Robust Distillation for Worst-class Performance: On the Interplay Between Teacher and Student Objectives (Supplementary Material)

Serena Wang[1,2] **Harikrishna Narasimhan**[1] **Yichen Zhou**[1] **Sara Hooker**[3] **Michal Lukasik**[1]
Aditya Krishna Menon[1]

[1]Google Research, Mountain View, California and New York, New York, USA,
[2]University of California, Berkeley, Berkeley, California, USA,
[3]Cohere For AI, Palo Alto, California, USA

A PROOFS

A.1 PROOF OF THEOREM 1

(i) The first result follows from the fact that the cross-entropy loss is a proper composite loss [Williamson et al., 2016] with the softmax function as the associated (inverse) link function.

(ii) For a proof of the second result, please see Menon et al. [2021b].

(iii) Below, we provide a proof for the third result.

The minimization of the robust objective in (3) over f can be re-written as a min-max optimization problem:

$$\min_{f:\mathcal{X}\to\mathbb{R}^m} L^{\text{rob}}(f) = \min_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \underbrace{\sum_{y=1}^{m} \frac{\lambda_y}{\pi_y}\mathbb{E}\left[\eta_y(X)\,\ell(y, f(X))\right]}_{\omega(\lambda, f)}. \tag{1}$$

The min-max objective $\omega(\lambda, f)$ is clearly linear in λ (for fixed f) and with ℓ chosen to be the cross-entropy loss, is convex in f (for fixed λ), i.e., $\omega(\lambda, \kappa f_1 + (1-\kappa) f_2) \leq \kappa\omega(\lambda, f_1) + (1-\kappa)\omega(\lambda, f_2)$, $\forall f_1, f_2 : \mathcal{X} \to \mathbb{R}^m, \kappa \in [0,1]$. Furthermore, Δ_m is a convex compact set, while the domain of f is convex. It follows from Sion's minimax theorem [Sion, 1958] that:

$$\min_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \omega(\lambda, f) = \max_{\lambda\in\Delta_m} \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda, f). \tag{2}$$

Let (λ^*, f^*) be such that:

$$\lambda^* \in \operatorname*{argmax}_{\lambda\in\Delta_m} \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda, f); \qquad f^* \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \omega(\lambda, f),$$

Such a λ^* exists for the following reason: for any fixed $\lambda \in \Delta_m$, owing to the use of the cross-entropy loss, a minimizer over always exists for $\omega(\lambda, f)$, and is given by $f_y(x) = \log\left(\frac{\lambda_y}{\pi_y}\eta_y(x)\right) + C$, for some $C \in \mathbb{R}$; therefore $\min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda, f)$ is bounded above for any λ, and Δ_m being compact set gives us there exits a maximizer λ^* over this set. Similarly, such an f^* exists for the following reason: the objective $\max_{\lambda\in\Delta_m} \omega(\lambda, f)$ takes a bounded value when $f = \eta$, and any minimizer of $\max_{\lambda\in\Delta_m} \omega(\lambda, f)$ yields a value below that; because $\omega(\lambda, f) \geq 0$ and is convex in f, the minimizer f^* exits.

We then have from (2):

$$\begin{aligned}
\omega(\lambda^*, f^*) &\leq \max_{\lambda\in\Delta_m} \omega(\lambda, f^*) \\
&= \min_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \omega(\lambda, f) \;=\; \max_{\lambda\in\Delta_m} \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda, f) \\
&= \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda^*, f) \;\leq\; \omega(\lambda^*, f^*),
\end{aligned}$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

which tells us that there exists (λ^*, f^*) is a saddle-point for (1), i.e.,

$$\omega(\lambda^*, f^*) = \max_{\lambda \in \Delta_m} \omega(\lambda, f^*) \;=\; \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda^*, f).$$

Consequently, we have:

$$L^{\text{rob}}(f^*) = \max_{\lambda \in \Delta_m} \omega(\lambda, f^*) \;=\; \min_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda \in \Delta_m} \omega(\lambda, f) \;=\; \min_{f:\mathcal{X}\to\mathbb{R}^m} L^{\text{rob}}(f).$$

We thus have that f^* is a minimizer of $L^{\text{rob}}(f)$. Furthermore, because f^* is also a minimizer of $\omega(\lambda^*, f)$ over f, i.e.,

$$f^* \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}^m} \sum_{y=1}^{m} \frac{\lambda^*_y}{\pi_y} \mathbb{E}\left[\eta_y(X)\, \ell(y, f(X))\right],$$

it follows that:

$$\text{softmax}_y(f^*(x)) \;\propto\; \frac{\lambda^*_y}{\pi_y} \eta_y(x).$$

(iv) For the fourth result, we expand the traded-off objective, and re-write it as:

$$\begin{aligned}
L^{\text{tdf}}(f) &= (1-\alpha)L^{\text{bal}}(f) + \alpha L^{\text{rob}}(f) \\
&= (1-\alpha)\frac{1}{m}\sum_{y=1}^{m} \frac{1}{\pi_y}\mathbb{E}\left[\eta_y(X)\,\ell(y, f(X))\right] \;+\; \alpha \max_{\lambda\in\Delta_m} \sum_{y=1}^{m} \frac{\lambda_y}{\pi_y}\mathbb{E}\left[\eta_y(X)\,\ell(y, f(X))\right] \\
&= \max_{\lambda\in\Delta_m} \underbrace{\sum_{y=1}^{m} \left((1-\alpha)\frac{1}{m} + \alpha\lambda_y\right)\frac{1}{\pi_y}\mathbb{E}\left[\eta_y(X)\,\ell(y, f(X))\right].}_{\omega(\lambda, f)}
\end{aligned}$$

For a fixed λ, $\omega(\lambda, f)$ is convex in f (as the loss ℓ is the cross-entropy loss), and for a fixed f, $\omega(\lambda, f)$ is linear in λ. Following the same steps as the proof of (iii), we have that there exists (λ^*, f^*) such that

$$L^{\text{tdf}}(f^*) \;=\; \max_{\lambda\in\Delta_m} \omega(\lambda, f^*) \;=\; \min_{f:\mathcal{X}\to\mathbb{R}^m} L^{\text{tdf}}(f),$$

and

$$f^* \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}^m} \sum_{y=1}^{m} \left((1-\alpha)\frac{1}{m} + \alpha\lambda^*_y\right)\frac{1}{\pi_y}\mathbb{E}\left[\eta_y(X)\,\ell(y, f(X))\right],$$

which, owing to the properties of the cross-entropy loss, then gives us the desired form for f^*.

A.2 PROOF OF THEOREM 2

Proof. Expanding the left-hand side, we have:

$$\begin{aligned}
|\hat{L}^{\text{rob-d}}(f) - L^{\text{rob}}(f)| &\le |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f) + L^{\text{rob-d}}(f) - L^{\text{rob}}(f)| \\
&\le |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + |L^{\text{rob-d}}(f) - L^{\text{rob}}(f)| \\
&= |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + \left|\max_{y\in[m]} \frac{\mathbb{E}_x\left[p^t_y(x)\,\ell(y, f(x))\right]}{\mathbb{E}_x\left[p^t_y(x)\right]} - \max_{y\in[m]} \frac{\mathbb{E}_x\left[\eta_y(x)\,\ell(y, f(x))\right]}{\pi_y}\right| \\
&\le |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + \max_{y\in[m]} \left|\frac{\mathbb{E}_x\left[p^t_y(x)\,\ell(y, f(x))\right]}{\mathbb{E}_x\left[p^t_y(x)\right]} - \frac{\mathbb{E}_x\left[\eta_y(x)\,\ell(y, f(x))\right]}{\pi_y}\right| \\
&\le |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + B \max_{y\in[m]} \mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\mathbb{E}_x\left[p^t_y(x)\right]} - \frac{\eta_y(x)}{\pi_y}\right| \ell(y, f(x))\right] \\
&\le |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + B \max_{y\in[m]} \mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\mathbb{E}_x\left[p^t_y(x)\right]} - \frac{\eta_y(x)}{\pi_y}\right|\right],
\end{aligned}$$

where the second-last step uses Jensen's inequality and the fact that $\ell(y, f(x)) \geq 0$, and the last step uses the fact that $\ell(y, f(x)) \leq B$.

Further expanding the first term,

$$\begin{aligned}
|\hat{L}^{\text{rob-d}}(f) - L^{\text{rob}}(f)| &\leq \left|\max_{y\in[m]} \phi_y(f) \,-\, \max_{y\in[m]} \hat{\phi}_y(f)\right| + B \max_{y\in[m]} \mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\mathbb{E}_x\left[p^t_y(x)\right]} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right] \\
&\leq \max_{y\in[m]} \left|\phi_y(f) \,-\, \hat{\phi}_y(f)\right| + B \max_{y\in[m]} \mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\mathbb{E}_x\left[p^t_y(x)\right]} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right],
\end{aligned}$$

as desired. □

A.3 CALIBRATION OF MARGIN-BASED LOSS

To show that minimizer of the margin-based objective in (9) also minimizes the balanced objective in (6), we state the following general result:

Lemma 1. *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Let*

$$\hat{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}^{\text{mar}}\left(p^t(x_i), f(x_i); \mathbf{c}\right) \tag{3}$$

for some cost vector $\mathbf{c} \in \mathbb{R}^m_+$. Then:

$$\hat{f}_y(x_i) = \log\left(c_y p^t_y(x_i)\right) + C_i, \;\; \forall i \in [n],$$

for some example-specific constant constants $C_i \in \mathbb{R}, \forall i \in [n]$. Furthermore, for any assignment of example weights of $w \in \mathbb{R}^n_+$, $\hat{f}$ is also the minimizer of the weighted objective:

$$\hat{f} \,\in\, \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} w_i \sum_{y=1}^{m} c_y\, p^t_y(x_i)\, \ell\left(y, f(x_i)\right). \tag{4}$$

Proof. Following Menon et al. [2021b] (e.g. proof of Theorem 1), we have that for class probabilities $\mathbf{p} \in \Delta_m$ and costs $\mathbf{c} \in \mathbb{R}^m_+$, the margin-based loss in (9)

$$\mathcal{L}^{\text{mar}}\left(\mathbf{p}, \mathbf{f}; \mathbf{c}\right) = \frac{1}{m}\sum_{y\in[m]} p_y \log\left(1 + \sum_{j\neq y} \exp\left(\log(c_y/c_j) \,-\, (f_y - f_j)\right)\right).$$

is minimized by:

$$f^*_y = \log\left(c_y p_y\right) + C,$$

for any $C > 0$. To see why this is true, note that the above loss can be equivalently written as:

$$\mathcal{L}^{\text{mar}}\left(\mathbf{p}, \mathbf{f}; \mathbf{c}\right) = -\frac{1}{m}\sum_{y\in[m]} p_y \log\left(\frac{\exp\left(f_y - \log(c_y)\right)}{\sum_{j=1}^{m}\exp\left(f_j - \log(c_j)\right)}\right).$$

This the same as the softmax cross-entropy loss with adjustments made to the logits, the minimizer for which is of the form:

$$f^*_y - \log(c_y) = \log\left(p_y\right) + C \quad \text{or} \quad f^*_y = \log\left(c_y p_y\right) + C.$$

It follows that any minimizer $\hat{f}$ of the average margin-based loss in (3) over sample S, would do so point-wise, and therefore

$$\hat{f}_y(x_i) = \log\left(c_y p^t_y(x_i)\right) + C_i, \;\; \forall i \in [n],$$

for some example-specific constant constants $C_i \in \mathbb{R}, \forall i \in [n]$.

To prove the second part, we note that for the minimizer $\hat{f}$ to also minimize the weighted objective:

$$\frac{1}{n}\sum_{i=1}^{n} w_i \sum_{y=1}^{m} c_y \, p_y^t(x_i) \, \ell\left(y, f(x_i)\right),$$

it would also have to do so point-wise for each $i \in [m]$, and so as long the weights w_i are non-negative, it suffices that

$$\hat{f}(x_i) \in \operatorname*{argmin}_{\mathbf{f} \in \mathbb{R}^m} \sum_{y=1}^{m} c_y \, p_y^t(x_i) \, \ell\left(y, f(x_i)\right).$$

This is indeed the case when ℓ is the softmax cross-entropy loss, where the point-wise minimizer for each $i \in [m]$ would be of the form $\text{softmax}_y(f(x)) = c_y p_y^t(x)$, which is satisfied by $\hat{f}$. □

A similar result also holds in the population limit, when (3) and (4) are computed in expectation, and the per-example weighting in (4) is replaced by an arbitrary weighting function $w(x) \in \mathbb{R}_+$. Any scorer of the following form would then minimize both objectives:

$$\hat{f}_y(x) = \log\left(c_y p_y^t(x)\right) + C(x), \;\; \forall x \in \mathcal{X},$$

where $C(x)$ is some example-specific constant.

A.4 PROOF OF PROPOSITION 3

Proposition (Restated). *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Then the final scoring function $\bar{f}^s(x) = \frac{1}{K}\sum_{k=1}^{K} f^k(x)$ output by Algorithm 1 is of the form:*

$$\text{softmax}_j(\bar{f}^s(x)) \propto \bar{\lambda}_j p_j^t(x), \quad \forall j \in [m], \, \forall (x, y) \in S,$$

where $\bar{\lambda}_y = \left(\prod_{k=1}^{K} \lambda_y^k / \pi_y^t\right)^{1/K}$.

