Label Noise: Correcting a Correction Loss

William Toner [1] **Amos Storkey** [1]

Abstract

Training neural network classifiers on datasets with label noise poses a risk of overfitting them to the noisy labels. To address this issue, researchers have explored alternative loss functions that aim to be more robust. However, many of these alternatives are heuristic in nature and still vulnerable to overfitting or underfitting. In this work, we propose a more direct approach to tackling overfitting caused by label noise. We observe that the presence of label noise implies a lower bound on the noisy generalised risk. Building upon this observation, we propose imposing a lower bound on the empirical risk during training to mitigate overfitting. Our main contribution is providing theoretical results that yield explicit, easily computable bounds on the minimum achievable noisy risk for different loss functions. We empirically demonstrate that using these bounds significantly enhances robustness in various settings, with virtually no additional computational cost.

1. Introduction

Over the last decade, we have seen an enormous improvement in the efficacy of machine learning methods for classification. Correspondingly, there has been an increased need for large labelled datasets to train these models. However, obtaining cleanly labelled datasets at the scale and quantity needed for industrial machine learning can be prohibitively expensive. For this reason, practitioners commonly rely on approaches which yield large datasets but contain high-label noise. Examples include web querying or crowd-sourcing systems. Even standard dataset collection methods are susceptible to noise introduced by fallible human labellers. This is especially true when data are hard to label, or labelling requires a specialist background (e.g. medical imaging). Such issues have led to immense interest in designing machine learning methods which can learn within the regime of noisy labels.

Most approaches for addressing the label noise problem consist of a mechanism for either removing or compensating for it. Unfortunately, many of these methods are elaborate or require pipelines involving multiple networks and stages (Li et al., 2020; Han et al., 2018; Jiang et al., 2018; Malach & Shalev-Shwartz, 2017; Li et al., 2023; Ren et al., 2018). This complexity damages their applicability in settings with user limitations on time, technical expertise or computational resources.

A simpler style of approach designs methods to be inherently resilient in the face of corrupted labels. The most prominent family of such methods are robust loss functions. Here the goal is to choose an objective function which allows training in the presence of noise without harming the generality of the learned classifier. An advantage of these methods is that they are simple; a robust loss can be easily implemented with minimal computational overhead. One such approach alters the cross-entropy objective to be less inclined to fit to noise (Wang et al., 2019; Zhang & Sabuncu, 2018; Ma et al., 2020). Similarly, regularisation and consistency-based approaches modify a loss with model-dependent terms or data cross-terms to restrict the network to avoid overfitting (Zhang et al., 2017; Liu et al., 2020; Englesson & Azizpour, 2021; Reed et al., 2014). Unfortunately, losses of these types are usually empirically rather than theoretically motivated, meaning the reasons for their robustness are rarely fully understood.

A more principled class of robust losses is *loss correction methods*. Here one estimates the noising distribution so that its impact may be subtracted from the training objective. (Patrini et al., 2017; Sukhbaatar et al., 2015; Goldberger & Ben-Reuven, 2016; Larsen et al., 1998; Mnih & Hinton, 2012). However, despite this correction, the training loss is minimised by fitting the noisy labels. Consequently, when using highly expressive neural network models, these losses remain susceptible to overfitting in the presence of label noise.

In this paper, we tackle the challenge of overfitting in popular robust losses by introducing a principled solution: bounding the allowable loss during training by recognising that the presence of label noise means the generalised noisy risk is lower bounded. The critical contribution of this paper is explicitly deriving these bounds and showing that their im-

[1]School of Informatics, University of Edinburgh, Edinburgh, UK. Correspondence to: William Toner <w.j.toner@sms.ed.ac.uk>, Amos Storkey <a.storkey.ed.ac.uk>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

plementation improves robustness. In addition, we provide a deeper understanding of existing robust loss functions, unifying correction losses with several other popular heuristic losses into a single family of generalised correction losses.

Key Idea: When a distribution contains label noise, this implies there is a minimum achievable risk. Current methods do not respect this bound, targeting a zero training loss instead. It is this which causes overfitting. By bounding the loss below during training, one may prevent overfitting. Crucially, these bounds are explicitly derivable.

