A Strong Duality of the CSL Problem

Assumption 1. The losses $\ell(\cdot, y)$ and $\ell'(\cdot, y)$, are convex functions for all $y \in \mathcal{Y}$.

In Assumption 1, the convexity of the losses is taken with respect to the model's output, and not the model parameters. The cross-entropy loss, commonly used in classification with a softmax layer satisfies strong convexity when considering a probability simplex [64]. Typical losses for regression (e.g., mean-squared error, L1 loss) also satisfy this assumption.

Assumption 2. The hypothesis class $\mathcal{F}$ is convex.

To obtain a convex hypothesis class, as required by Assumption 2, it suffices to take the convex hull of the function class originally considered.

Assumption 3. There exists $f \in \mathcal{F}$ strictly feasible for (CSL) (i.e., $\ell'\left(f(\boldsymbol{x}), y\right) < \epsilon(\boldsymbol{x}), \quad \mathfrak{D}$–a.e.)

Assumption 3 guarantees that the problem (CSL) is feasible and that its dual is well-posed.

Proposition A.1. *Under Assumptions 1-3,* (CSL) *and* (D-CSL) *are strongly dual, i.e.,* $P^* = D^*$.

Proof. Note that assumptions 1 and 2 imply that (CSL) is a *convex* program. Under the strict feasibility assumption, (CSL) satisfies the constraint qualification known as *Slater's condition*, from which strongly duality follows [65, 66]. □

B Proof of Theorem 3.2 (Sensitivity of $P^\star$)

This result stems from a sensitivity analysis on the constraint of problem (CSL) and is well-known in the convex optimization literature. More general versions of this theorem are shown in [67] (Section 4), [68] or [69].

We start by viewing (CSL) as an optimization problem parameterized by the function $\epsilon(\boldsymbol{x})$:

$$\begin{aligned} P^\star(\epsilon(\boldsymbol{x})) = \min_{f \in \mathcal{F}} \quad & \mathbb{E}_{\mathfrak{D}}\left[\ell\left(f(\boldsymbol{x}), y\right)\right] \\ \text{s.t.} \quad & \ell'\left(f(\boldsymbol{x}), y\right) \leq \epsilon(\boldsymbol{x}), \quad \mathfrak{D}_{\mathbf{x}}\text{–a.e.} \end{aligned}$$

Define the Lagrangian $L(f, \lambda(\boldsymbol{x}); \epsilon(\boldsymbol{x}))$ as

$$L(f, \lambda(\boldsymbol{x}); \epsilon(\boldsymbol{x})) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\left[\ell(f(\boldsymbol{x}), y) + \lambda(\boldsymbol{x})(\ell'(f(\boldsymbol{x}), y) - \epsilon(\boldsymbol{x}))\right],$$

where the dependence on $\epsilon(\boldsymbol{x})$ is explicitly shown. Then, following the definition of $P^\star(\epsilon(\boldsymbol{x}))$ and using strong duality, we have

$$P^\star(\epsilon(\boldsymbol{x})) = \min_f L(f, \lambda^\star(\boldsymbol{x}; \epsilon(\boldsymbol{x})); \epsilon(\boldsymbol{x})) \leq L(f, \lambda^\star(\boldsymbol{x}; \epsilon(\boldsymbol{x})); \epsilon(\boldsymbol{x}))$$

with the inequality being true for any function f, and where the dependence of $\lambda^\star$ on $\epsilon(\boldsymbol{x})$ is also explicitly shown. Now, consider an arbitrary function $\epsilon'(\boldsymbol{x})$ and the respective primal function $f^\star(\cdot; \epsilon'(\boldsymbol{x}))$ which minimizes its corresponding Lagrangian. Plugging $f^\star(\cdot; \epsilon'(\boldsymbol{x}))$ into the above inequality, we have

$$\begin{aligned} P^\star(\epsilon(\boldsymbol{x})) &\leq L(f^\star(\cdot; \epsilon'(\boldsymbol{x})), \lambda^\star(\boldsymbol{x}; \epsilon(\boldsymbol{x})); \epsilon(\boldsymbol{x})) \\ &= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\left[\ell(f^\star(\boldsymbol{x}; \epsilon'(\boldsymbol{x})), y) + \lambda^\star(\boldsymbol{x}; \epsilon(\boldsymbol{x}))(\ell'(f^\star(\boldsymbol{x}; \epsilon'(\boldsymbol{x})), y) - \epsilon(\boldsymbol{x}))\right] \end{aligned}$$