Proof. The proof follows from Lemma 1 with the costs $\mathbf{c}$ set to λ^k/π^t for each iteration k. The lemma tells us that each f^k is of the form:

$$f^k(x') = \log\left(\frac{\lambda_y^k}{\pi_y^t} p_y^t(x')\right) + C(x'), \;\; \forall (x', y') \in S,$$

for some example-specific constant $C(x') \in \mathbb{R}$. Consequently, we have that:

$$\bar{f}_y^s(x') = \log(\bar{\lambda}_y p_y^t(x')) + \bar{C}(x'), \;\; \forall (x', y') \in S,$$

where $\bar{\lambda}_y = \left(\prod_{k=1}^{K} \lambda_y^k / \pi_y^t\right)^{1/K}$ and $\bar{C}(x') \in \mathbb{R}$. Applying a softmax to $\bar{f}^s$ results in the desired form. □

A.5 PROOF OF THEOREM 4

Theorem (Restated). *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Suppose ℓ is the softmax cross-entropy loss ℓ^{xent}, $\ell(y, z) \leq B$ and $\max_{y \in [m]} \frac{1}{\pi_y^t} \leq Z$, for some $B, Z > 0$. Furthermore, suppose for any $\delta \in (0, 1)$, the following bound holds on the estimation error in Theorem 2: with probability at least $1 - \delta$ (over draw of $S \sim D^n$), for all $f \in \mathcal{F}$,*

$$\max_{y \in [m]} \left|\phi_y(f) - \hat{\phi}_y(f)\right| \leq \Delta(n, \delta),$$

for some $\Delta(n,\delta) \in \mathbb{R}_+$ *that is increasing in* $1/\delta$*, and goes to 0 as* $n \to \infty$*. Fix* $\delta \in (0,1)$*. Then when the step size* $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$ *and* $n^{\text{val}} \geq 8Z\log(2m/\delta)$*, with probability at least* $1-\delta$ *(over draw of* $S \sim D^n$ *and* $S^{\text{val}} \sim D^{n^{\text{val}}}$*)*

$$L^{\text{rob}}(\bar{f}^s) \leq \min_{f\in\mathcal{F}} L^{\text{rob}}(f) \,+\, \underbrace{2B\max_{y\in[m]}\mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\pi^t_y} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right]}_{\textit{Approximation error}}$$

$$+\, \underbrace{2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2)}_{\textit{Estimation error}} + \underbrace{4BZ\sqrt{\frac{\log(m)}{K}}}_{\textit{EG convergence}}.$$

Before proceeding to the proof, we will find it useful to define:

$$\hat{\phi}^{\text{val}}_y(f^s) = \frac{1}{\hat{\pi}^{t,\text{val}}_y}\frac{1}{n^{\text{val}}}\sum_{(x',y')\in S^{\text{val}}} p^t_y(x')\,\ell\left(y, f^s(x')\right).$$

We then state a useful lemma.

Lemma 2. *Suppose the conditions in Theorem 4 hold. Then with probability* $\leq 1-\delta$ *(over draw of* $S \sim D^n$ *and* $S^{\text{val}} \sim D^{n^{\text{val}}}$*), at each iteration k,*

$$\sum_{y=1}^m \lambda^{k+1}_y \phi_y(f^{k+1}) \,-\, \min_{f\in\mathcal{F}}\sum_{y=1}^m \lambda^{k+1}_y\phi_y(f) \leq 2\Delta(n,\delta);$$

and for any $\lambda \in \Delta_m$*:*

$$\left|\sum_{y=1}^m \lambda_y\hat{\phi}^{\text{val}}_y(f^{k+1}) \,-\, \sum_{y=1}^m \lambda_y\phi_y(f^{k+1})\right| \,\leq\, \Delta(n^{\text{val}},\delta).$$

Proof. We first note that by applying Lemma 1 with $w_i = 1, \forall i$, we have that f^{k+1} is the minimizer of $\sum_{y=1}^m \lambda^{k+1}_y\hat{\phi}_y(f)$ over all $f \in \mathcal{F}$, and therefore:

$$\sum_{y=1}^m \lambda^{k+1}_y\hat{\phi}_y(f^{k+1}) \,\leq\, \sum_{y=1}^m \lambda^{k+1}_y\hat{\phi}_y(f),\; \forall f \in \mathcal{F}. \tag{5}$$

Further, for a fixed iteration k, let us denote $\tilde{f} \in \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{y=1}^m \lambda^{k+1}_y\phi_y(f)$. Then for the first part, we have:

$$\begin{aligned}
&\sum_{y=1}^m \lambda^{k+1}_y\phi_y(f^{k+1}) \,-\, \sum_{y=1}^m \lambda^{k+1}_y\phi_y(\tilde{f}) \\
&\leq \sum_{y=1}^m \lambda^{k+1}_y\phi_y(f^{k+1}) \,-\, \sum_{y=1}^m \lambda^{k+1}_y\hat{\phi}_y(f^{k+1}) \,+\, \sum_{y=1}^m \lambda^{k+1}_y\hat{\phi}_y(f^{k+1}) \,-\, \sum_{y=1}^m \lambda^{k+1}_y\phi_y(\tilde{f}) \\
&\leq \sum_{y=1}^m \lambda^{k+1}_y\phi_y(f^{k+1}) \,-\, \sum_{y=1}^m \lambda^{k+1}_y\hat{\phi}_y(f^{k+1}) \,+\, \sum_{y=1}^m \lambda^{k+1}_y\hat{\phi}_y(\tilde{f}) \,-\, \sum_{y=1}^m \lambda^{k+1}_y\phi_y(\tilde{f}) \\
&\leq 2\sup_{f\in\mathcal{F}}\left|\sum_{y=1}^m \lambda^{k+1}_y\hat{\phi}_y(f) \,-\, \sum_{y=1}^m \lambda^{k+1}_y\phi_y(f)\right| \\
&\leq 2\sup_{f\in\mathcal{F}}\max_{\lambda\in\Delta_m}\left|\sum_{y=1}^m \lambda_y\hat{\phi}_y(f) \,-\, \sum_{y=1}^m \lambda_y\phi_y(f)\right| \\
&\leq 2\sup_{f\in\mathcal{F}}\max_{\lambda\in\Delta_m}\sum_{y=1}^m \lambda_y\left|\hat{\phi}_y(f) \,-\, \phi_y(f)\right| \\
&= 2\sup_{f\in\mathcal{F}}\max_{y\in[m]}\left|\hat{\phi}_y(f) - \phi_y(f)\right|.
\end{aligned}$$

where for the second inequality, we use (5). Applying the generalization bound assumed in Theorem 4, we have with probability $\leq 1 - \delta$ (over draw of $S \sim D^n$), for all iterations $k \in [K]$,

$$\sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(\tilde{f}) \,\leq\, 2\Delta(n, \delta),$$

For the second part, note that for any $\lambda \in \Delta_m$,

$$\begin{aligned}\left|\sum_{y=1}^{m} \lambda_y \hat{\phi}_y^{\text{val}}(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y \phi_y(f^{k+1})\right| &\leq \sum_{y=1}^{m} \lambda_y \left|\hat{\phi}_y^{\text{val}}(f^{k+1}) \,-\, \phi_y(f^{k+1})\right| \\ &\leq \max_{y \in [m]} \left|\hat{\phi}_y^{\text{val}}(f^{k+1}) \,-\, \phi_y(f^{k+1})\right| \\ &\leq \sup_{f \in \mathcal{F}} \max_{y \in [m]} \left|\hat{\phi}_y^{\text{val}}(f) \,-\, \phi_y(f)\right|.\end{aligned}$$

An application of the generalization bound assumed in Theorem 4 to empirical estimates from the validation sample completes the proof. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. Note that because $\min_{y \in [m]} \pi_y^t \geq \frac{1}{Z}$ and $n^{\text{val}} \geq 8Z \log(2m/\delta)$, we have by a direct application of Chernoff's bound (along with a union bound over all m classes) that with probability at least $1 - \delta/2$:

$$\min_{y \in [m]} \hat{\pi}_y^{t,\text{val}} \geq \frac{1}{2Z}, \forall y \in [m]$$

and consequently, $\hat{\phi}_y^{\text{val}}(f) \leq 2BZ, \forall f \in \mathcal{F}$. The boundedness of $\hat{\phi}_y^{\text{val}}$ will then allow us to apply standard convergence guarantees for exponentiated gradient ascent [Shalev-Shwartz et al., 2011]. For $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$, the updates on λ will give us with probability at least $1 - \delta/2$:

$$\max_{\lambda \in \Delta_m} \frac{1}{K} \sum_{k=1}^{K} \sum_{y=1}^{m} \lambda_y \hat{\phi}_y^{\text{val}}(f^k) \,\leq\, \frac{1}{K} \sum_{k=1}^{K} \sum_{y=1}^{m} \lambda_y^k \hat{\phi}_y^{\text{val}}(f^k) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \tag{6}$$

Applying the second part of Lemma 2 to each iteration k, we have with probability at least $1 - \delta$:

$$\max_{\lambda \in \Delta_m} \frac{1}{K} \sum_{k=1}^{K} \sum_{y=1}^{m} \lambda_y \phi_y(f^k) \,\leq\, \frac{1}{K} \sum_{k=1}^{K} \sum_{y=1}^{m} \lambda_y^k \phi_y(f^k) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2),$$

and applying the first part of Lemma 2 to the RHS, we have with the same probability:

$$\begin{aligned}&\max_{\lambda \in \Delta_m} \frac{1}{K} \sum_{k=1}^{K} \sum_{y=1}^{m} \lambda_y \phi_y(f^k) \\ &\leq \frac{1}{K} \sum_{k=1}^{K} \min_{f \in \mathcal{F}} \sum_{y=1}^{m} \lambda_y^k \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2) \\ &\leq \min_{f \in \mathcal{F}} \frac{1}{K} \sum_{k=1}^{K} \sum_{y=1}^{m} \lambda_y^k \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2).\end{aligned}$$

Note that we have taken a union bound over the high probability statement in (6) and that in Lemma 2. Using the convexity of $\phi(\cdot)$ in $f(x)$ and Jensen's inequality, we have that $\sum_{y=1}^{m} \lambda_y \phi_y(\bar{f}^s) \leq \frac{1}{K} \sum_{k=1}^{K} \sum_{y=1}^{m} \lambda_y \phi_y(f^k)$. We use this to further lower bound the LHS in terms of the averaged scoring function $\bar{f}^s(x) = \frac{1}{K} \sum_{k=1}^{K} f^k(x)$:

$$\max_{\lambda \in \Delta_m} \sum_{y=1}^{m} \lambda_y \phi_y(\bar{f}^s)$$

$$
\begin{aligned}
&\leq \min_{f\in\mathcal{F}} \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda_y^k\phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2)\\
&= \min_{f\in\mathcal{F}} \sum_{y=1}^{m}\tilde{\lambda}_y\phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2)\\
&\leq \max_{\lambda\in\Delta_m}\min_{f\in\mathcal{F}} \sum_{y=1}^{m}\lambda_y\phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2)\\
&= \min_{f\in\mathcal{F}}\max_{\lambda\in\Delta_m} \sum_{y=1}^{m}\lambda_y\phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2)\\
&= \min_{f\in\mathcal{F}}\max_{y\in[m]} \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2), \qquad (7)
\end{aligned}
$$

where in the second step $\tilde{\lambda}_y = \frac{1}{K}\sum_{k=1}^{K}\lambda_y^k$; in the fourth step, we swap the 'min' and 'max' using Sion's minimax theorem [Sion, 1958]. We further have from (7),

$$
\max_{y\in[m]} \phi_y(\bar{f}^s) \leq \min_{f\in\mathcal{F}}\max_{y\in[m]} \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2).
$$

In other words,

$$
L^{\text{rob-d}}(\bar{f}^s) \leq \min_{f\in\mathcal{F}} L^{\text{rob-d}}(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2).
$$

To complete the proof, we need to turn this into a guarantee on the original robust objective L^{rob} in (3):

$$
\begin{aligned}
&L^{\text{rob}}(\bar{f}^s)\\
&\leq \min_{f\in\mathcal{F}} L^{\text{rob}}(f) \,+\, 2\max_{f\in\mathcal{F}}\left|L^{\text{rob}}(f) - L^{\text{rob-d}}(f)\right| \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2)\\
&\leq \min_{f\in\mathcal{F}} L^{\text{rob}}(f) \,+\, 2B\max_{y\in[m]}\mathbb{E}_x\left[\left|\frac{p_y^t(x)}{\pi_y^t} - \frac{\eta_y(x)}{\pi_y}\right|\right] \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}},\delta/2) \,+\, 2\Delta(n,\delta/2),
\end{aligned}
$$

where we have used the bound on the approximation error in the proof of Theorem 2. This completes the proof. □

B STUDENT ESTIMATION ERROR

We now provide a bound on the estimation error in Theorem 4 using a generalization bound from Menon et al. [2021a].

Lemma 3. *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a given class of scoring functions. Let $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{X}}$ denote the class of loss functions $v(x,y) = \ell(y, f(x))$ induced by scorers $f \in \mathcal{F}$. Let $\mathcal{M}_n = \mathcal{N}_\infty(\frac{1}{n}, \mathcal{V}, 2n)$ denote the uniform L_∞ covering number for $\mathcal{V}$. Fix $\delta \in (0,1)$. Suppose $\ell(y,z) \leq B$, $\pi_y^t \leq \frac{1}{Z}, \forall y \in [m]$, and the number of samples $n \geq 8Z\log(4m/\delta)$. Then with probability $\geq 1-\delta$ over draw of $S \sim D^n$, for any $f \in \mathcal{F}$ and $y \in [m]$:*

$$
\left|\phi_y(f) - \hat{\phi}_y(f)\right| \leq CZ\left(\sqrt{\mathbb{V}_{n,y}(f)\frac{\log(m\mathcal{M}_n/\delta)}{n}} + \frac{\log(m\mathcal{M}_n/\delta)}{n} + B\sqrt{\frac{\log(m/\delta)}{n}}\right),
$$

where $\mathbb{V}_{n,y}(f)$ denotes the empirical variance of the loss values $\{p_y^t(x_i)\cdot\ell(y, f(x_i))\}_{i=1}^n$ for class y, and $C > 0$ is a distribution-independent constant.