The paper is structured as follows. Following a summary of the literature Section 2, we define in Section 3 a class of generalised correction losses, unifying a number of robust losses into one family, noting they still have the propensity to overfit. We then observe that the presence of label noise implies a lower bound on the achievable risk. We leverage this by introducing a *budget-corrected loss*, which lower bounds the loss during training. Later, in Section 4, we present a formula for generating these budgets. Finally, in Section 5, we experimentally test these budgets showing that they improve robustness across datasets and noise types.

Notation and Terminology: See Appendix A for a comprehensive overview of definitions, notation, and terminology.

2. Related Work

Corruption identification methods: Methods in this class handle label noise by identifying corrupted samples and then re-weighting, refining or removing these from the dataset. Many such methods rely on the heuristic that noisy samples have higher losses, especially earlier in training (Arpit et al., 2017). Song et al. (2019) use the entropy of the historical prediction distribution to identify refurbishable samples. Arazo et al. (2019) deploy a beta mixture model in the loss space and use the posterior probabilities that a sample is corrupted in the parameters of a bootstrapping loss. Zhou et al. (2020) define a loss which ignores samples that incur a higher loss value. Other approaches include a two-network model (Li et al., 2020) in which a Gaussian mixture model selects clean samples based on their loss values. These samples are then taken and used to train the other network. A number of other two-network models work on similar lines. Co-teaching (Han et al., 2018) trains two networks, one on the outputs of the other with the lowest loss values. Decoupling (Malach & Shalev-Shwartz, 2017) has the two networks update on the basis of disagreement with each other. Mentor-Net (Jiang et al., 2018) harnesses a teacher network for training a student network by re-weighting probably correct samples.

Other corruption identification methods may use the latent space to identify out-of-distribution data using an eigendecomposition (Kim et al., 2021) or KNN (Feng et al., 2021). Alternatives achieve consistency between different views of a data sample by training on convex combinations of data-label pairs (Zhang et al., 2017) or by minimising a Jensen-Shannon divergence between different augmentations (Englesson & Azizpour, 2021)

Loss-Based Methods: Methods in this class achieve robustness by altering the loss function to avoid overfitting to noise. One of the advantages here is the simplicity of these methods, as they do not require multiple networks or complex noise detection pipelines. This makes them suitable for plug-and-play use in any setting. **Correction-based methods** 'correct' the loss to compensate for the noising process (Larsen et al., 1998; Goldberger & Ben-Reuven, 2016). This procedure involves using noisy (Patrini et al., 2017) or clean data (Hendrycks et al., 2018) to infer the noise transition matrix. Despite the correction, such losses are still capable of overfitting (Patrini et al., 2017). An alternative set of methods consists of looking for innately **robust loss functions**. These methods are based on the observation that cross-entropy results in overfitting in the presence of label noise (Janocha & Czarnecki, 2016). Wang et al. (2019) propose a solution to this by adding a 'reverse cross-entropy' (RCE) term to the usual cross-entropy (CE) term. Janocha & Czarnecki (2016) observe that L^p-losses typically used for regression show good robustness in a classification setting. This is particularly true for the MAE loss (Mean Absolute Error), which exhibits good robustness albeit with a tendency to under-fit and train slowly (Ma et al., 2020). Following this observation, Generalised Cross-Entropy (Zhang & Sabuncu, 2018) construct a family of losses which interpolate between CE and L^1 in order to get the best of both. Ghosh & Kumar (2017) offer some theoretical insights, suggesting robustness may be obtained by choosing losses which are bounded or satisfy a 'symmetry' property. On the back of this Ma et al. (2020) show that one can take unbounded loss functions and re-normalise them to achieve this objective. Other methods soften or mix labels to avoid overfitting (Reed et al., 2014; Szegedy et al., 2016; Thiel, 2008), or use regularisers (Liu et al., 2020; Tanaka et al., 2018). Ishida et al. (2020), in line with our work, bound the loss to prevent overfitting. However, their method is only briefly discussed in relation to label noise and provides no mechanism for selecting this bound. We ground this work firmly in the context of label noise and provide theoretical results for producing bounds on the loss.