Now, since $f^\star(\cdot; \epsilon'(\boldsymbol{x}))$ is *optimal* for constraint bounds given by $\epsilon'(\boldsymbol{x})$ and complementary slackness holds, we have $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\left[\ell(f^\star(\boldsymbol{x}; \epsilon'(\boldsymbol{x})), y)\right] = P^\star(\epsilon'(\boldsymbol{x}))$. Moreover, since $f^\star(\cdot; \epsilon'(\boldsymbol{x}))$ is feasible for constraint bounds given by $\epsilon'(\boldsymbol{x})$, we have $\ell'(f^\star(\boldsymbol{x}; \epsilon'(\boldsymbol{x})), y) \leq \epsilon'(\boldsymbol{x}), \ \mathfrak{D}_{\mathbf{x}}$-a.e. Combining the above, we get

$$\begin{aligned} P^\star(\epsilon(\boldsymbol{x})) &\leq P^\star(\epsilon'(\boldsymbol{x})) + \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\left[\lambda^\star(\boldsymbol{x}; \epsilon(\boldsymbol{x}))(\epsilon'(\boldsymbol{x}) - \epsilon(\boldsymbol{x}))\right] \\ &= P^\star(\epsilon'(\boldsymbol{x})) + \langle\lambda^\star(\boldsymbol{x}; \epsilon(\boldsymbol{x}))(\epsilon'(\boldsymbol{x}) - \epsilon(\boldsymbol{x}))\rangle, \end{aligned}$$

or eqivalently,

$$P^\star(\epsilon'(\boldsymbol{x})) - P^\star(\epsilon(\boldsymbol{x})) \geq \langle -\lambda^\star(\boldsymbol{x}; \epsilon(\boldsymbol{x}))(\epsilon'(\boldsymbol{x}) - \epsilon(\boldsymbol{x}))\rangle,$$

which exactly matches the definition of the Fréchet subdifferential in Definition 3.1, hence completing the proof.

This shows that $-\lambda^\star(\boldsymbol{x})$ is a sub-gradient of $P^*(\epsilon(\boldsymbol{x}))$. Let ∂P^* be the sub-differential of P^* (i.e., the set of all sub-gradients). By definition, P^* is differentiable if its sub-differential is a singleton: $\partial P^* = \{-\lambda^\star(\boldsymbol{x})\}$. This holds in our problem if the Lagrangian minimizers are unique, which is the case for a strongly convex objective. However, it does not hold in the general case. In [68], right-side differentiability is shown by further assuming that $\exists \alpha > P^*$ and a compact set S such that set of feasible points where the objective does not exceed α is contained in S.

C The Statistical Bias Induced by Active Sampling

As mentioned in Section 3.3, when performing several active learning iterations, we undertake a biased version of (BAL):

$$\mathcal{B}^* = \underset{\mathcal{B}\subseteq\mathcal{U}_t\,:\,|\mathcal{B}|\leq b}{\arg\min} \; \min_{f\in\mathcal{F}} \mathbb{E}_{(\mathbf{x},y)\sim\mathfrak{A}^{(t)}} \left[\ell\left(f(\mathbf{x}; \mathcal{L}_t \cup \mathcal{B}), y\right)\right], \tag{b-BAL}$$

where $\mathfrak{A}^{(t)}$ represents the biased/shifted distribution underlying the actively sampled set $\mathcal{L}^{(t)}$.. This makes the learnt predictor $f(\mathbf{x}; \mathcal{L}^{(t)} \cup \mathcal{B})$ sub-optimal for the natural data distribution $\mathfrak{D}$.

We focus our attention on the setting with $\ell = \ell'$. Observe that if $\mathfrak{D}_{\mathbf{x}}$ and $\mathfrak{A}_{\mathbf{x}}$ have the same support, then the feasibility formulation of the inner minimization is the same whether we consider the natural or biased data distribution. That is,

$$\begin{aligned} P^\star = \min_{f\in\mathcal{F}} \quad & 0 \\ \text{s.t.} \quad & \ell\left(f(\boldsymbol{x}), y\right) \leq \epsilon(\boldsymbol{x}), \quad \mathfrak{D}_{\boldsymbol{x}}\text{–a.e.} \end{aligned} \qquad \Longleftrightarrow \qquad \begin{aligned} P^\star = \min_{f\in\mathcal{F}} \quad & 0 \\ \text{s.t.} \quad & \ell\left(f(\boldsymbol{x}), y\right) \leq \epsilon(\boldsymbol{x}), \quad \mathfrak{A}_{\boldsymbol{x}}\text{–a.e.} \end{aligned}$$