Notice the dependence on the *variance* that the teacher's predictions induce on the loss. This suggests that the lower the variance in the teacher's predictions, the better is the student's generalization. Similar to Menon et al. [2021a], one can further show that when the teacher closely approximates the Bayes-probabilities $\eta(x)$, the distilled loss $p_y^t(x_i)\cdot\ell(y, f(x_i))$ has a lower empirical variance that the loss $\ell(y_i, f(x_i))$ computed from one-hot labels.

Proof of Lemma 3. We begin by defining the following intermediate term:

$$\tilde{\phi}_y(f) = \frac{1}{\pi^t_y}\frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell\left(y, f(x_i)\right).$$

Then for any $y \in [m]$,

$$\left|\phi_y(f) - \hat{\phi}_y(f)\right| \le \left|\phi_y(f) - \tilde{\phi}_y(f)\right| + \left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right|. \tag{8}$$

We next bound each of the terms in (8), starting with the first term:

$$\begin{aligned}\left|\phi_y(f) - \tilde{\phi}_y(f)\right| &= \frac{1}{\pi^t_y}\left|\mathbb{E}_x\left[p^t_y(x)\,\ell\left(y, f(x)\right)\right] - \frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell\left(y, f(x_i)\right)\right|\\ &\le Z\left|\mathbb{E}_x\left[p^t_y(x)\,\ell\left(y, f(x)\right)\right] - \frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell\left(y, f(x_i)\right)\right|,\end{aligned}$$

where we use the fact that $\pi^t_y \le \frac{1}{Z}, \forall y$. Applying the generalization bound from Menon et al. [2021a, Proposition 2], along with a union bound over all m classes, we have with probability at least $1 - \delta/2$ over the draw of $S \sim D^n$, for all $y \in [m]$:

$$\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \le C'Z\left(\sqrt{\mathbb{V}_{n,y}(f)\frac{\log(m\mathcal{M}_n/\delta)}{n}} + \frac{\log(m\mathcal{M}_n/\delta)}{n}\right), \tag{9}$$

for a distribution-independent constant $C' > 0$.

We next bound the second term in (8):

$$\begin{aligned}\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| &= \left|\frac{1}{\pi^t_y} - \frac{1}{\hat{\pi}^t_y}\right|\frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\cdot\ell\left(y, f(x_i)\right)\\ &\le B\left|\frac{1}{\pi^t_y} - \frac{1}{\hat{\pi}^t_y}\right|\\ &= \frac{B}{\pi^t_y\hat{\pi}^t_y}\left|\pi^t_y - \hat{\pi}^t_y\right|,\end{aligned}$$

where in the second step we use the fact that $\ell(y, f(x)) \le B$ and $p^t_y(x) \le 1$.

Further note that because $\min_{y\in[m]}\pi^t_y \ge \frac{1}{Z}$ and $n \ge 8Z\log(4m/\delta)$, we have by a direct application of Chernoff's bound (and a union bound over m classes) that with probability at least $1 - \delta/4$:

$$\min_{y\in[m]}\hat{\pi}^t_y \ge \frac{1}{2Z}, \forall y \in [m]. \tag{10}$$

Therefore for any $y \in [m]$:

$$\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \le 2BZ^2\left|\pi^t_y - \hat{\pi}^t_y\right|.$$

Conditioned on the above statement, a simple application of Hoeffding's inequality and a union bound over all $y \in [m]$ gives us that with probability at least $1 - \delta/4$ over the draw of $S \sim D^n$, for all $y \in [m]$:

$$\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \le 2BZ^2\left(\frac{1}{Z}\sqrt{\frac{\log(8m/\delta)}{2n}}\right) = 2BZ\sqrt{\frac{\log(8m/\delta)}{2n}}. \tag{11}$$

A union bound over the high probability statements in (9–11) completes the proof. To see this, note that, for any $\epsilon > 0$ and $y \in [m]$,

$$\mathbb{P}\left(\left|\phi_y(f) - \hat{\phi}_y(f)\right| \ge \epsilon\right)$$

Algorithm 1 Distilled Margin-based DRO with One-hot Validation Labels

```
Inputs: Teacher $p^t$, Student hypothesis class $\mathcal{F}$, Training set $S$, Validation set $S^{\text{val}}$, Step-size $\gamma \in \mathbb{R}_+$, Number of iterations $K$, Loss $\ell$
Initialize: Student $f^0 \in \mathcal{F}$, Multipliers $\lambda^0 \in \Delta_m$
For $k = 0$ to $K - 1$
    $\tilde{\lambda}_j^{k+1} = \lambda_j^k \exp\left(\gamma \hat{R}_j\right), \forall j \in [m]$
            where $\hat{R}_j = \frac{1}{n^{\text{val}}} \frac{1}{\hat{\pi}_j^{\text{val}}} \sum_{(x,y) \in S^{\text{val}}} \ell(y, f^k(x))$ and $\hat{\pi}_j^{\text{val}} = \frac{1}{n^{\text{val}}} \sum_{(x,y) \in S^{\text{val}}} \mathbf{1}(y = j)$
    $\lambda_y^{k+1} = \frac{\tilde{\lambda}_y^{k+1}}{\sum_{j=1}^m \tilde{\lambda}_j^{k+1}}, \forall y$
    $f^{k+1} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}^{\text{mar}}\left(p^t(x_i), f(x_i); \frac{\lambda^{k+1}}{\hat{\pi}^t}\right)$   // Replaced with a few steps of SGD
End For
Output: $\bar{f}^s : x \mapsto \frac{1}{K} \sum_{k=1}^K f^k(x)$
```

$$
\begin{aligned}
&\leq \mathbb{P}\left(\left(\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \geq \epsilon\right) \vee \left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \epsilon\right)\right) \\
&\leq \mathbb{P}\left(\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \geq \epsilon\right) + \mathbb{P}\left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \epsilon\right) \\
&\leq \mathbb{P}\left(\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \geq \epsilon\right) + \mathbb{P}\left(\hat{\pi}_y^t \leq \frac{1}{Z}\right) \cdot \mathbb{P}\left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \epsilon \,\middle|\, \hat{\pi}_y^t \leq \frac{1}{Z}\right) \\
&\qquad\qquad + \mathbb{P}\left(\hat{\pi}_y^t \geq \frac{1}{Z}\right) \cdot \mathbb{P}\left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \epsilon \,\middle|\, \hat{\pi}_y^t \geq \frac{1}{Z}\right) \\
&\leq \mathbb{P}\left(\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \geq \epsilon\right) + \mathbb{P}\left(\hat{\pi}_y^t \leq \frac{1}{Z}\right) + \mathbb{P}\left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \epsilon \,\middle|\, \hat{\pi}_y^t \geq \frac{1}{Z}\right),
\end{aligned}
$$

which implies that a union bound over (9–11) would give us the desired result in Lemma 3. □

C DRO WITH ONE-HOT VALIDATION LABELS

The updates on λ in Algorithm 1 use a validation set labeled by the teacher. One could instead perform these updates with a curated validation set containing the original one-hot labels. Each of these choices presents different merits. The use of a teacher-labeled validation set is useful in many real world scenarios where labeled data is hard to obtain, while unlabeled data abounds. In contrast, the use of one-hot validation labels, although more expensive to obtain, may make the student more immune to errors in the teacher's predictions, as the coefficients λs are now based on an unbiased estimate of the student's performance on each class.

Algorithm 1 contains a version of the margin-based DRO described in Section 3.1, where instead of teacher labels the original one-hot labels are used in the validation set.

Before proceeding to providing a convergence guarantee for this algorithm, we will find it useful to define the following one-hot metrics:

$$
\begin{aligned}
\phi_y^{\text{oh}}(f^s) &= \frac{1}{\pi_y} \mathbb{E}_x \left[\eta_y(x)\, \ell\left(y, f^s(x)\right)\right] \\
\hat{\phi}_y^{\text{oh,val}}(f^s) &= \frac{1}{\hat{\pi}_y} \frac{1}{n^{\text{val}}} \sum_{(x',y') \in S^{\text{val}}} \mathbf{1}(y' = y)\, \ell\left(y', f^s(x')\right).
\end{aligned}
$$

Theorem 4. *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Then the final scoring function $\bar{f}^s(x) = \frac{1}{K}\sum_{k=1}^K f^k(x)$ output by Algorithm 1 is of the form:*

$$
\text{softmax}_y(\bar{f}^s(x')) \propto \bar{\lambda}_y p_y^t(x'), \quad \forall (x', y') \in S,
$$

where $\bar{\lambda}_y = \left(\prod_{k=1}^{K} \lambda^k_y / \pi^t_y\right)^{1/K}$*. Furthermore, suppose* ℓ *is the softmax cross-entropy loss in* ℓ^{xent}*,* $\ell(y, z) \leq B$*, for some* $B > 0$*, and* $\max_{y \in [m]} \frac{1}{\pi_y} \leq Z$*, for some* $Z > 0$*. Suppose for any* $\delta \in (0, 1)$*, the following holds: with probability at least* $1 - \delta$ *(over draw of* $S \sim D^n$*), for all* $f \in \mathcal{F}$*,*

$$\max_{y \in [m]} \left|\phi^{\text{oh}}_y(f) - \hat{\phi}^{\text{oh}}_y(f)\right| \leq \Delta^{\text{oh}}(n, \delta); \qquad \max_{y \in [m]} \left|\phi_y(f) - \hat{\phi}_y(f)\right| \leq \Delta(n, \delta),$$

for some $\Delta^{\text{oh}}(n, \delta), \Delta(n, \delta) \in \mathbb{R}_+$ *that is increasing in* $1/\delta$*, and goes to 0 as* $n \to \infty$*. Fix* $\delta \in (0, 1)$*. Then when the step size* $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$ *and* $n^{\text{val}} \geq 8Z \log(2m/\delta)$*, with probability at least* $1 - \delta$ *(over draw of* $S \sim D^n$ *and* $S^{\text{val}} \sim D^{n^{\text{val}}}$*), for any* $\tau \in \mathbb{R}_+$*,*

$$L^{\text{rob}}(\bar{f}^s) \leq \min_{f \in \mathcal{F}} L^{\text{rob}}(f) \;+\; \underbrace{2B \max_{y \in [m]} \mathbb{E}_x\left[\left|\tau \cdot \frac{p^t_y(x)}{\pi^t_y} \;-\; \frac{\eta_y(x)}{\pi_y}\right|\right]}_{\textit{Approximation error}}$$

$$+ \;\underbrace{2\tau \cdot \Delta^{\text{oh}}(n^{\text{val}}, \delta/2) \;+\; 2\Delta(n, \delta/2)}_{\textit{Estimation error}} \;+\; \underbrace{4BZ\sqrt{\frac{\log(m)}{K}}}_{\textit{EG convergence}}.$$

Comparing this to the bound in Theorem 4, we can see that there is an additional scaling factor τ against the teacher probabilities $p^t_y(x)$ and in the approximation error. When we set $\tau = 1$, the bound looks very similar to Theorem 4, except that the estimation error term Δ^{oh} now involves one-hot labels. Therefore the estimation error may incur a slower convergence with sample size as it no longer benefits from the lower variance that the teacher predictions may offer (see Appendix B for details).

The τ-scaling in the approximation error also means that the teacher is no longer required to exactly match the (normalized) class probabilities $\eta(x)$. In fact, one can set τ to a value for which the approximation error is the lowest, and in general to a value that minimizes the upper bound in Theorem 4, potentially providing us with a tighter convergence rate than Theorem 4.

The proof of Theorem 4 is similar to that of Theorem 4, but requires a modified version of Lemma 2:

Lemma 5. *Suppose the conditions in Theorem 4 hold. With probability* $\leq 1 - \delta$ *(over draw of* $S \sim D^n$ *and* $S^{\text{val}} \sim D^{n^{\text{val}}}$*), at each iteration* k *and for any* $\tau \in \mathbb{R}_+$*,*

$$\sum_{y=1}^{m} \lambda^{k+1}_y \phi^{\text{oh}}_y(f^{k+1}) \;-\; \min_{f \in \mathcal{F}} \sum_{y=1}^{m} \lambda^{k+1}_y \phi^{\text{oh}}_y(f) \leq 2\tau \cdot \Delta(n, \delta) \;+\; 2B \max_{y \in [m]} \mathbb{E}_x\left[\left|\tau \frac{p^t_y(x)}{\pi^t_y} \;-\; \frac{\eta_y(x)}{\pi_y}\right|\right].$$

Furthermore, with the same probability, for any $\lambda \in \Delta_m$*:*

$$\left|\sum_{y=1}^{m} \lambda_y \hat{\phi}^{\text{oh,val}}_y(f^{k+1}) \;-\; \sum_{y=1}^{m} \lambda_y \phi^{\text{oh}}_y(f^{k+1})\right| \;\leq\; \Delta^{\text{oh}}(n^{\text{val}}, \delta).$$

Proof. We first note from Lemma 1 that because $f^{k+1} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}^{\text{mar}}\Big(p^t(x_i), f(x_i); \frac{\lambda^{k+1}}{\hat{\pi}}\Big)$, we have for the example-weighting $w_i = \tau, \forall i$:

$$\tau \sum_{y=1}^{m} \lambda^{k+1}_y \hat{\phi}_y(f^{k+1}) \;\leq\; \tau \sum_{y=1}^{m} \lambda^{k+1}_y \hat{\phi}_y(f), \;\forall f \in \mathcal{F}. \tag{12}$$