3. Robust Losses

In this section, we show how two types of losses may be partially unified by generalising correction-based losses to allow non-linear noise models. We call this family *f-proper losses*. This unification will aid in our subsequent analysis where we: (a) remark that despite enjoying certain theoretical guarantees, these losses are still prone to overfitting, (b) argue that the presence of label noise implies a lower bound

on the achievable risk, and (c) hypothesise that utilising this bound may mitigate overfitting. This leads us to propose bounding the empirical risk during training. We call this a *budget-corrected loss*.

Robust losses are a popular approach to tackling label noise by selecting losses less prone to fit the entire training set. In essence, one trades-off fitting power to gain improved robustness. Two well-known examples are the Generalised Cross-Entropy (GCE) and Symmetric Cross-Entropy (SCE) defined $\mathcal{L}_{GCE}(\vec{q},y=k) := \frac{1-q_k^a}{a}$ and $\mathcal{L}_{SCE}(\vec{q},y=k) = -log(q_k) + A(1-q_k)$ respectively.

Correction-based losses are an alternative motivated by the observation that, under label noise, the (noisy) empirical risk of a model is no longer a suitable proxy for its generalised unnoised risk (Defn A.4). However, by altering the loss through incorporating the noise model, one may fix this discrepancy. The most effective method is the *forward-corrected* loss (Patrini et al., 2017). Given a base loss $\mathcal{L}$, the forward-corrected loss is defined $\mathcal{L}^{\mapsto}(T^{-1}\vec{q}(x),\tilde{y}=k) := \mathcal{L}(\vec{q}(x),\tilde{y}=k)$ where T is a stochastic matrix approximating $p(\tilde{y}|y)$.

Unification. For forward-corrected losses, the loss and corrected loss are related by a linear transform T. We now show that generalising this, to allow T to be a non-linear transformation, we construct a family which unifies the forward-corrected loss with losses such as SCE and GCE. We call the resulting class of losses *f-proper*. This name reflects our requirement that the base loss be proper.

Definition 3.1 ($f-$proper Losses)**.** Let $\mathcal{L}$ be an elementwise loss and $f : \Delta^{c-1} \to \Delta^{c-1}$ be a bijective function on the probability simplex. We define $\mathcal{L}$ as (strictly) **f-proper** if there exists a (strictly) proper loss $\mathcal{L}_{\text{proper}}$ (as defined in A.3) such that for all $\vec{q} \in \Delta^{c-1}$, $\mathcal{L}(f(\vec{q}),i) = \mathcal{L}_{\text{proper}}(\vec{q},i)$. We refer to $\mathcal{L}_{\text{proper}}$ as the **base loss**.

This definition describes that a loss is a proper loss under an appropriate choice of transformation. E.g., in the context of label noise, this transformation can be a noise model. We can therefore interpret *$f-$proper losses* as a **generalised class of correction losses** where we permit non-linear noise models. This definition is broad, trivially including all proper losses such as cross-entropy (CE) when $f = id$. We now demonstrate that, in addition, the GCE, SCE and forward-corrected CE (FCE) losses mentioned above are all f-proper, deriving expressions for the transformations f. For the plots of the functions and a discussion of Definition 3.1, we refer to Appendices C.3 and C.2 respectively.

Lemma 3.2. *The GCE, SCE and FCE losses are all strictly $f-$proper where* $f_{GCE}(\vec{p})_i = \frac{p_i^{\frac{1}{1-a}}}{\sum_{i=1}^{c} p_i^{\frac{1}{1-a}}}$, $f_{SCE}(\vec{p})_i = \frac{p_i}{\lambda - A p_i}$ *and* $f_{FCE}(\vec{p}) = T^{-1}\vec{p}$. *Here T is the invertible stochastic matrix used to define the correction, and λ is a constant selected to ensure the correct normalisation.*

Lemma 3.2 demonstrates that GCE and SCE are non-linear correction-based losses; the noise model is represented by the function f^{-1}. We stress that these are by no means the only robust losses which adhere to Definition 3.1. However, these losses permit us to compute f explicitly. For this reason, they provide useful examples when empirically demonstrating the results of Section 4.

3.1. The Problem of Overfitting

The unification established by Definition 3.1 provides two main benefits. It improves our understanding of losses like GCE since we show that they encode an implicit noise model. More importantly, this grants a common framework for analysing the failure modes of these losses. In this section, our analysis demonstrates that correcting for the noise model alone is insufficient. To achieve robustness, we must additionally correct our loss by imposing a lower bound on the training loss to account for the randomness introduced by label noise.