This is because,

$$\begin{aligned} \ell(f(\mathbf{x}), y) \leq \epsilon(\mathbf{x}) \quad & \mathfrak{D}_{\boldsymbol{x}}\text{–a.e.} \qquad \Longleftrightarrow \qquad \ell(f(\mathbf{x}), y) p_{\mathfrak{D}_{\boldsymbol{x}}}(\mathbf{x}) \leq \epsilon(\mathbf{x}) p_{\mathfrak{D}_{\boldsymbol{x}}}(\mathbf{x}) \quad \forall \mathbf{x} \\ \ell(f(\mathbf{x}), y) \leq \epsilon(\mathbf{x}) \quad & \mathfrak{A}_{\boldsymbol{x}}\text{–a.e.} \qquad \Longleftrightarrow \qquad \ell(f(\mathbf{x}), y) p_{\mathfrak{D}'_{\boldsymbol{x}}}(\mathbf{x}) \leq \epsilon(\mathbf{x}) p_{\mathfrak{A}_{\boldsymbol{x}}}(\mathbf{x}) \quad \forall \mathbf{x} \end{aligned}$$

Notice that for both distributions, the constraints are *only* enforced when the respective density is non-zero. When the density is exactly zero, the constraint is satisfied trivially. Therefore, if

$$\{x \,:\, p_{\mathfrak{D}_{\boldsymbol{x}}}(\mathbf{x}) > 0\} = \{x \,:\, p_{\mathfrak{A}_{\boldsymbol{x}}}(\mathbf{x}) > 0\}$$

then the feasible set is the same for both problems. Since the objectives also coincide, the problems are identical.
In fact, we can replace the objective in the above problems by any functional $R(f)$ that does not depend on the data distribution (e.g: regularizers or complexity measures such as $\|f\|$) and the equivalence still holds. When setting a regularizing functional as the objective, the resulting dual problem is:

$$D^\star = \max_{\lambda\in\Lambda} \; \min_{f\in\mathcal{F}} R(f) + \lambda(\boldsymbol{x})(\ell(f(\boldsymbol{x}), y) - \epsilon(\boldsymbol{x})) \tag{D-FSL}$$

We know that the primal problem of (D-FSL) is agnostic to the underlying data distribution. If we further assume the existence of a strictly feasible solution, strong duality holds (see Appendix A). Thus, the solution of (D-FSL) is also unimpacted by the distribution shift induced by active sampling.

Observe that the dual problem in the original ALLY formulation, with $\ell = \ell'$, is:

$$D^\star = \max_{\lambda\in\Lambda} \; \min_{f\in\mathcal{F}} (\lambda(\boldsymbol{x}) + \mathbf{1})\ell(f(\boldsymbol{x}), y) - \lambda(\boldsymbol{x})\epsilon(\boldsymbol{x}) \tag{D-CSL}$$

Note that the dual problems (D-CSL) and (D-FSL) differ in two aspects: the presence of a regularizer $R(f)$ and a unit shift in the dual variable function $\lambda(\mathbf{x})$. However, a unit shift in the dual variable

function does not affect the relative impact of each sample on the expected loss. That is, the ranking of informativeness scores remains unchanged. Thus, assuming the supports of $\mathfrak{D}_{\mathbf{x}}$ and $\mathfrak{A}_{\mathbf{x}}$ coincide, ALLY is equivalent to a statistically consistent active learning method modulus the use of a dual function regularizer.

D Empirical Primal-Dual Learning Procedure in Algorithm 1

Even in the case were (CSL) and (D-CSL) are strongly dual, in practice we undertake the empirical version of (D-CSL) due to the challenges mentioned in Section 4.1 (i.e: $\mathfrak{D}$ is unknown and $\mathcal{F}$ is infinite dimensional). This requires introducing a *parameterization* of the hypothesis class $\mathcal{F}$ as $\mathcal{P} = \{f_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in \Theta\}$ and replacing expectations by sample means. In the following section, we present a high-level overview of some results from [30] on the implications of these two changes and the solution yielded by Algorithm PDCL.

D.1 Approximation Error

On one hand, approximating the function class $\mathcal{F}$ with a finite dimensional parametrization $\mathcal{P}$ transforms the dual problem (D-CSL) into:

$$\tilde{D}^{\star} = \max_{\lambda \in \Lambda} \min_{\theta \in \Theta} \tilde{L}\left(f_{\theta}, \lambda(\boldsymbol{x})\right), \tag{$\tilde{D}$-CSL}$$

where

$$\tilde{L}\left(f_{\theta}, \lambda(\boldsymbol{x})\right) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\left[\ell(f_{\theta}(\boldsymbol{x}), y) + \lambda(\boldsymbol{x})(\ell'(f_{\theta}(\boldsymbol{x}), y) - \epsilon(\boldsymbol{x}))\right].$$

Assuming that $\mathcal{P}$ is PAC learnable and that there is $\nu > 0$ such that for each $f \in \mathcal{F}$ there exists $f_{\theta} \in P$ that satisifes $\sup_{\boldsymbol{x}\in\mathcal{X}} |f_{\theta}(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \nu$, $f_{\theta^{\star}}$ is a *near optimal solution* of (CSL) [30] (Proposition 2).