For a fixed iteration k, let us denote $\tilde{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{y=1}^{m} \lambda^{k+1}_y \phi_y(f)$. Then for the first part, we have for any $\tau \in \mathbb{R}_+$:

$$\sum_{y=1}^{m} \lambda^{k+1}_y \phi^{\text{oh}}_y(f^{k+1}) \;-\; \sum_{y=1}^{m} \lambda^{k+1}_y \phi^{\text{oh}}_y(\tilde{f})$$

$$
\begin{aligned}
&\le \tau\left(\sum_{y=1}^{m} \lambda_y^{k+1}\phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1}\phi_y(\tilde{f})\right) \,+\, \sum_{y=1}^{m} \lambda_y^{k+1}\left|\phi_y^{\text{oh}}(f^{k+1}) - \tau\phi_y(f^{k+1})\right| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \sum_{y=1}^{m} \lambda_y^{k+1}\left|\phi_y^{\text{oh}}(\tilde{f}) - \tau\phi_y(\tilde{f})\right| \\
&\le \tau\left(\sum_{y=1}^{m} \lambda_y^{k+1}\phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1}\phi_y(\tilde{f})\right) \,+\, 2\max_{f\in\mathcal{F}} \sum_{y=1}^{m} \lambda_y^{k+1}\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right| \\
&\le \tau\left(\sum_{y=1}^{m} \lambda_y^{k+1}\phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1}\phi_y(\tilde{f})\right) \,+\, 2\max_{f\in\mathcal{F}}\max_{\lambda\in\Delta_m} \sum_{y=1}^{m} \lambda\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right| \\
&\le \tau\left(\sum_{y=1}^{m} \lambda_y^{k+1}\phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1}\phi_y(\tilde{f})\right) \,+\, 2\max_{f\in\mathcal{F}}\max_{y\in[m]} \left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right| \\
&\le 2\tau \sup_{f\in\mathcal{F}}\max_{y\in[m]} \left|\hat{\phi}_y(f) - \phi_y(f)\right| \,+\, 2\max_{f\in\mathcal{F}}\max_{y\in[m]} \left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right|.
\end{aligned}
$$

where the last inequality re-traces the steps in Lemma 2. Further applying the generalization bound assumed in Theorem 4, we have with probability $\le 1-\delta$ (over draw of $S \sim D^n$), for all iterations $k \in [K]$ and any $\tau \in \mathbb{R}_+$,

$$
\sum_{y=1}^{m} \lambda_y^{k+1}\phi_y^{\text{oh}}(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1}\phi_y^{\text{oh}}(\tilde{f}) \;\le\; 2\tau\Delta(n,\delta) + 2\max_{f\in\mathcal{F}}\max_{y\in[m]} \left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right|. \tag{13}
$$

All that remains is to bound the second term in (13). For any $f \in \mathcal{F}$ and $y \in [m]$,

$$
\begin{aligned}
\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right| &\le \left|\frac{1}{\pi_y}\mathbb{E}_x\left[\eta_y(x)\,\ell\left(y, f(x)\right)\right] \,-\, \frac{\tau}{\pi_y^t}\mathbb{E}_x\left[p_y^t(x)\,\ell\left(y, f(x)\right)\right]\right| \\
&\le \mathbb{E}_x\left[\left|\frac{1}{\pi_y}\eta_y(x)\,\ell\left(y, f(x)\right) \,-\, \frac{\tau}{\pi_y^t}p_y^t(x)\,\ell\left(y, f(x)\right)\right|\right] \\
&= \mathbb{E}_x\left[\left|\frac{1}{\pi_y}\eta_y(x) \,-\, \frac{\tau}{\pi_y^t}p_y^t(x)\right|\ell\left(y, f^s(x)\right)\right] \\
&\le B\mathbb{E}_x\left[\left|\frac{\eta_y(x)}{\pi_y} \,-\, \tau\frac{p_y^t(x)}{\pi_y^t}\right|\right],
\end{aligned}
$$

where we use Jensen's inequality in the second step, the fact that $\ell(y,z) \le B$ is non-negative in the second step, and the fact that $\ell(y,z) \le B$ in the last step. Substituting this upper bound back into (13) completes the proof of the first part.

The second part follows from a direct application of the bound on the per-class estimation error $\max_{y\in[m]}\left|\phi_y^{\text{oh}}(f) - \hat{\phi}_y^{\text{oh,val}}(f)\right|$. □

Proof of Theorem 4. The proof traces the same steps as Proposition 3 and Theorem 4, except that it applies Lemma 5 instead of Lemma 2.

Note that because $\min_{y\in[m]}\pi_y \ge \frac{1}{Z}$ and $n^{\text{val}} \ge 8Z\log(2m/\delta)$, we have by a direct application of Chernoff's bound (along with a union bound over all m classes) that with probability at least $1-\delta/2$:

$$
\min_{y\in[m]} \hat{\pi}_y^{\text{oh,val}} \ge \frac{1}{2Z}, \forall y \in [m],
$$

and consequently, $\hat{\phi}_y^{\text{oh,val}}(f) \le 2BZ, \forall f \in \mathcal{F}$. The boundedness of $\hat{\phi}_y^{\text{oh,val}}$ will then allow us to apply standard convergence guarantees for exponentiated gradient ascent [Shalev-Shwartz et al., 2011]. For $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$, the updates on λ will give us:

$$
\max_{\lambda\in\Delta_m} \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m} \lambda_y \hat{\phi}_y^{\text{oh,val}}(f^k) \;\le\; \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m} \lambda_y^k \hat{\phi}_y^{\text{oh,val}}(f^k) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}}
$$

Algorithm 2 Distilled Margin-based DRO for Traded-off Objective

```
**Inputs:** Teacher $p^t$, Student hypothesis class $\mathcal{F}$, Training set $S$, Validation set $S^{\text{val}}$, Step-size $\gamma \in \mathbb{R}_+$, Number of iterations $K$, Loss $\ell$, Trade-off parameter $\alpha$
**Initialize:** Student $f^0 \in \mathcal{F}$, Multipliers $\lambda^0 \in \Delta_m$
**For** $k = 0$ to $K - 1$
  $\tilde{\lambda}_j^{k+1} \,=\, \lambda_j^k \exp\left(\gamma\alpha\hat{R}_j\right), \forall j \in [m]$  where $\hat{R}_j = \frac{1}{n^{\text{val}}}\frac{1}{\hat{\pi}_j^{t,\text{val}}}\sum_{(x,y)\in S^{\text{val}}} p_j^t(x_i)\,\ell(j, f^k(x))$
  $\lambda_y^{k+1} \,=\, \frac{\tilde{\lambda}_y^{k+1}}{\sum_{j=1}^m \tilde{\lambda}_j^{k+1}}, \forall y$
  $\beta_y^{k+1} \,=\, (1-\alpha)\frac{1}{m} \,+\, \alpha\lambda_y^{k+1}$
  $f^{k+1} \,\in\, \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathcal{L}^{\text{mar}}\left(p^t(x_i), f(x_i); \frac{\beta^{k+1}}{\hat{\pi}^t}\right)$   // Replaced with a few steps of SGD
**End For**
**Output:** $\bar{f}^s : x \mapsto \frac{1}{K}\sum_{k=1}^K f^k(x)$
```

Applying the second part of Lemma 2 to each iteration k, we have with probability at least $1 - \delta$:

$$\max_{\lambda\in\Delta_m} \frac{1}{K}\sum_{k=1}^K\sum_{y=1}^m \lambda_y\phi_y^{\text{oh}}(f^k) \,\le\, \frac{1}{K}\sum_{k=1}^K\sum_{y=1}^m \lambda_y^k\phi_y^{\text{oh}}(f^k) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta^{\text{oh}}(n^{\text{val}}, \delta/2),$$

and applying the first part of Lemma 2 to the RHS, we have with the same probability, for any $\tau \in \mathbb{R}_+$:

$$\begin{aligned}
\max_{\lambda\in\Delta_m} \frac{1}{K}\sum_{k=1}^K\sum_{y=1}^m \lambda_y\phi_y^{\text{oh}}(f^k) &\le \frac{1}{K}\sum_{k=1}^K \min_{f\in\mathcal{F}}\sum_{y=1}^m \lambda_y^k\phi_y^{\text{oh}}(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta^{\text{oh}}(n^{\text{val}}, \delta/2) \\
&\qquad + \, 2\tau\Delta(n, \delta/2) \,+\, 2B\max_{y\in[m]}\mathbb{E}_x\left[\left|\tau\frac{p_y^t(x)}{\pi_y^t} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right] \\
&\le \min_{f\in\mathcal{F}} \frac{1}{K}\sum_{k=1}^K\sum_{y=1}^m \lambda_y^k\phi_y^{\text{oh}}(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta^{\text{oh}}(n^{\text{val}}, \delta/2) \\
&\qquad + \, 2\tau\Delta(n, \delta/2) \,+\, 2B\max_{y\in[m]}\mathbb{E}_x\left[\left|\tau\frac{p_y^t(x)}{\pi_y^t} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right].
\end{aligned}$$

Using the convexity of $\phi(\cdot)$ in $f(x)$ and Jensen's inequality, we have that $\sum_{y=1}^m \lambda_y\phi_y(\bar{f}^s) \le \frac{1}{K}\sum_{k=1}^K\sum_{y=1}^m \lambda_y\phi_y(f^k)$. We use this to further lower bound the LHS in terms of the averaged scoring function $\bar{f}^s(x) = \frac{1}{K}\sum_{k=1}^K f^k(x)$, and re-trace the steps in Theorem 4 to get"

$$\begin{aligned}
\max_{y\in[m]} \phi_y^{\text{oh}}(\bar{f}^s) &\le \min_{f\in\mathcal{F}}\max_{y\in[m]} \phi_y^{\text{oh}}(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta^{\text{oh}}(n^{\text{val}}, \delta/2) \\
&\qquad + \, 2\tau\Delta(n, \delta/2) \,+\, 2B\max_{y\in[m]}\mathbb{E}_x\left[\left|\tau\frac{p_y^t(x)}{\pi_y^t} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right].
\end{aligned}$$

Noting that $L^{\text{rob}}(f) = \max_{y\in[m]} \phi_y^{\text{oh}}(f)$ completes the proof. □

D DRO FOR TRADED-OFF OBJECTIVE

We present a variant of the margin-based DRO algorithm described in Section 3.1 that seeks to minimize a trade-off between the balanced and robust student objectives:

$$\hat{L}^{\text{tdf-d}}(f^s) = (1-\alpha)\hat{L}^{\text{bal-d}}(f^s) + \alpha\hat{L}^{\text{rob-d}}(f^s),$$

for some $\alpha \in [0, 1]$.

Expanding this, we have:

$$\begin{aligned}
L^{\text{tdf-d}}(f) &= (1-\alpha)\frac{1}{m}\sum_{y=1}^{m}\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell(y, f(x_i)) \;+\; \alpha \max_{y\in[m]} \sum_{y=1}^{m}\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell(y, f(x_i)) \\
&= (1-\alpha)\frac{1}{m}\sum_{y=1}^{m}\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell(y, f(x_i)) \;+\; \alpha \max_{\lambda\in\Delta_m} \sum_{y=1}^{m}\frac{\lambda_y}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell(y, f(x_i)) \\
&= \max_{\lambda\in\Delta_m} \sum_{y=1}^{m}\left((1-\alpha)\frac{1}{m} + \alpha\lambda_y\right)\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell(y, f(x_i)).
\end{aligned}$$

The minimization of $L^{\text{tdf-d}}(f)$ over f can then be a cast as a min-max problem:

$$\min_{f:\mathcal{X}\to\mathbb{R}^m} L^{\text{tdf-d}}(f) = \min_{f:\mathcal{X}\to\mathbb{R}^m}\max_{\lambda\in\Delta_m} \sum_{y=1}^{m}\left((1-\alpha)\frac{1}{m} + \alpha\lambda_y\right)\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)\,\ell(y, f(x_i)).$$

Retracing the steps in the derivation of Algorithm 1 in Section 3.1, we have the following updates on λ and f to solve the above min-max problem:

$$\begin{aligned}
\tilde{\lambda}^{k+1}_y &= \lambda^k_y \exp\left(\gamma\alpha\frac{1}{n\hat{\pi}^t_y}\sum_{i=1}^{n} p^t_y(x_i)\,\ell\left(y, f^k(x_i)\right)\right), \forall y \\
\lambda^{k+1}_y &= \frac{\tilde{\lambda}^{k+1}_y}{\sum_{j=1}^{m}\tilde{\lambda}^{k+1}_j}, \forall y \\
\beta^{k+1}_y &= (1-\alpha)\frac{1}{m} \,+\, \alpha\lambda^{k+1}_y \\
f^{k+1} &\in \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{y\in[m]}\frac{\beta^{k+1}_y}{n\hat{\pi}^t_y}\sum_{i=1}^{n} p^t_y(x_i)\,\ell\left(y, f(x_i)\right),
\end{aligned}$$

for step-size parameter $\gamma > 0$. To better handle training of over-parameterized students, we will perform the updates on λ using a held-out validation set, and employ a margin-based surrogate for performing the minimization over f. This procedure is outlined in Algorithm 2.

D.1 CONNECTION TO POST-HOC ADJUSTMENT

The form of the student in Proposition 3 raises an interesting question. Instead of training an explicit student model, why not directly construct a new scoring model by making post-hoc adjustments to the teacher's predictions? Specifically, one could optimize over functions of the form $f^s_y(x) = \log(\gamma_y p^t_y(x))$, where the teacher p^t is fixed, and pick the coefficients $\gamma \in \mathbb{R}^m$ so that resulting scoring function yields the best worst-class accuracy on a held-out dataset. This simple *post-hoc adjustment* strategy may not be feasible if the goal is to distill to a student that is considerably smaller than the teacher. Often, this is the case in settings where distillation is used as a compression technique. Yet, this post-hoc method serves as good baseline to compare with.