The original motivation for noise-corrected losses recognises that, in the presence of label noise, the noisy empirical risk ($\hat{R}^{\eta}_{\mathcal{L}}$) no longer adequately approximates the generalised clean risk $R_{\mathcal{L}}$ (Defn A.5). However, by modifying the loss as in Defn. 3.1 by defining $\mathcal{L}(f(\vec{q}),i) := \mathcal{L}_{pr.}(\vec{q},i)$, one can ensure that $R^{\eta}_{\mathcal{L}}(q) = R_{\mathcal{L}_{pr.}}(q)$. This holds when f adequately models the true noising process. Despite this correction, the empirical risk is minimised by precisely fitting the noisy labels. Consequently, training with a highly expressive neural network still incurs overfitting, similar to that observed with the uncorrected losses.

Ideally, we want to train our model to fit the clean labels without overfitting the noisy ones. Our insight is realising that we can do this in a principled way. When a label distribution contains noise, there is a lower bound on the optimal noisy risk any model can achieve. An analogy to this is that no forecaster can predict the outcome of a biased coin flip 100% of the time: For example, if the bias is 70%, we can't expect any model to do better 70% over a large number of flips. Likewise, even if an optimal model $q^*(x)$ exists, which minimises the noisy risk, we *still* expect it to incur a non-zero loss on a randomly sampled noisy training set. Consequently, a model whose training loss is much below this irreducible error has necessarily overfit to the noise.

We propose, therefore, that the principled way to handle label noise is to limit the minimum allowable risk on the training set. Specifically, we define a budget B and train so that our loss does not go below this value. Explicitly we define this as follows:

Definition 3.3 (Budget Loss)**.** Let $\mathcal{L}$ be an elementwise loss. Let $\mathcal{D}$ be a batch of N data-label pairs (x_i,y_i). Define the

$B-$**corrected** (batchwise) loss $\mathcal{L}_B$ as follows:

$$\mathcal{L}_B(\vec{q}(x),\mathcal{D}) := |B - \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\vec{q}(x_i),y_i)| \tag{1}$$

Previously, Ishida et al. (2020) have explored the idea of targeting a specific non-zero loss value. We want to remark our contribution goes beyond that, providing a theoretically justified toolset for selecting the budget for a broad family of loss functions. We achieve this by constructing bounds for the minimum achievable noisy risk of $f-$ proper losses.

4. Risk Bounds

In the last section, we remarked that despite theoretical motivation, $f-$proper losses are still prone to overfitting. We proposed training against a budget to prevent this, noting that the optimal generalised noisy risk is non-zero. In this section, we make this more concrete, giving bounds for the generalised noisy $\mathcal{L}$-risk of the optimal probability estimator for $f-$proper losses. We conclude by using these bounds to derive a formula for choosing a budget B to train against in Defn. 3.3. We call this the *noise-corrected budget*. This budget is simple to compute, depending only on the loss, the aggregate noise rate η and the number of classes c. Throughout this section, we adhere to the notation and definitions given in Appendix A.

We begin by presenting our bound in its most general form. Following this, we make a simplifying assumption, which allows us to derive more useable bounds. The proof is given in Appendix B.

Lemma 4.1. . *Let $\mathcal{L}$ be an elementwise, strictly $f-$proper loss. For any probability estimator $\vec{q}(x)$, we may derive a bound for the noisy pointwise risk that is tight and obtained when $\vec{q}(x)=\tilde{p}(\tilde{y}|x)$. Specifically, for all $x \in supp(p(x))$ we have;*

$$R^{\eta}_{\mathcal{L}}(\vec{q})(x) \geq R^{\eta}_{\mathcal{L}}(f(\tilde{p}(\tilde{y}|x))) \tag{2}$$

The key assumption which we employ to simplify Lemma 4.1 is that the clean label distribution is (approximately) deterministic. We define this as follows.

Definition 4.2 (Deterministic)**.** We say that $p(y|x)$ is **deterministic** when, for each $x \in supp(p(x))$, there exists $k(x) \in \mathcal{Y}$ such that $p(y=k(x)|x)=1$.