Note that the assumption mentioned above is connected to the richness of the parameterization $\mathcal{P}$. For instance, $\mathcal{F}$ can denote C^0 (i.e: space of continuous functions) and $\mathcal{P}$ be a neural network, which meets the universal approximation assumption [70].

D.2 Estimation Error

On the other hand, approximating expectations by their sample means transforms the statistical dual problem ($\tilde{D}$-CSL) into its empirical counterpart (D-CERM):

$$\hat{D}^{\star} = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}), \tag{D-CERM}$$

This modification creates a difference between the optimal value of the parametrized dual problem and the optimal value of the empirical dual problem. However, assuming that the losses in question are $[0, B]$-valued and M-Lipschitz continuous -conditions which can be satisfied in the case of Cross-Entropy and Mean-Squared Error by setting bounds- this difference is bounded by a constant. In addition, this constant depends on the number of samples, the VC dimension of the parametrization $\mathcal{P}$, B and M. See [30] (Proposition 3) for a more detailed analysis.

In [30] (Theorem 3), the approximation and estimation errors are combined, and the sub-optimality of Algorithm PDCL with respect to (CSL) is bounded.

E Additional Experimental Settings

We do not use warm-starting or data augmentation. We employ early stopping on a validation set from the unlabelled pool and run each experiment five times with different random seeds. The patience in the early stopping callback was set to 2 epochs. We report the mean and standard deviations across the different seeds. We smooth the curves using exponential moving average for clarity and focus on discriminative regions of the learning curves. The dropout parameters for the dual prediction head were determined by cross-validation and are set to 0.3 and 0.25. The parameters of the neural networks are updated once at each iteration ($T_p = 1$) using the ADAM optimizer [71], and a learning

rate of $\eta_p = 0.001$ for SVHN, STL and CIFAR and $\eta_p = 0.005$ on MNIST. We adapt the architecture for STL-10 images ($96 \times 96 \times 3$) and split the dataset so as to have 11000 train samples and 2000 test samples (the original version has 5000 training samples and 8000 testing samples). This architecture has the same embedding size as ResNet-18 (i.e: 512). The dual variables are updated with stochastic gradient ascent and a learning rate of $\eta_d = 0.05$. During primal and dual steps we use a diminishing update rule for the step size. The number of layers of the dual variable prediction head was set by cross validation. All experiments were carried out on a system with the following specifications: Ubuntu 20.04, AMD Threadripper 3960X CPU and RTX 3090 GPU.

F Ablation on the Number of Clusters

We perform an ablation study on the number of clusters used by the k-MEANS algorithm. We use the MNIST dataset and the MLP architecture described in section 5. As shown in Figure 6, the performance of ALLY improves as the number of clusters grows. This suggests that, in this setting, prioritizing diversity over individual sample informativeness is beneficial in terms of average test accuracy.

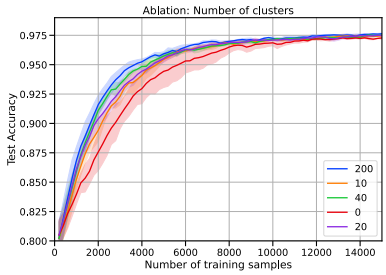

Figure 6: Ablation study on the number of clusters k used in ALLY.

G Experiment on Tiny ImageNet

Figure 7 compares the performance of ALLY with baselines methods on the Tiny ImageNet dataset [72], which is a subset of the ImageNet dataset (ILSVRC2012) consisting of $64 \times 64 \times 3$ images categorized in 200 classes.

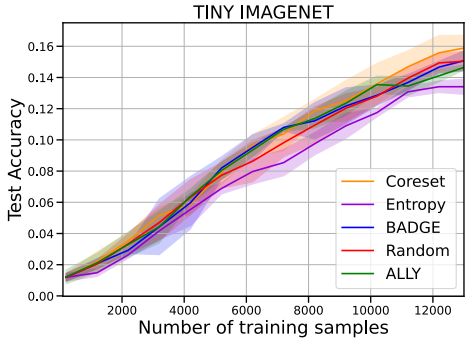

Figure 7: Empirical evaluation on the Tiny ImageNet dataset with $b = 1000$.