E ADDITIONAL EXPERIMENT DETAILS

This section contains further experiment details about the datasets, hyperparameters, and baselines.

E.1 ADDITIONAL DETAILS ABOUT DATASETS

E.1.1 Building long tailed datasets

The long-tailed datasets were created from the original datasets following Cui et al. [2019] by downsampling examples with an exponential decay in the per-class sizes. As done by Narasimhan and Menon [2021], we set the imbalance ratio

$\frac{\max_i P(y=i)}{\min_i P(y=i)}$ to 100 for CIFAR-10 and CIFAR-100, and to 83 for TinyImageNet (the slightly smaller ratio is to ensure that the smallest class is of a reasonable size). We use the long-tail version of ImageNet generated by Liu et al. [2017].

E.1.2 Dataset splits

The original test samples for CIFAR-10, CIFAR-10-LT, CIFAR-100, CIFAR-100-LT, TinyImageNet (200 classes), TinyImageNet-LT (200 classes), and ImageNet (1000 classes) are all balanced. Following Narasimhan and Menon [2021], we randomly split them in half and use half the samples as a validation set, and the other half as a test set. For the CIFAR and TineImageNet datasets, this amounts to using a validation set of size 5000. For the ImageNet dataset, we sample a subset of 5000 examples from the validation set each time we update the Lagrange multipliers in Algorithm 1.

In keeping with prior work Menon et al. [2021b], Narasimhan and Menon [2021], Lukasik et al. [2022], we use the same validation and test sets for the long-tailed training sets as we do for the original versions. For the long tailed training sets, this simulates a scenario where the training data follows a long tailed distribution due to practical data collection limitations, but the test distribution of interest still comes from the original data distribution. In plots, the "balanced accuracy" that we report for the long-tail datasets (e.g., CIFAR-10-LT) is actually the standard accuracy calculated over the balanced test set, which is shared with the original balanced dataset (e.g., CIFAR-10).

Both teacher and student were always trained on the same training set.

The CIFAR datasets had images of size 32 × 32, while the TinyImageNet and ImageNet datasets dataset had images of size 224 × 224.

These datasets do not contain personally identifiable information or offensive content. The CIFAR-10 and CIFAR-100 datasets are licensed under the MIT License. The terms of access for ImageNet are given at `https://www.image-net.org/download.php`.

E.2 ADDITIONAL DETAILS ABOUT TRAINING AND HYPERPARAMETERS

E.2.1 Training details and hyperparameters

Temperature hyperparameters. We apply temperature scaling to the teacher scores on both the training set and validation set when training the student, i.e., compute $p^t(x) = \text{softmax}(f^t(x)/\gamma)$, and vary the temperature parameter γ over a range of $\{1, 3, 5\}$. When training with teacher labels on the validation set (Algorithm 1), we vary the temperature parameters independently for the training set and the validation set. That is, we apply $p^t(x) = \text{softmax}(f^t(x)/\gamma_{\text{train}})$ over the training set and $p^t(x) = \text{softmax}(f^t(x)/\gamma_{\text{val}})$ over the validation set. When teacher labels are applied to the validation set, we additionally include a temperature of 0.1 on the teacher's validation set labels to approximate a hard thresholding of the teacher probabilities. Thus, the final hyperparameter search spaces are $\gamma_{\text{train}} \in \{1, 3, 5\}$, and $\gamma_{\text{val}} \in \{0.1, 1, 3, 5\}$.

Unless otherwise specified, in all tables, the temperature hyperparameters were chosen to achieve the best worst-class accuracy on the validation set. In all scatter plots such as Figure 1, for each α^t, α^s combination, temperature hyperparameters were selected to achieve the best worst-class accuracy on the validation set.

Learning rate hyperparameters. All models were trained using SGD with momentum of 0.9 [Lukasik et al., 2022, Narasimhan and Menon, 2021].

The learning rate schedule were chosen to mimic the settings in prior work Narasimhan and Menon [2021], Lukasik et al. [2022]. For CIFAR-10 and CIFAR-100 datasets, we ran the optimizer for 450 epochs, linearly warming up the learning rate till the 15th epoch, and then applied a step-size decay of 0.1 after the 200th, 300th and 400th epochs, as done by Lukasik et al. [2022]. For the long-tail versions of these datasets, we trained for 256 epochs, linearly warming up the learning rate till the 15th epoch, and then applied a step-size decay of 0.1 after the 96th, 192nd and 224th epochs, as done by Narasimhan and Menon [2021]. Similarly, for the TinyImageNet datasets, we train for 200 epochs, linearly warming up the learning rate till the 5th epoch, and then applying a decay of 0.1 after the 75th and 135th epochs, as done by Narasimhan and Menon [2021]. For ImageNet, we train for 90 epochs, linearly warming up the learning rate till the 5th epoch, then applying a decay of 0.1 after the 30th, 60th and 80th epochs, as done by Lukasik et al. [2022]. We used a batch size of 128 for the CIFAR-10 and the long-tailed TinyImageNet datasets [Narasimhan and Menon, 2021], a batch size of 512 for the balanced ImageNet dataset, a batch size of 2048 for the balanced TinyImageNet dataset, and a batch size of 1024 for other datasets Lukasik et al. [2022].

We apply an L_2 weight decay of 10^{-4} in all our SGD updates Lukasik et al. [2022]. This amounts to applying an L_2 *regularization* on the model parameters, and has the effect of keeping the model parameters (and as a result the loss function) bounded.

When training with the margin-based robust objective (see Algorithm 1), a separate step size α was applied for training the main model function f, and for updating the multipliers λ. We set α to 0.1 in all experiments.

Hardware. Model training was done using TPUv2.

E.2.2 Repeats

For all comparative baselines without distillation (Group DRO, Post shift, and all teachers alone), we provide average results over m retrained models ($m = 5$ for ImageNet / TinyImageNet, or $m = 10$ for CIFAR datasets). For students on all CIFAR* datasets, unless otherwise specified, we train the teacher once and run the student training 10 times using the same arbitrarily chosen fixed teacher. We compute the mean and standard error of metrics over these $m = 10$ runs. For the resource-heavy TinyImageNet and ImageNet students, we reduce the number of repeats to $m = 5$. This methodology captures variation in the student retrainings while holding the teacher fixed. To capture the end-to-end variation in both teacher and student training, we include Appendix F.4 and Table 3 which contains a rerun of the CIFAR experiments in Table 1 using a distinct teacher for each student retraining. The overall best teacher/student objective combinations did not change for most datasets, with the only exception coming from a difference in the use of validation set labels.

E.3 ADDITIONAL DETAILS ABOUT ALGORITHMS AND BASELINES

E.3.1 Practical improvements to Algorithms 1–2

Algorithms 1–2 currently return a scorer that averages over all K iterates $\bar{f}^s(x) = \frac{1}{K}\sum_{k=1}^{K} f^k(x)$. While this averaging was required for our theoretical robustness guarantees to hold, in our experiments, we find it sufficient to simply return the last model f^K. Another practical improvement that we make to these algorithms following Cotter et al. [2019], is to employ the 0-1 loss while performing updates on λ, i.e., set $\ell = \ell^{\text{0-1}}$ in the λ-update step. We are able to do this because the convergence of the exponentiated gradient updates on λ does not depend on ℓ being differentiable. This modification allows λs to better reflect the model's per-class performance on the validation sample.

E.3.2 Discussion on post-shifting baseline

We implement the post-shifting method in Narasimhan and Menon [2021] (Algorithm 3 in their paper), which provides for an efficient way to construct a scoring function of the form $f^s_y(x) = \log(\gamma_y p^t_y(x))$, for a fixed teacher p^t, where the coefficients $\gamma \in \mathbb{R}^m$ are chosen to maximize the worst-class accuracy on the validation dataset. Interestingly, in our experiments, we find this approach to do exceedingly well on the validation sample, but this does not always translate to good worst-class test performance. In contrast, some of the teacher-student combinations that we experiment with were seen to over-fit less to the validation sample, and as a result were able to generalize better to the test set. This could perhaps indicate that the teacher labels we use in these combinations benefit the student in a way that it improves its generalization. The variance reduction effect that Menon et al. [2021a] postulate may be one possible explanation for why we see this behavior.

F ADDITIONAL EXPERIMENTAL RESULTS

This section contains additional experimental results.

F.1 EXTENDED TABLES FOR OBJECTIVE COMBINATIONS

We include extended tables comparing worst-class performance for different combinations of teacher and student objectives. The mean and standard errors are reported over repeat trainings as described in Appendix E.2.2.

Table 1 is an extended version of Table 1 that includes standard errors for both worst-k accuracy and average accuracy.

Table 2 includes similar comparisons when the student is compressed – that is, the student's architecture is smaller than the teacher's architecture.

Table 1: Worst-class accuracy comparison of self-distilled teacher/student combos on test. The "none" row indicates the performance of the teacher alone. Worst-class accuracy is shown above (or worst-10 accuracy for TinyImageNet-LT), and average is accuracy shown in parentheses below. The combination with the best worst-class accuracy is in **bold**. We include results for the robust student using either a teacher labeled validation set ("teacher val"), or true one-hot class labels in the validation set ("one-hot val"), as outlined in Appendix C. Perhaps counterintuitively, the teacher with the best worst-class accuracy alone (the "none" row) did not always produce the student with the highest worst-class accuracy.

		CIFAR-10 Teacher Obj.		**CIFAR-100** Teacher Obj.		**TinyImageNet** Teacher Obj.	
		L^{std}	L^{rob}	L^{std}	L^{rob}	L^{std}	L^{rob}
	none	86.48 ± 0.32 (93.74 ± 0.05)	90.09 ± 0.22 (92.67 ± 0.09)	42.22 ± 0.90 72.42 ± 0.16	43.42 ± 1.03 68.81 ± 0.11	8.42 ± 1.88 (56.79 ± 0.33)	11.87 ± 1.74 (48.40 ± 0.15)
Student Obj.	$L^{\text{std-d}}$	87.66 ± 0.40 (94.34 ± 0.07)	90.12 ± 0.23 (94.07 ± 0.07)	43.81 ± 0.58 (74.61 ± 0.15)	$\mathbf{48.20} \pm 1.15$ (73.23 ± 0.07)	6.32 ± 2.31 (57.83 ± 0.13)	10.53 ± 1.49 (55.36 ± 0.16)
	$L^{\text{rob-d}}$ (teacher val)	$\mathbf{90.94} \pm 0.16$ (92.54 ± 0.05)	85.14 ± 0.47 (89.58 ± 0.11)	39.18 ± 1.58 (63.49 ± 0.29)	30.42 ± 1.30 (55.77 ± 0.39)	9.98 ± 1.87 (49.84 ± 0.21)	16.58 ± 1.23 (46.11 ± 0.37)
	$L^{\text{rob-d}}$ (one-hot val)	89.37 ± 0.17 (91.63 ± 0.06)	87.32 ± 0.21 (91.16 ± 0.10)	44.61 ± 1.55 (69.02 ± 0.30)	42.68 ± 0.74 (62.03 ± 0.24)	16.27 ± 0.43 (48.06 ± 0.24)	$\mathbf{17.36} \pm 1.32$ (43.92 ± 0.30)

		CIFAR-10-LT Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
		L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
	None	57.26 ± 0.55 (76.27 ± 0.20)	68.52 ± 0.52 (79.85 ± 0.20)	74.8 ± 0.30 (80.29 ± 0.12)	0.00 ± 0.00 (43.33 ± 0.16)	3.75 ± 0.62 (47.55 ± 0.17)	10.33 ± 0.82 (44.27 ± 0.13)
Student Obj.	$L^{\text{std-d}}$	36.67 ± 0.28 (69.5 ± 0.13)	66.96 ± 0.43 (79.25 ± 0.10)	71.15 ± 0.24 (80.95 ± 0.11)	0.00 ± 0.00 (43.86 ± 0.14)	2.39 ± 0.24 (48.95 ± 0.15)	7.32 ± 0.47 (47.93 ± 0.11)
	$L^{\text{bal-d}}$	71.23 ± 0.44 (80.5 ± 0.12)	70.52 ± 0.20 (81.12 ± 0.08)	72.96 ± 0.53 (80.71 ± 0.07)	4.39 ± 0.65 (50.4 ± 0.11)	7.08 ± 0.80 (50.1 ± 0.09)	7.19 ± 0.79 (47.51 ± 0.20)
	$L^{\text{rob-d}}$ (teacher val)	63.85 ± 0.21 (76.81 ± 0.08)	$\mathbf{75.56} \pm 0.19$ (80.81 ± 0.08)	69.21 ± 0.45 (76.72 ± 0.19)	9.05 ± 0.71 (33.75 ± 0.10)	12.52 ± 0.98 (34.05 ± 0.09)	10.32 ± 0.76 (36.83 ± 0.15)
	$L^{\text{rob-d}}$ (one-hot val)	73.59 ± 0.25 (77.92 ± 0.05)	75.43 ± 0.38 (79.02 ± 0.07)	74.7 ± 0.19 $(77.99 \pm 0.10$	12.28 ± 0.46 (30.79 ± 0.18)	11.94 ± 0.80 (29.8 ± 0.20)	$\mathbf{13.18} \pm 0.61$ (31.88 ± 0.20)

		TinyImageNet-LT Teacher Obj.		
		L^{std}	L^{bal}	L^{rob}
	None	0.00 ± 0.00 (33.15 ± 0.17)	2.11 ± 0.37 (35.96 ± 0.12)	4.92 ± 0.66 (27.23 ± 0.15)
Student Obj.	$L^{\text{std-d}}$	0.00 ± 0.00 (26.05 ± 0.18)	0.00 ± 0.00 (27.21 ± 0.15)	1.87 ± 0.23 (25.34 ± 0.13)
	$L^{\text{bal-d}}$	0.20 ± 0.18 (30.43 ± 0.06)	2.82 ± 0.14 (39.41 ± 0.15)	4.77 ± 0.41 (38.41 ± 0.15)
	$L^{\text{rob-d}}$ (teacher val)	0.00 ± 0.00 (22.66 ± 0.08)	4.93 ± 0.38 (35.43 ± 0.18)	3.32 ± 0.43 (25.11 ± 0.17)
	$L^{\text{rob-d}}$ (one-hot val)	1.55 ± 0.37 (21.59 ± 0.19)	6.11 ± 0.39 (28.24 ± 0.17)	$\mathbf{6.19} \pm 0.25$ (25.30 ± 0.18)

F.1.1 Robust distillation with a onehot-labeled validation set

Tables 1 and 2 also include results when the robust student is trained using a validation set using onehot labels, as described in Appendix C. We report the accuracies for this robust student for different teachers trained with the standard, balanced, and robust objectives in the last rows of Tables 1 and 2 ($L^{\text{rob-d}}$ (one-hot val)). We compare these to the robust student trained using teacher labels on the validation set ($L^{\text{rob-d}}$ (teacher val)), which require less labeled data.

