 (0.38)	**5.19** (0.44)	7.73 (0.72)	7.73 (0.72)	7.27 (0.57)	5.71 (0.94)
RxRx1-WILDS	Natural	5.21 (0.26)	8.44 (0.15)	8.09 (0.16)	7.48 (0.26)	6.53 (0.10)	6.86 (0.28)	**3.25** (0.16)	5.82 (0.31)
Amazon-WILDS	Natural	2.62 (0.16)	2.35 (0.06)	2.34 (0.06)	17.04 (0.84)	1.63 (0.1)	**1.54** (0.11)	2.43 (0.04)	2.01 (0.42)
CivilCom.-WILDS	Natural	1.54 (0.23)	0.96 (0.19)	**0.86** (0.20)	8.7 (0.14)	2.3 (0.34)	2.3 (0.34)	1.23 (0.05)	4.68 (0.39)

where $\beta_{(\vec{y})} = \alpha \cdot P_T(\vec{y})$. Finally, based on $P_{\tilde{T}}(\vec{y})$, we sample a new set of test samples for which we estimate the performance. We conducted this mild label shift experiment for CIFAR10, CIFAR100, ImageNet, Living17, Nonliving26, Entity13, and Entity30 as these datasets have the same source and target marginal. We showed the results in Table 2. As we can see, our methods still dominate existing methods under this relaxed condition.

B.5 When does thresholding improve over averaging?

In this section, we provide some intuitions on when using a threshold provides better estimates than taking the average. From Fig. 2, we show that thresholding yields larger and more accurate error estimates when the cost distribution on the OOD data is more spread out and less concentrated around 0. By contrast, when the cost distribution is mostly near 0, thresholding leads to similar estimates as averaging. Interestingly, even on OOD data where the model has very low performance, there is still a decent amount of samples whose cost is near 0. Thus, when taking the average, we will end up with a smaller value which suggests a low error. In these cases, thresholding will give larger error estimates than averaging.

Table 2: Mean Absolute Error (MAE) between the estimated error and ground truth error to compare different methods under mild label shift. The results are averaged over 3 random seeds. We highlight the best-performing method. The number in the parentheses denotes the standard deviation.

Dataset	Shift	Baselines						Ours	
		AC	**DoC**	**IM**	**GDE**	**ATC-MC**	**ATC-NE**	**COT**	**COTT**
CIFAR10	Natural	5.58 (0.26)	4.99 (0.23)	5.50 (0.22)	5.69 (0.04)	2.76 (0.32)	2.47 (0.43)	3.75 (0.28)	**1.68** (0.34)
	Synthetic	8.67 (0.29)	8.10 (0.31)	8.82 (0.38)	8.47 (0.15)	3.93 (0.38)	3.13 (0.32)	**2.76** (0.04)	4.0 (0.30)
CIFAR100	Synthetic	10.89 (0.15)	8.85 (0.22)	12.14 (0.37)	11.33 (0.23)	6.93 (0.44)	6.76 (0.48)	**1.89** (0.30)	2.81 (0.07)
ImageNet	Natural	8.36 (0.37)	7.29 (0.42)	8.46 (0.48)	5.54 (0.36)	3.53 (0.48)	2.47 (0.71)	3.74 (0.20)	**2.05** (0.26)
	Synthetic	10.26 (0.83)	9.19 (0.86)	12.79 (0.77)	6.50 (0.40)	**1.61** (0.08)	3.51 (0.52)	3.03 (0.25)	1.75 (0.33)
Entity13	Same	15.50 (0.38)	14.60 (0.34)	15.49 (0.22)	15.18 (1.03)	8.51 (0.80)	7.40 (0.57)	4.59 (0.23)	**3.24** (0.12)
	Novel	24.39 (0.23)	23.49 (0.19)	24.56 (0.05)	23.48 (0.62)	14.99 (0.82)	12.45 (0.64)	11.19 (0.34)	**4.6** (0.38)
Entity30	Same	15.46 (0.70)	13.93 (0.65)	15.55 (0.74)	13.83 (0.27)	8.80 (0.64)	8.26 (0.83)	4.75 (0.29)	**2.16** (0.15)
	Novel	25.98 (0.53)	24.45 (0.46)	26.72 (0.68)	23.28 (0.14)	15.56 (0.56)	13.21 (0.90)	13.96 (0.17)	**7.07** (0.27)
Living17	Same	11.38 (0.67)	10.90 (0.48)	11.83 (1.31)	9.85 (0.41)	4.46 (0.31)	4.39 (0.18)	4.40 (0.34)	**2.71** (0.81)
	Novel	25.72 (0.46)	25.13 (0.76)	26.32 (1.98)	21.61 (0.68)	14.09 (2.31)	11.48 (1.99)	16.94 (0.78)	**9.31** (1.79)
Nonliving26	Same	16.28 (0.37)	14.48 (0.29)	15.69 (0.19)	12.88 (0.84)	9.63 (0.43)	9.69 (0.66	5.33 (0.73)	**2.18** (0.23)
	Novel	27.93 (0.08)	26.13 (0.25)	27.76 (0.29)	23.25 (0.69)	18.15 (0.38)	16.08 (0.38)	15.66 (0.45)	**8.71** (0.42)