Definition 4.2 describes a scenario where the label is considered a deterministic function of the input. i.e. there is no randomness in the label distribution. While an idealised assumption, it is a reasonable approximation for many real-world image classification tasks where clear images of a single subject dominate the dataset. In domains with high inherent randomness, such as medical diagnostics, this assumption is less suitable. In the following, we leverage this assumption to construct a number of bounds. Later we discuss the validity of this assumption and the impact on the derived bounds.

4.1. Entropy Bounds

The following characterisation of proper losses is indispensable for the statements and proofs of the following results.

Theorem 4.3 (Savage (1971))**.** *Let $\mathcal{L}$ be some elementwise loss. $\mathcal{L}$ is proper if and only if there exists a concave function $J:\Delta^{c-1}\to\mathbb{R}$ such that $\mathcal{L}(\vec{q},i) := J(\vec{q}) + (\vec{e}_i - \vec{q})\cdot\nabla J(\vec{q})$. We call J the **entropy function** of $\mathcal{L}$ where $J(\vec{q}) = \sum_{i=1}^{c} q_i \mathcal{L}(\vec{q},i)$. Example: When $J(\vec{q}) := -\sum_{i=1}^{c} q_i log(q_i)$ (Shannon Entropy) one recovers the cross-entropy loss.*

Definition 4.4. We call a loss **symmetric** if it is invariant under a permutation of the labels. In practice this means the loss has no inherent bias toward any particular class. We call a proper loss symmetric if the entropy J is a symmetric function of its variables, e.g. $J(a,b)=J(b,a)$ for all a,b.

Lemma 4.5. *Let $p(x,y)$ be a distribution where $p(y|x)$ is deterministic, and let $\tilde{p}(x,\tilde{y})$ be a noisy distribution obtained by applying label noise to $p(x,y)$. Assume that $\mathcal{L}$ is a symmetric (strictly) f-proper loss and let J denote the entropy function of its base loss. For any probability estimator q, we can derive a bound for the noisy risk when the label noise is symmetric that is achieved (uniquely) if $q(x)=\tilde{p}(\tilde{y}|x)$. Specifically, for each x, we have the following lower bound, where $\eta(x)$ denotes the noise rate at x:*

$$R^{\eta}_{\mathcal{L}}(q) \geq \mathbb{E}_{x\sim p(x)}[J\Big((1-\eta(x), \frac{\eta(x)}{c-1}, \frac{\eta(x)}{c-1}, ..., \frac{\eta(x)}{c-1})\Big)]$$

Corollary 4.6. *When the label noise is uniform and symmetric with rate η, the following bound on the risk of any probability estimator is tight and achieved only when $q(x)=\tilde{p}(\tilde{y}\,|\,x)$.*

$$R^{\eta}_{\mathcal{L}}(q) \geq J\Big((1-\eta, \frac{\eta}{c-1}, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1})\Big) \tag{3}$$

Corollary 4.7. *Let $\mathcal{L}$ be a symmetric (strictly) f-proper loss and let $\vec{u}_c(\eta) := (1-\eta, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1})$. We have the following bound on the noisy risk if the label noise is symmetric and uniform, which is obtained (uniquely) by setting $\vec{q}(x)=\tilde{p}(\tilde{y}|x)$ for all $x \in \text{supp}(p(x))$.*

$$R^{\eta}_{\mathcal{L}}(q) \geq (1-\eta)\mathcal{L}(f(\vec{u}_c(\eta)),1) + \eta\mathcal{L}(f(\vec{u}_c(\eta)), i\neq 1)$$

Proofs are given in Appendix B.

Corollaries 4.6, 4.7 give bounds on the noisy risk in terms of the aggregate noise rate and the number of classes (η,c respectively). Specifically, the risk of any model q is bounded below by the mean entropy of the noisy label distribution. These bounds hold if the label noise is symmetric and uniform (Defn A.5). When noise deviates

from these idealised assumptions, the minimum achievable risk is (typically) lower (See Corollary C.2). How much lower is determined by the entropy function J. This topic is discussed further in Appendix C.

Using Corollary 4.7, we define the budget for our budget-corrected loss (Defn. 3.3).