Table 2: Comparison of ResNet-56→ResNet-32 distilled teacher/student combos on test on CIFAR datasets. Worst-class accuracy shown above, and average accuracy shown in parentheses below. The combination with the best worst-class accuracy is bolded. Mean and standard error are reported over 10 repeats. We include results for the robust student using either a teacher labeled validation set ("teacher val"), or true one-hot class labels in the validation set ("one-hot val"), as outlined in Section 3.1.

		CIFAR-10 Teacher Obj.		**CIFAR-100** Teacher Obj.	
		L^{std}	L^{rob}	L^{std}	L^{rob}
Student Obj. (ResNet-32)	$L^{\text{std-d}}$	86.4 ± 0.27 (93.73 ± 0.05)	89.56 ± 0.20 (93.38 ± 0.05)	41.82 ± 1.12 73.19 ± 0.10	$\mathbf{45.7} \pm 1.13$ 71.42 ± 0.22
	$L^{\text{rob-d}}$ (teacher val)	$\mathbf{89.61} \pm 0.27$ (92.20 ± 0.08)	83.8 ± 0.95 (88.71 ± 0.24)	38.94 ± 2.61 (62.28 ± 0.40)	19.15 ± 0.00 (52.9 ± 0.00)
	$L^{\text{rob-d}}$ (one-hot val)	87.92 ± 0.23 (90.89 ± 0.12)	86.57 ± 0.24 (90.54 ± 0.11)	33.19 ± 1.29 (57.43 ± 0.29)	41.23 ± 0.84 (61.14 ± 0.24)

		CIFAR-10-LT Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
		L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
Student Obj. (ResNet-32)	$L^{\text{std-d}}$	57.23 ± 0.53 (75.76 ± 0.12)	66.80 ± 0.25 (78.99 ± 0.06)	72.36 ± 0.39 (80.74 ± 0.09)	0.00 ± 0.00 (44.33 ± 0.11)	1.38 ± 0.39 (47.28 ± 0.13)	7.99 ± 0.48 (47.34 ± 0.08)
	$L^{\text{bal-d}}$	71.37 ± 0.50 (81.13 ± 0.12)	71.00 ± 0.45 (81.12 ± 0.15)	72.17 ± 0.40 (79.91 ± 0.08)	3.57 ± 0.58 (49.21 ± 0.10)	4.28 ± 0.45 (46.56 ± 0.13)	5.58 ± 0.53 (48.58 ± 0.09)
	$L^{\text{rob-d}}$ (teacher val)	64.1 ± 0.36 (76.34 ± 0.12)	73.51 ± 0.33 (80.10 ± 0.10)	69.90 ± 0.42 (76.37 ± 0.14)	10.24 ± 0.71 (33.55 ± 0.16)	$\mathbf{13.41} \pm 0.72$ (33.37 ± 0.17)	11.27 ± 0.61 (36.14 ± 0.19)
	$L^{\text{rob-d}}$ (one-hot val)	72.65 ± 0.27 (77.69 ± 0.11)	74.39 ± 0.34 (78.68 ± 0.16)	$\mathbf{74.45} \pm 0.26$ $(77.97 \pm 0.10$	10.93 ± 0.65 (29.48 ± 0.22)	12.2 ± 0.65 (30.27 ± 0.18)	12.93 ± 0.62 $(31.83 \pm 0.17$

Perhaps surprisingly, it did not always benefit the robust student to utilize the true one-hot labels in the validation set. Instead, training the robust student with teacher labels on the validation set was often sufficient to achieve the best or close to the best worst-class performance. This is promising from a data efficiency standpoint, since it can be expensive to build up a labeled dataset for validation, especially if the training data is long-tailed.

F.2 ADDITIONAL PLOTS FOR ALL TRADE-OFF PARAMETER COMBINATIONS

Figure 1 show accuracies for all α^t, α^s the equivalent of Figure 1 but for all datasets.

F.3 COMPARISON TO BASELINES OF ALL PARETO EFFICIENT TRADE-OFF PARAMETERS

To supplement the comparison to baselines in Table 2, Figures 2 and 3 show all Pareto efficient α^t and α^s combinations on test. Whereas only a single α^t, α^s combination was selected on the validation set and reported in Table 2, Figures 2 and 3 show that there were many more combinations of α^t, α^s that could have Pareto dominated all baselines.

Figures 2 and 3 also give more insight into which values of α^t work best for different values of α^s. Whereas Figure 1 shows that α^s is highly correlated with average accuracy, the same is not true for α^t. Worst-class accuracy generally increases with α^s, but the teachers that achieve the Pareto efficient points all have $\alpha^t < 1$. This reveals counter-intuitively that the teacher's worst-class accuracy is not a direct predictor of the robustness of a subsequent student. This couples with our theoretical understanding in Section 5, which showed that the ability of a teacher to train robust students is determined by the calibration of scores within each class.

Trading off average vs. worst-class accuracy. Figures 2 and 3 show that when we allow for more nuanced L^{tdf} objective combinations, the resulting models may have higher average accuracy and worst-class accuracy than standard distillation. Interestingly, the models with the most "even" trade-offs between average accuracy and worst-class accuracy tend to have low α^t (around 0.25) and low α^s (also around 0.25). Higher values of α^t tended to lead to more extreme points on the trade-off curve, either with higher average accuracy at the expense of worst-class accuracy, or vice versa. Overall, the

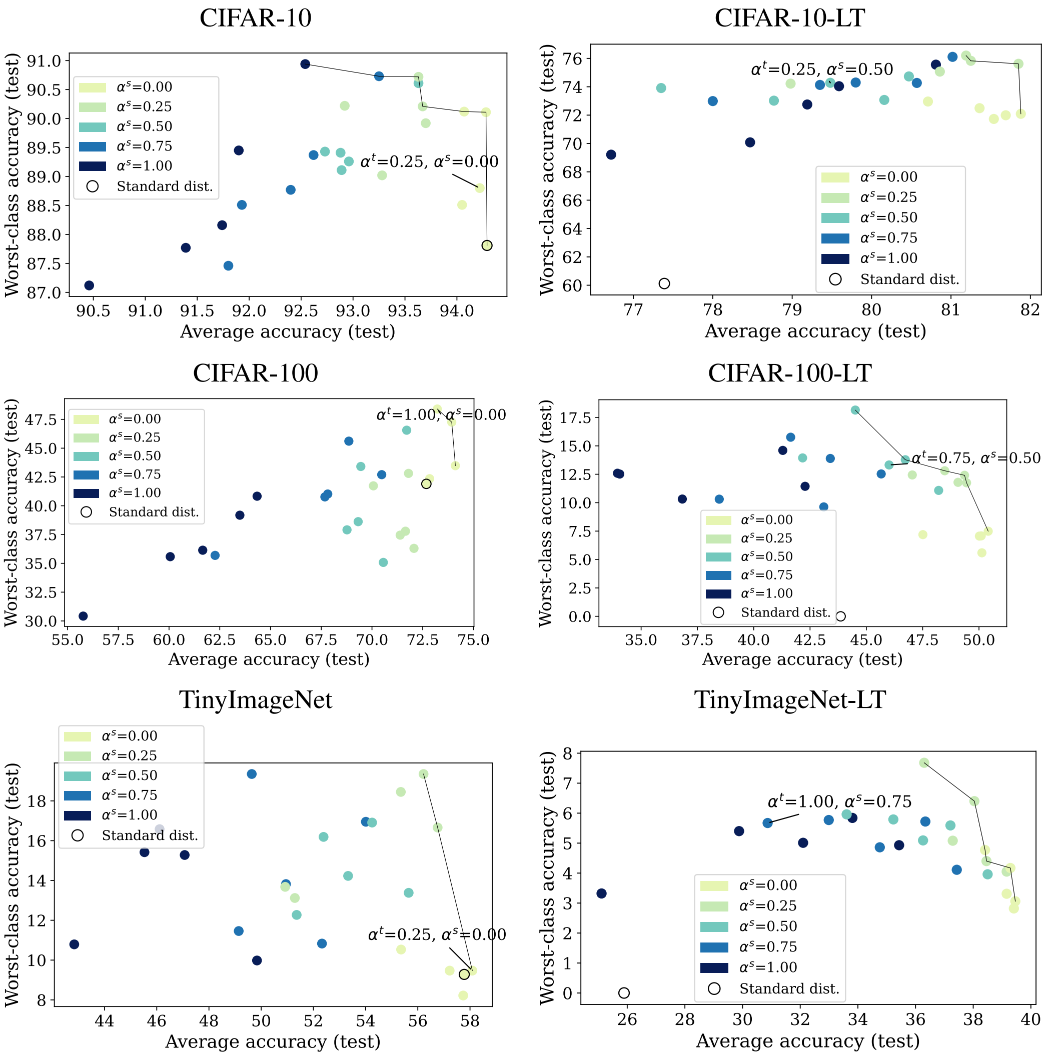
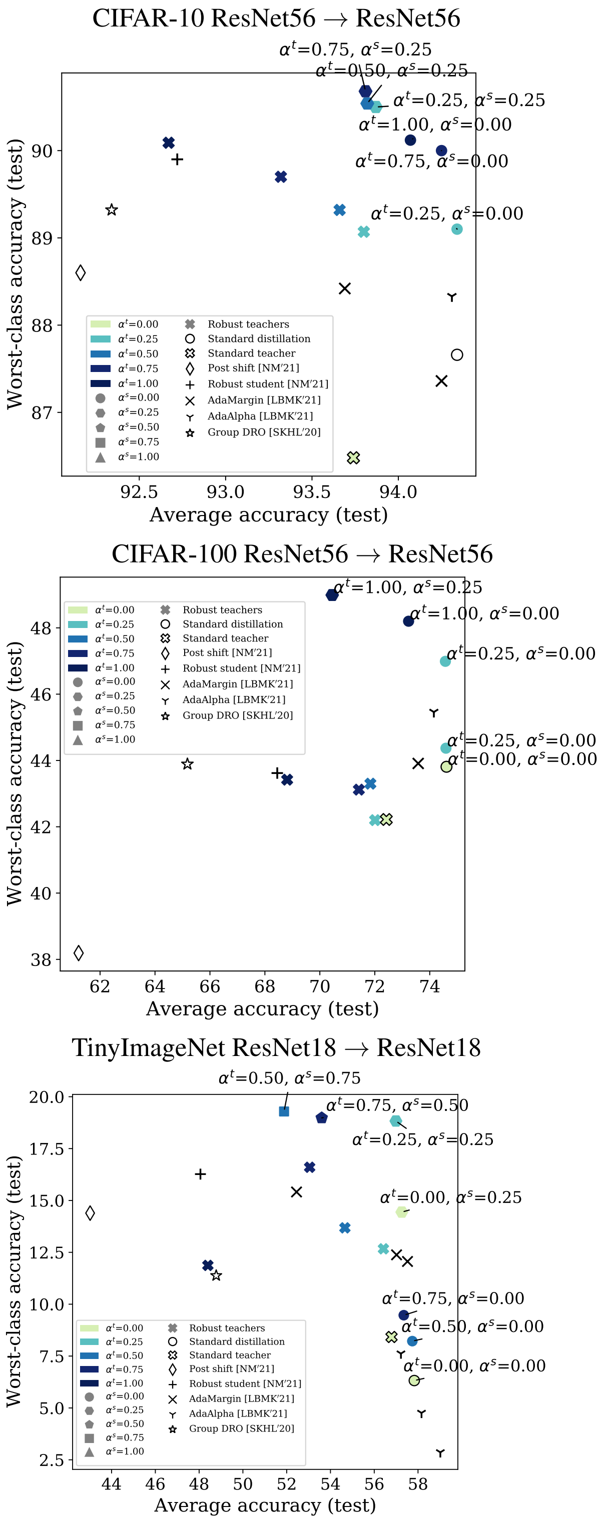


Figure 1: All α^t, α^s combinations for all datasets on test. The black line traces out the Pareto frontier. Average accuracy is roughly determined by α^s. The labeled point corresponds to the "best" combination selected in Table 2 based on validation criteria, but other domain-specific trade-off criteria could yield any of these other points.

robust L^{tdf} combinations also Pareto dominated most of the baselines that all used the standard teacher. Together, these results highlight the fact that in robust distillation, the teacher's training objective is important and should be tailored to the desired final accuracy/robustness trade-off (perhaps using a held-out validation sample with some domain-specific criteria in practice). Figure 4 confirms that these results also hold up in a compression setting, where the compressed models can actually even beat their larger teachers.