Table 3: Coefficients of determination (R^2) and rank correlations (ρ) to measure the linear correlation between a method's output quantity and the true target error (the higher the better). COT achieves superior performance than all existing methods across different models and datasets.

Dataset	Network	AC		Entropy		GDE		ATC		ProjNorm		COT	
		R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ
CIFAR10	ResNet18	0.825	0.980	0.862	0.982	0.842	0.981	0.875	0.987	0.947	0.988	**0.996**	**0.998**
	ResNet50	0.950	0.995	0.949	0.995	0.959	0.995	0.885	0.989	0.936	0.989	**0.993**	**0.996**
	VGG11	0.710	0.938	0.762	0.958	0.723	0.948	0.548	0.851	0.756	0.949	**0.994**	**0.993**
	Average	0.828	0.971	0.858	0.978	0.841	0.975	0.769	0.942	0.880	0.975	**0.994**	**0.996**
CIFAR100	ResNet18	0.943	0.987	0.932	0.984	0.950	0.988	0.927	0.985	0.969	0.974	**0.995**	**0.997**
	ResNet50	0.957	0.987	0.948	0.984	0.962	0.989	0.955	0.991	0.982	0.991	**0.992**	**0.996**
	VGG11	0.794	0.959	0.821	0.973	0.870	0.978	0.736	0.975	0.653	0.849	**0.996**	**0.997**
	Average	0.898	0.978	0.900	0.980	0.927	0.985	0.873	0.984	0.868	0.938	**0.994**	**0.997**