Definition 4.8 (Noise-Corrected Budget)**.** Let $\mathcal{L}$ be a symmetric $f-$proper loss whose base loss has entropy function J. Let $\vec{u}_c(\eta) := (1-\eta, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1})$. We define the **noise-corrected budget** as:

$$\boxed{B(\eta,c) := (1-\eta)\mathcal{L}(f(\vec{u}_c(\eta)),1) + \eta\mathcal{L}(f(\vec{u}_c(\eta)),2)} \quad (4)$$

This leads us to the main proposal of this paper. When our dataset has label noise, we propose using the budget loss (Eqn. 1) with B set to $B(\eta,c)$, the noise-corrected budget from Eqn. 4. We call this the **noise-corrected budget loss**. For CE and FCE, the noise-corrected budget corresponds to the Shannon Entropy of the distribution $(1-\eta, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1})$. For SCE and GCE, we substitute expressions for f derived in Lemma 3.2 into Eqn. 4 to generate budgets. These are given explicitly in Appendix C.4.

5. Experiments

Losses: In this section, we empirically investigate the effectiveness of the noise-corrected budget loss (Eqn. 1 with Eqn. 4) for improving robustness to label noise. We consider several loss functions: CE, SCE, forward-corrected CE (FCE), and GCE. Additionally, we explore a variant of CE that includes a prior on the model probabilities (CEP). Our experiments all follow a similar structure. We use a dataset containing intrinsic or synthetic label noise in the training set. We train models using each loss on this noisy training set and evaluate their performance on a clean test set. We compare the results obtained without budgeting and when using noise-corrected budgets. The latter are denoted by a '+B' after the loss name (e.g., CE+B). We also explore treating the budget as a hyperparameter in order to assess the optimality of the noise-corrected budget. These experiments are indicated by an asterisk (e.g., CE+B*). We compare against other standard baseline losses; details may be found in Appendix D.1.

Datasets: We evaluate each loss on various datasets with different label noise types. We consider two versions of Cifar100; one corrupted with symmetric label noise at rates of 0.2, 0.4 and one with asymmetric label noise (Asym-Cifar100) at the same rates. We also evaluate on a version of EMNIST corrupted by non-uniform noise. Precise experimental details (including how the label noise is constructed for these datasets) and further experiments with synthetic noise (EMNIST, FashionMNIST, Cifar10, MNIST) and real, intrinsic open-set noise (TinyImageNet and Animals-10N), are given in Appendix D.

5.1. Results

The results of our experiments are presented in Table 1. Additional results for other datasets are presented in tables in Appendix D. Each table follows a similar structure, with losses listed in rows and datasets in columns. The baselines are grouped together at the top. Our main losses are organised into triplets, such as CE, CE+B, and CE+B*. The rows that use the noise-corrected budget (e.g., GCE+B) are highlighted in grey to enhance readability. If using our noise-corrected budget leads to higher mean accuracy compared to training without the budget, this is indicated by a box. The best overall model for each dataset is highlighted by colouring the given cell light yellow.

With few exceptions, our budget leads to improved performance compared to the standard version of each loss. For the Asym-Cifar100 and Non-Uniform-EMNIST datasets, our CE+B loss performs worse than regular CE. This outcome is expected since our derived budgets are only optimal for symmetric noise and may be suboptimal for non-symmetric noise. This discrepancy is especially pronounced for losses based on Shannon-Entropy like CE (Appendix C.1). In contrast, the other f-proper losses, as we had anticipated, exhibit greater resilience to the precise noise structure and consistently outperform the baseline across different types of noise.

Our optimal budget is attained by doing a grid search in a small vicinity of the noise-corrected budget. When this doesn't yield an improvement, the starred and unstarred accuracy values are the same. In around half of our experiments, we find that we achieve an improvement by perturbing the budget. This improvement is generally minor. Our assumption that the underlying clean dataset is deterministic (Defn 4.2) means one should be able to improve performance by raising the budget to account for the additional randomness in the label distributions. Generally, we find this to be so. An exception to this are the non-uniform and asymmetric datasets. In these cases, one typically benefits from marginally lowering the budget. These observations are consistent with our expectations, as the bounds are tight only for symmetric noise and will otherwise be overly strict. The values of the optimal budgets may be found in a table in Appendix D.1.2.