F.4 DIFFERENT TEACHERS ON REPEAT TRAININGS

Distillation experimental results in the main paper use the same teacher for all repeat trainings of the student. This captures the variance in the student training process while omitting the variance in the teacher training process. To capture the variance in the full training pipeline, we ran an additional set of experiments where students were trained on different retrained teachers, rather than on the same teacher. We report results on all CIFAR datasets in Table 3. The best teacher/student combinations are identical for all datasets except for CIFAR-10-LT, for which the best teacher/student combinations from Table 3 and Table 1 were both a robust student trained with a balanced teacher, and only differed in whether the validation set contained teacher labels or one-hot labels ($L^{\text{bal}}/L^{\text{rob-d}}$ (one-hot val) in Table 3 vs. $L^{\text{bal}}/L^{\text{rob-d}}$ (teacher val) in Table 1). Note that the first and second rows of Table 1 are already averaged over m retrained teachers ($m = 5$ for TinyImageNet, or



Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.75	0.25	93.81 ± 0.07	90.68 ± 0.20
0.50	0.25	93.82 ± 0.09	90.54 ± 0.22
0.25	0.25	93.87 ± 0.08	90.50 ± 0.18
1.00	0.00	94.07 ± 0.07	90.12 ± 0.23
0.75	0.00	94.25 ± 0.05	90.00 ± 0.17
0.25	0.00	94.34 ± 0.06	89.10 ± 0.31

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	94.34 ± 0.07	87.66 ± 0.40
Post shift [NM'21]	92.16 ± 0.18	88.60 ± 0.35
Robust student [NM'21]	92.72 ± 0.05	89.90 ± 0.21
AdaMargin [LBMK'21]	93.69 ± 0.06	88.42 ± 0.36
AdaAlpha [LBMK'21]	94.31 ± 0.01	88.33 ± 0.14
Group DRO [SKHL'20]	92.34 ± 0.07	89.32 ± 0.21

Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
1.00	0.25	70.45 ± 0.16	48.99 ± 0.72
1.00	0.00	73.23 ± 0.07	48.20 ± 1.15
0.25	0.00	74.57 ± 0.12	46.99 ± 1.09
0.25	0.00	74.59 ± 0.09	44.37 ± 0.58
0.00	0.00	74.61 ± 0.15	43.81 ± 0.58

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	74.61 ± 0.15	43.81 ± 0.58
Post shift [NM'21]	61.22 ± 0.36	38.19 ± 0.40
Robust student [NM'21]	68.45 ± 0.13	43.62 ± 1.27
AdaMargin [LBMK'21]	73.58 ± 0.11	43.91 ± 1.11
AdaAlpha [LBMK'21]	74.15 ± 0.08	45.46 ± 0.67
Group DRO [SKHL'20]	65.18 ± 0.08	43.89 ± 1.12

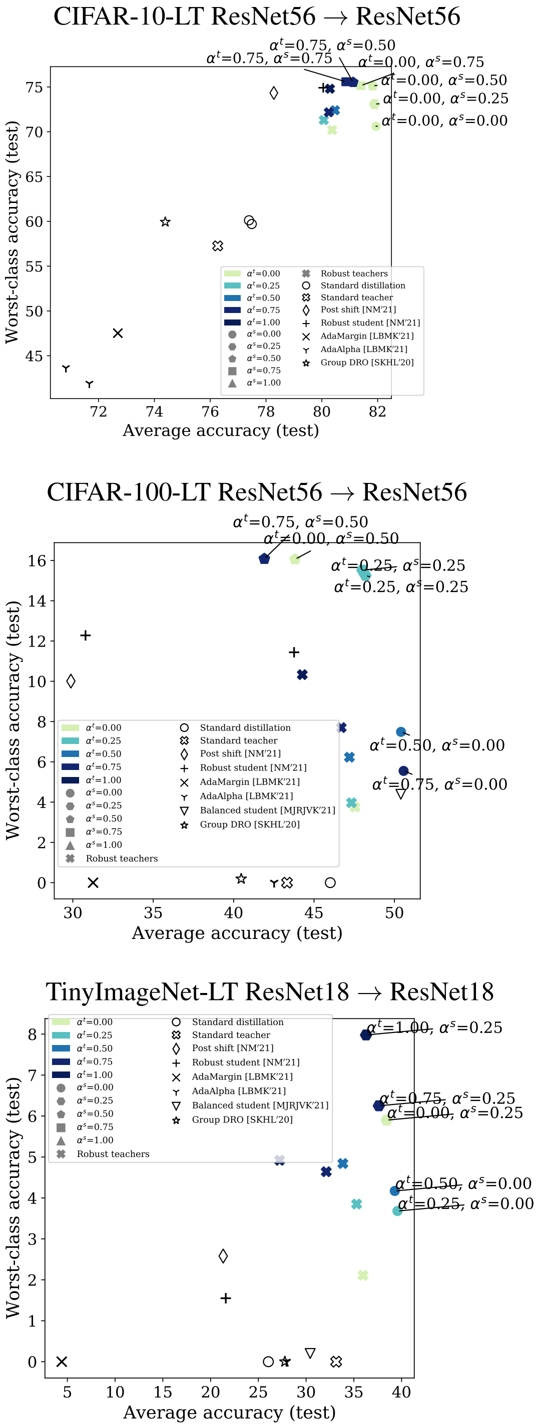
Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.50	0.75	51.88 ± 0.18	19.29 ± 1.27
0.75	0.50	53.60 ± 0.31	18.98 ± 0.86
0.25	0.25	56.99 ± 0.14	18.83 ± 0.85
0.00	0.25	57.26 ± 0.15	14.44 ± 0.91
0.75	0.00	57.35 ± 0.17	9.47 ± 1.76
0.50	0.00	57.74 ± 0.20	8.22 ± 1.09

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	57.83 ± 0.13	6.32 ± 2.31
Post shift [NM'21]	43.02 ± 0.79	14.39 ± 1.13
Robust student [NM'21]	48.06 ± 0.24	16.27 ± 0.43
AdaMargin [LBMK'21]	52.45 ± 0.08	15.41 ± 0.71
AdaAlpha [LBMK'21]	57.22 ± 0.08	7.62 ± 2.17
Group DRO [SKHL'20]	48.78 ± 0.21	11.38 ± 1.79

Figure 2: Trade-offs in worst-class test accuracy vs. average test accuracy for CIFAR-10 and CIFAR-100 distilling from ResNet-56 to ResNet-56, and TinyImageNet distilling from ResNet-18 to ResNet-18. All baseline results that require a teacher use the "standard teacher" (trained using L^{std}), as done in the original papers. For methods run multiple times with multiple hyperparameters (e.g. temperatures), all Pareto efficient results are shown in the plot, but the tables show only the baseline results with the best worst-class accuracy (on the validation set). The highlighted row indicates the model with the highest worst-class accuracy that also achieves at least as high average accuracy as *standard distillation*.



Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.75	0.75	80.86 ± 0.09	75.58 ± 0.17
0.75	0.50	81.12 ± 0.11	75.52 ± 0.22
0.00	0.75	81.40 ± 0.10	75.15 ± 0.38
0.00	0.50	81.82 ± 0.11	75.13 ± 0.24
0.00	0.25	81.89 ± 0.08	73.09 ± 0.32
0.00	0.00	81.94 ± 0.16	70.61 ± 0.39

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	77.39 ± 0.10	60.12 ± 0.56
Post shift [NM'21]	78.28 ± 0.05	74.33 ± 0.09
Robust student [NM'21]	80.05 ± 0.13	74.91 ± 0.24
AdaMargin [LBMK'21]	72.69 ± 0.24	47.52 ± 0.95
AdaAlpha [LBMK'21]	70.83 ± 0.28	43.64 ± 1.09
Group DRO [SKHL'20]	74.39 ± 0.17	59.93 ± 0.59

Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.75	0.50	41.91 ± 0.15	16.08 ± 0.52
0.00	0.50	43.82 ± 0.14	16.06 ± 0.89
0.25	0.25	48.01 ± 0.09	15.52 ± 0.41
0.25	0.25	48.20 ± 0.11	15.26 ± 0.73
0.50	0.00	50.41 ± 0.11	7.49 ± 0.72
0.75	0.00	50.57 ± 0.18	5.55 ± 0.54

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	46.01 ± 0.16	0.00 ± 0.00
Post shift [NM'21]	29.88 ± 0.61	10.01 ± 0.72
Robust student [NM'21]	30.79 ± 0.18	12.28 ± 0.46
AdaMargin [LBMK'21]	31.26 ± 0.21	0.00 ± 0.00
AdaAlpha [LBMK'21]	42.52 ± 0.08	0.00 ± 0.00
Balanced student [MJRJVK'21]	50.40 ± 0.12	4.39 ± 0.66
Group DRO [SKHL'20]	40.47 ± 0.17	0.19 ± 0.17

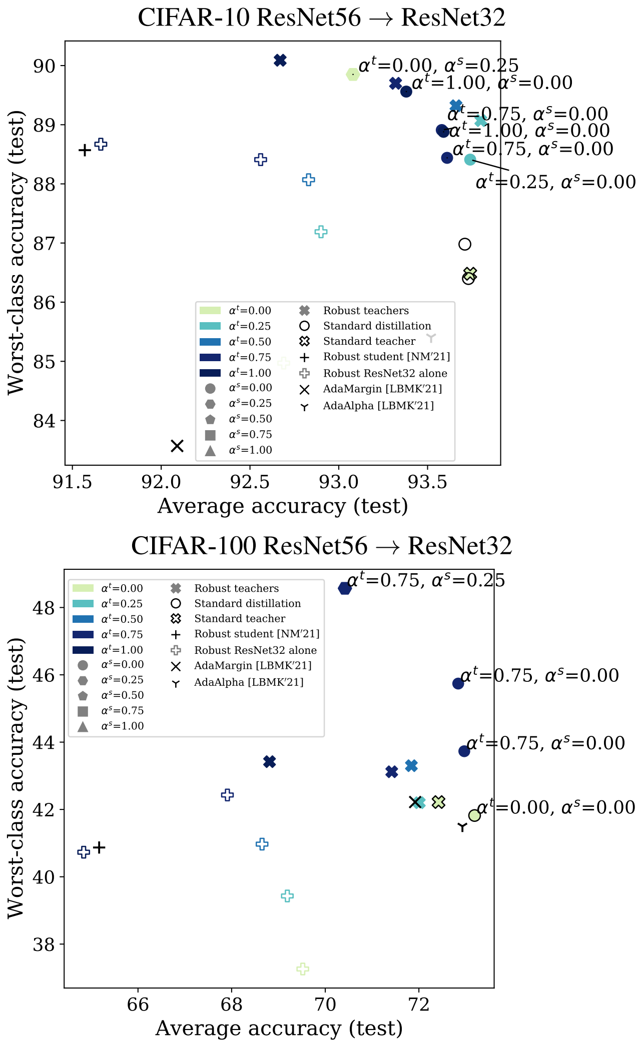
Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-10 acc.
1.00	0.25	36.28 ± 0.17	7.98 ± 0.21
0.75	0.25	37.62 ± 0.15	6.25 ± 0.12
0.00	0.25	38.44 ± 0.13	5.90 ± 0.45
0.50	0.00	39.29 ± 0.09	4.17 ± 0.34
0.25	0.00	39.57 ± 0.06	3.68 ± 0.30

Baseline results (test)

Baseline	Average acc.	Worst-10 acc.
Standard distillation	26.05 ± 0.18	0.00 ± 0.00
Post shift [NM'21]	21.32 ± 0.49	2.58 ± 0.42
Robust student [NM'21]	21.59 ± 0.19	1.55 ± 0.37
AdaMargin [LBMK'21]	4.41 ± 0.09	0.00 ± 0.00
AdaAlpha [LBMK'21]	27.95 ± 0.14	0.00 ± 0.00
Balanced student [MJRJVK'21]	30.43 ± 0.06	0.20 ± 0.18
Group DRO [SKHL'20]	27.78 ± 0.13	0.00 ± 0.00

Figure 3: Trade-offs in worst-class test accuracy vs. average test accuracy for CIFAR-10-LT, CIFAR-100-LT, and TinyImageNet-LT under self-distillation. All baseline results that require a teacher use the "standard teacher" (trained using L^{std}), as done in the original papers. For methods run multiple times with multiple hyperparameters (e.g. temperatures), all Pareto efficient results are shown in the plot, but the tables show only the baseline results with the best worst-class accuracy (on the validation set). The highlighted row indicates the model with the highest worst-class (or worst-10) accuracy that also achieves at least as high average accuracy as *standard distillation* (within error margins). Note that the for the LT datasets, L^{tdf} mixes between L^{bal} and L^{rob}.



Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.00	0.25	93.08 ± 0.07	89.85 ± 0.22
1.00	0.00	93.38 ± 0.05	89.56 ± 0.20
0.75	0.00	93.58 ± 0.09	88.91 ± 0.25
1.00	0.00	93.59 ± 0.06	88.88 ± 0.36
0.75	0.00	93.61 ± 0.05	88.44 ± 0.33
0.25	0.00	93.74 ± 0.07	88.41 ± 0.32

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	93.71 ± 0.05	86.98 ± 0.36
Robust student [NM'21]	91.57 ± 0.08	88.57 ± 0.18
AdaMargin [LBMK'21]	92.09 ± 0.09	83.57 ± 0.64
AdaAlpha [LBMK'21]	93.52 ± 0.11	85.41 ± 0.45

Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.75	0.25	70.42 ± 0.14	48.57 ± 0.55
0.75	0.00	72.84 ± 0.22	45.74 ± 1.57
0.75	0.00	72.97 ± 0.18	43.73 ± 1.72

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	73.19 ± 0.10	41.82 ± 1.12
Robust student [NM'21]	65.17 ± 0.11	40.87 ± 0.89
AdaMargin [LBMK'21]	71.92 ± 0.17	42.22 ± 1.65
AdaAlpha [LBMK'21]	72.93 ± 0.09	41.50 ± 1.14

Figure 4: Trade-offs in worst-class test accuracy vs. average test accuracy for CIFAR-10 and CIFAR-100 distilling from ResNet-56 to ResNet-32. All baseline results that require a teacher use the "standard teacher" (trained using L^{std}), as done in the original papers. For methods run multiple times with multiple hyperparameters (e.g. temperatures), all Pareto efficient results are shown in the plot, but the tables show only the baseline results with the best worst-class accuracy (on the validation set). The highlighted row indicates the model with the highest worst-class accuracy that also achieves at least as high average accuracy as *standard distillation* (within error margins).