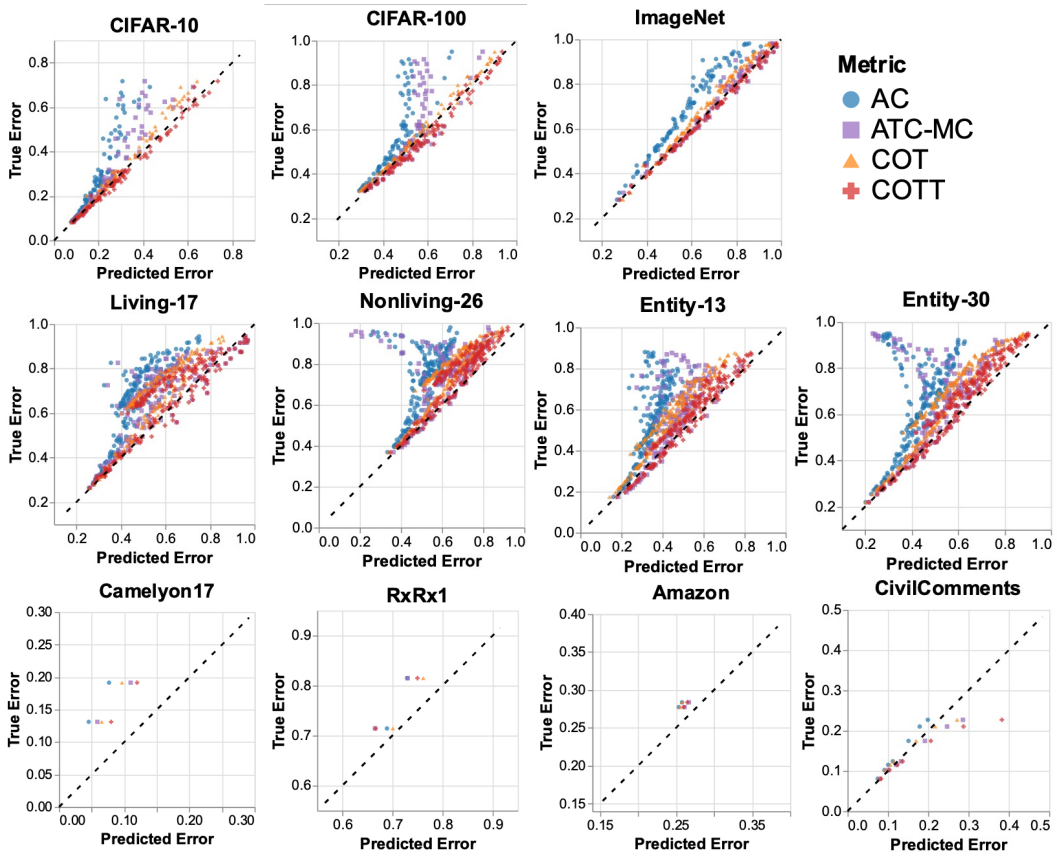

Figure 1: Qualitative results for AC, ATC, COT, and COTT. In these scatterplots, the x-axis is the target error estimate and the y-axis is the ground truth target error. Accurate estimates should be close to $y = x$ (dashed black line). We can see that for all datasets, COT and COTT avoid the severe underestimation seen on ATC.

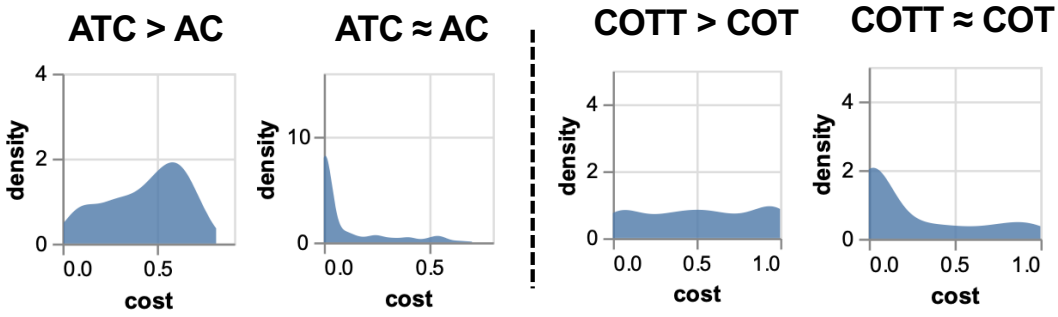

Figure 2: We demonstrate cases where using thresholding improves over taking averages. The x-axis denotes the max norm between a confidence vector and the corresponding one-hot label. For AC and ATC-MC, the corresponding label is always the argmax of the confidence vector as mentioned in section 2.3. For COT and COTT, the corresponding label is the one matched via optimal transport. We observe that thresholding improves over averaging when the cost distribution is less concentrated around 0, which corresponds to situations where the model is very confident on most samples.

C Datasets

CIFAR10: The synthetic shifts included 19 common visual corruptions across 5 levels of severity from [2]. The natural shift is CIFAR10-V2 [7].