$m = 10$ for CIFAR datasets), and those same m teachers are used in the repeat trainings in Table 3.

Table 3: Comparison using different teachers for student retrainings for self-distilled teacher/student combos on test. For each student/teacher objective pair, we train $m = 10$ students total on each of $m = 10$ distinct retrained teachers. For comparability, the same set of m teachers is used for each student. This differs from Table 1 in that in Table 1, the students are retrained on each repeat using the same teacher (arbitrarily selected). Otherwise, setups are the same as in Table 1.

		CIFAR-10 Teacher Obj.		**CIFAR-100**Teacher Obj.	
		L^{std}	L^{rob}	L^{std}	L^{rob}
Student Obj.	$L^{\text{std-d}}$	87.09 ± 0.51 (93.78 ± 0.22)	89.68 ± 0.20 (93.74 ± 0.07)	44.21 ± 0.57 74.6 ± 0.11	$\mathbf{47.79} \pm 0.82$ 73.48 ± 0.11
	$L^{\text{rob-d}}$ (teacher val)	$\mathbf{90.62} \pm 0.19$ (92.58 ± 0.08)	87.12 ± 0.38 (90.46 ± 0.08)	39.7 ± 1.32 (64.28 ± 0.41)	31.09 ± 1.21 (55.39 ± 0.28)
	$L^{\text{rob-d}}$ (one-hot val)	88.15 ± 0.66 (91.03 ± 0.47)	86.44 ± 0.52 (90.16 ± 0.42)	39.44 ± 0.94 (61.23 ± 0.36)	39.65 ± 0.59 (60.89 ± 0.29)

		CIFAR-10-LT Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
		L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
Student Obj.	$L^{\text{std-d}}$	60.12 ± 0.56 (77.39 ± 0.10)	66.13 ± 0.47 (79.16 ± 0.20)	69.75 ± 0.52 (80.73 ± 0.08)	0.00 ± 0.00 (45.84 ± 0.13)	1.41 ± 0.41 (49.67 ± 0.20)	9.17 ± 0.74 (48.55 ± 0.14)
	$L^{\text{bal-d}}$	72.41 ± 0.52 (81.97 ± 0.11)	71.49 ± 0.30 (81.20 ± 0.15)	71.70 ± 0.33 (80.29 ± 0.11)	5.83 ± 0.54 (50.58 ± 0.15)	5.94 ± 0.50 (50.85 ± 0.14)	8.37 ± 0.72 (48.16 ± 0.20)
	$L^{\text{rob-d}}$ (teacher val)	62.77 ± 0.58 (77.18 ± 0.15)	73.09 ± 0.34 (80.03 ± 0.22)	68.04 ± 0.47 (75.36 ± 0.25)	10.53 ± 0.76 (33.69 ± 0.14)	12.04 ± 0.89 (34.08 ± 0.12)	9.66 ± 1.15 (37.10 ± 0.15)
	$L^{\text{rob-d}}$ (one-hot val)	$\mathbf{75.10} \pm 0.36$ (79.27 ± 0.13)	$\mathbf{75.10} \pm 0.50$ (79.07 ± 0.20)	74.16 ± 0.34 $(78.11 \pm 0.14$	10.74 ± 0.44 (30.36 ± 0.39)	11.95 ± 0.69 (31.00 ± 0.16)	$\mathbf{12.87} \pm 0.81$ $(31.62 \pm 0.34$

F.5 ADAALPHA AND ADAMARGIN COMPARISONS WITH DIFFERENT TEACHERS

We include and discuss additional comparisons to the AdaMargin and AdaAlpha methods Lukasik et al. [2022], which each define additional ways to modify the student training algorithm (see Section 4). In Table 2, we show results with each of these methods using the standard teacher, as done in the original paper. However, in this section we extend these results by also applying AdaMargin and AdaAlpha with different teachers trained with the robust and balanced objectives. Table 4 compares the results of AdaMargin and AdaAlpha for these different teachers under the same self distillation setup as Table 1.

Overall, the use of a robust teacher leads to marked improvements for students trained by AdaMargin and AdaAlpha. For the balanced datasets, AdaMargin was competitive with the robust and standard students: on CIFAR-100 and TinyImageNet, AdaMargin combined with the robust teacher and the standard teacher (respectively) achieved worst-class accuracies that are statistically comparable to the best worst-class accuracies in Table 1. However, on the long-tailed datasets, AdaAlpha and AdaMargin did not achieve worst-class accuracies as high as other teacher/student combinations. This suggests that the AdaMargin method can work well on balanced datasets in combination with a robust teacher, but other combinations of standard/balanced/robust objectives are valuable for long-tailed datasets.

Relative to each other, AdaMargin usually achieved higher worst-class accuracy than AdaAlpha, whereas AdaAlpha often achieved higher average accuracy.

F.6 GROUP DRO COMPARISON

Sagawa et al. [2020] propose a group DRO algorithm to improve long tail performance without distillation. In this section we present additional experimental comparisons to Algorithm 1 from Sagawa et al. [2020]. This differs from our robust optimization methodology in Section 3.1 in two key ways: *(i)* we apply a margin-based surrogates of Menon et al. [2021b], and *(ii)* we use a validation set to update the Lagrange multipliers λ in Algorithm 2. Table 5 shows results from running group DRO directly as specified in Algorithm 1 in Sagawa et al. [2020], as well as a variant where we use the validation set

Table 4: Results for AdaAlpha and AdaMargin baselines for different teachers under self-distillation. For all CIFAR datasets, self-distillation is done from ResNet56 → ResNet56. For TinyImageNet, self-distillation is done from ResNet18 → ResNet18. Worst-class accuracy shown above (or worst-10 accuracy for TinyImageNet-LT), and average accuracy is shown in parentheses below. The temperature hyperparameter was tuned to maximize worst-class accuracy on the held-out validation set. Mean and standard error are reported over 5 repeats for all datasets.

	CIFAR-10 Teacher Obj.		**CIFAR-100** Teacher Obj.		**TinyImageNet** Teacher Obj.	
	L^{std}	L^{rob}	L^{std}	L^{rob}	L^{std}	L^{rob}
Ada Alpha	88.33 ± 0.14 (94.31 ± 0.01)	89.96 ± 0.44 (93.97 ± 0.07)	43.50 ± 0.62 73.96 ± 0.09	45.59 ± 0.82 71.42 ± 0.14	11.11 ± 1.29 61.13 ± 0.09	16.58 ± 1.67 56.84 ± 0.15
Ada Margin	87.36 ±0.06 (94.25 ± 0.02)	90.37±0.26 (94.02 ± 0.12)	43.91 ±1.11 (73.58 ± 0.11)	47.78 ±0.96 (70.92 ± 0.09)	18.17 ±3.89 (61.3 ± 0.28)	17.84 ±1.77 (55.77 ± 0.32)

	CIFAR-10-LT Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
	L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
Ada Alpha	41.90 ± 0.44 (71.67 ± 0.08)	66.23 ± 0.39 (77.87 ± 0.16)	71.17 ± 0.32 (79.66 ± 0.13)	0.00 ± 0.00 (42.52 ± 0.08)	1.46 ± 0.61 (45.44 ± 0.14)	9.15 ± 0.54 (45.64 ± 0.11)
Ada Margin	47.52 ±0.95 (72.69 ± 0.24)	66.74 ±0.35 (78.20 ± 0.09)	70.33 ±0.50 (78.87 ± 0.12)	0.00 ±0.00 (31.26 ± 0.21)	0.00 ±0.00 (34.06 ± 0.12)	12.46 ±0.36 (42.90 ± 0.07)

	TinyImageNet-LT Teacher Obj.		
	L^{std}	L^{bal}	L^{rob}
Ada Alpha	0.00 ± 0.00 (28.14 ± 0.12)	0.00 ± 0.00 (0.50 ± 0.00)	0.00 ± 0.00 (0.50 ± 0.00)
Ada Margin	0.00 ± 0.00 (9.18 ± 0.09)	0.00 ± 0.00 (7.92 ± 0.10)	0.41 ± 0.17 (23.08 ± 0.15)

to update Lagrange multipliers in group DRO (labeled as "with vali" in Table 5). Table 5 shows that this latter variant "with vali" performs better than the original version without a validation set; thus, for the results in Figures 2 and 3, we report these better results marked in Table 5 as "with vali." Overall, this comparison shows that L^{rob} is comparable to group DRO, and that robust distillation protocols can outperform group DRO alone.

Table 5: Results from comparison to group DRO (Algorithm 1 in Sagawa et al. [2020]) without distillation. "No vali" uses the training set to update group Lagrange multipliers, as done originally by Sagawa et al. [2020]. "With vali" uses the validation set to compute group Lagrange multipliers as done in all other experiments in our paper. Worst-class accuracy is shown above, and balanced accuracy is shown in parentheses below. Mean and standard error are shown over 5 repeats.

CIFAR-10 group DRO		**CIFAR-100** group DRO		**TinyImageNet** group DRO	
No vali	With vali	No vali	With vali	No vali	With vali
86.65 ±0.49 (93.61 ± 0.09)	89.32 ±0.21 (92.34 ± 0.07)	40.35 ±1.18 70.25 ± 0.17	43.89 ±1.12 65.18 ± 0.08	0.00 ±0.00 (6.55 ± 0.41)	9.17 ±1.55 (47.67 ± 0.22)

CIFAR-10-LT group DRO		**CIFAR-100-LT** group DRO		**TinyImageNet-LT** group DRO	
No vali	With vali	No vali	With vali	No vali	With vali
51.59 ±2.49 (71.94 ± 0.75)	59.93 ±0.59 (74.39 ± 0.17)	0.00 ±0.00 (39.81 ± 0.23)	0.19 ±0.17 (40.47 ± 0.17)	0.00 ±0.00 (9.79 ± 0.40)	0.00 ±0.00 (22.49 ± 0.10)

F.7 ADDITIONAL IMAGENET COMPARISONS

Here we present additional results when training ResNet-18 teachers and students on ImageNet. Table 6 includes measures of worst-1 accuracy, worst-10 accuracy, worst-50 accuracy, and worst-100 accuracy.

Table 6: ImageNet comparison of ResNet-18 teacher/student combos on test. Average worst-1/10/100 accuracy shown above, standard accuracy shown in parentheses below. The combination with the best worst-class accuracy is bolded. Mean and standard error are reported over up to 5 repeats.

		Worst-1 Accuracy		Worst-10 Accuracy		Worst-50 Accuracy		Worst-100 Accuracy	
		ImageNet Teacher Obj.		**ImageNet** Teacher Obj.		**ImageNet** Teacher Obj.		**ImageNet** Teacher Obj.	
		L^{std}	L^{rob}	L^{std}	L^{rob}	L^{std}	L^{rob}	L^{std}	L^{rob}
	none	0.00 (67.29)	10.71 (63.10)	11.54 (67.29)	17.13 (63.10)	24.01 67.29	25.08 63.10	30.35 (67.29)	29.96 (63.10)
Student Obj.	Post shift	8.70 (48.62)	3.57 (48.83)	16.15 (48.62)	11.64 (48.83)	21.85 (48.62)	18.86 (48.83)	25.58 (48.62)	23.17 (48.83)
	$L^{\text{std-d}}$	3.20 ± 1.33 (65.46 ± 0.05)	3.79 ± 0.11 (64.54 ± 0.01)	10.07 ± 0.27 (65.46 ± 0.05)	10.22 ± 0.33 (64.54 ± 0.01)	20.30 ± 0.39 (65.46 ± 0.05)	22.61 ± 0.32 (64.54 ± 0.01)	26.45 ± 0.25 (65.46 ± 0.05)	29.01 ± 0.21 (64.54 ± 0.01)
	$L^{\text{rob-d}}$ (teacher val)	0.00 ± 0.00 (59.60 ± 0.10)	0.00 ± 0.00 (51.01 ± 0.12)	1.18 ± 0.02 (59.60 ± 0.10)	1.47 ± 0.04 (51.01 ± 0.12)	13.02 ± 0.16 (59.60 ± 0.10)	6.85 ± 0.13 (51.01 ± 0.12)	21.00 ± 0.19 (59.60 ± 0.10)	11.26 ± 0.16 (51.01 ± 0.12)
	$L^{\text{rob-d}}$ (one-hot val)	0.00 ± 0.00 (59.65 ± 0.01)	0.00 ± 0.00 (55.34 ± 0.00)	8.32 ± 1.04 (59.65 ± 0.01)	5.99 ± 0.00 (55.34 ± 0.00)	18.77 ± 0.03 (59.65 ± 0.01)	16.82 ± 0.00 (55.34 ± 0.00)	23.95 ± 0.31 (59.65 ± 0.01)	22.16 ± 0.00 (55.34 ± 0.00)

References

Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Seungil You, Karthik Sridharan, and Maya R. Gupta. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research (JMLR)*, 20(172):1–59, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: understanding and mitigating biases in distillation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=ph3AYXpwEb`.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning*, pages 7632–7642. PMLR, 2021a.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021b.

Harikrishna Narasimhan and Aditya K Menon. Training over-parameterized models with non-decomposable objectives. *Advances in Neural Information Processing Systems*, 34, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4 (2):107–194, 2011.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:1–52, 2016.