CIFAR100: The synthetic shifts included 19 common visual corruptions across 5 levels of severity from [2].

ImageNet: The synthetic shifts included 19 common visual corruptions across 5 levels of severity from [2]. The natural shifts include 4 datasets from ImageNet-V2 [8] and ImageNet-Sketch [11].

BREEDS: The BREEDS benchmark contains 4 datasets, Living-17, Nonliving26, Entity13, Entity30. For each of the datasets, the same subpopulation shifts include the corrupted versions of the test set with the same subpopulation; the novel subpopulation shifts include the clean as well as corrupted versions [2] of the test set with novel subpopulation.

WILDS: For all WILDS datasets, we used the official OOD datasets provided in their paper [4].

D Experiment Setup

We performed training in PyTorch [6], and we used RTX 6000 Ada GPUs.

For datasets without an official validation set, we randomly sampled a subset of the official training set as the validation set to perform calibration and learn thresholds for ATC and COTT. We trained 3 models for each dataset with random seeds $\{0, 1, 10\}$.

CIFAR10 and CIFAR100: We reserved 10000 images from the training set as the validation set. We trained ResNet18 from scratch, using SGD with momentum equal to 0.9 for 300 epochs. We set weight decay to 5×10^{-4} and batch size to 200. We set the initial learning rate to 0.1 and multiply it by 0.1 every 100 epochs.

ImageNet: We reserved 50000 images from the training set as the validation set. We used ResNet50. While ImageNet pretrained weights are available in PyTorch, we needed multiple ones trained using different initializations. Due to limited computation resources, we reused the upper layer weights but reinitialized the last layer with different random seeds. We finetuned the whole model using Adam [3] with a batch size of 64 and a learning rate of 10^{-4}, for 10 epochs.

BREEDS: We used the intersection set of images that are both in the ImageNet validation images we set aside and the BREEDS dataset as the validation set. For all BREEDS datasets (Living17, Nonliving26, Entity13, Entity30), we trained ResNet50 from scratch.

For Living17 and Nonliving26, we used SGD with weight decay of 10^{-4} and batch size of 128. We trained for 450 epochs. We set the initial learning rate to 0.1 and multiplied it by 0.1 every 150 epochs.

For Entity13 and Entity30, we used SGD with weight decay of 10^{-4} and batch size of 128. We trained for 300 epochs. We set the initial learning rate to 0.1 and multiplied it by 0.1 every 100 epochs.

Camelyon17-WILDS: We used the `id_val` group as the validation set. We fine-tuned ImageNet pretrained ResNet50 using SGD with momentum of 0.9, weight decay of 5×10^{-4}, and batch size of 32, for 5 epochs.

RxRx1-WILDS: We used the `id_test` group as the validation set. We followed [4] to fine-tune an ImageNet pretrained ResNet50. We used Adam with weight decay of 10^{-5} and batch size of 75, for 90 epochs. We increased the learning rate from 0 to 10^{-4} linearly for the first 10 epochs and decayed it following a cosine learning rate schedule.

Amazon-WILDS: We used the `id_val` group as the validation set. We followed [4] to fine-tune a DistilBERT-base-uncased model [9]. We used AdamW [5] with weight decay of 10^{-2}, learning rate of 10^{-5}, and batch size of 8, for 3 epochs. We set the maximum number of tokens to 512.

CivilComments-WILDS: We used the `val` group as the validation set. We followed [4] to fine-tune a DistilBERT-base-uncased model [9]. We used AdamW [5] with weight decay of 10^{-2}, learning rate of 10^{-5}, and batch size of 16, for 5 epochs. We set the maximum number of tokens to 300.

References

[1] Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. RLSBench: A large-scale empirical study of domain adaptation under relaxed label shift. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

[2] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[7] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

[8] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[10] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[11] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[12] Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. *arXiv preprint arXiv:2202.05834*, 2022.