Figure 1 presents a visualisation of how test accuracy changes with the budget used during training on noisy CIFAR10 (left) and EMNIST (right) datasets. For CIFAR10 we train using an SCE loss. For the EMNIST dataset, we use a CE loss. The training set of each dataset has been corrupted with symmetric noise at a rate of $\eta = 0.4$. We employ a budget with an offset from the noise-corrected budget: $B(\epsilon) := B(\eta = 0.4, c = 10) + \epsilon$, and plot the clean test performance against the value of ϵ. Thus $\epsilon = 0$ corresponds to the budget given in 4. The graphs in Figure 1 exhibit similar patterns. As ϵ increases, the test performance improves due to the budget preventing overfitting. A peak is reached at around $\epsilon = 0.35, 0.15$ respectively,

Losses	CIFAR100 0.2 Top1	CIFAR100 0.2 Top5	CIFAR100 0.4 Top1	CIFAR100 0.4 Top5	ASYM-CIFAR100 0.2 Top1	ASYM-CIFAR100 0.2 Top5	ASYM-CIFAR100 0.4 Top1	ASYM-CIFAR100 0.4 Top5	Non-Uniform-EMNIST 0.6 Top 1	Non-Uniform-EMNIST 0.6 Top 5
MSE	$57.2_{\pm 0.93}$	$78.6_{\pm 0.25}$	$40.6_{\pm 0.38}$	$63.0_{\pm 0.24}$	$56.3_{\pm 0.11}$	$82.6_{\pm 0.22}$	$40.7_{\pm 0.12}$	$74.4_{\pm 0.25}$	$44.7_{\pm 2.66}$	$86.7_{\pm 3.10}$
MAE	$10.0_{\pm 0.11}$	$13.8_{\pm 0.28}$	$7.6_{\pm 1.89}$	$11.6_{\pm 1.25}$	$7.1_{\pm 6.02}$	$11.1_{\pm 6.6}$	$11.1_{\pm 5.43}$	$25.1_{\pm 5.76}$	$9.8_{\pm 1.74}$	$23.1_{\pm 1.80}$
NCE	$38.7_{\pm 3.13}$	$51.8_{\pm 3.77}$	$19.1_{\pm 0.20}$	$28.8_{\pm 0.15}$	$16.3_{\pm 1.24}$	$25.4_{\pm 1.80}$	$21.8_{\pm 1.24}$	$37.2_{\pm 1.80}$	$18.0_{\pm 1.17}$	$38.8_{\pm 1.93}$
MixUp	$59.6_{\pm 0.31}$	$81.5_{\pm 0.39}$	$51.3_{\pm 8.63}$	$75.8_{\pm 8.09}$	$61.2_{\pm 0.88}$	$86.0_{\pm 1.12}$	$47.2_{\pm 0.60}$	$81.3_{\pm 0.23}$	$\mathbf{52.4}_{\pm 0.80}$	$95.5_{\pm 0.08}$
Sph.	$57.7_{\pm 0.18}$	$82.9_{\pm 0.54}$	$48.8_{\pm 0.51}$	$74.3_{\pm 0.73}$	$54.2_{\pm 0.32}$	$81.2_{\pm 0.29}$	$39.2_{\pm 0.31}$	$72.1_{\pm 0.15}$	$41.9_{\pm 0.10}$	$94.4_{\pm 0.04}$
Boot.	$54.0_{\pm 0.37}$	$76.4_{\pm 0.39}$	$37.7_{\pm 0.89}$	$60.9_{\pm 1.52}$	$56.0_{\pm 0.34}$	$83.8_{\pm 0.03}$	$43.2_{\pm 0.35}$	$78.3_{\pm 0.20}$	$49.1_{\pm 0.29}$	$95.3_{\pm 0.42}$
Trunc.	$58.1_{\pm 0.36}$	$82.7_{\pm 0.37}$	$50.9_{\pm 1.17}$	$77.2_{\pm 0.59}$	$56.3_{\pm 0.62}$	$82.3_{\pm 0.61}$	$45.2_{\pm 0.81}$	$75.6_{\pm 0.29}$	$23.7_{\pm 0.98}$	$40.1_{\pm 1.24}$
CL	$53.0_{\pm 0.21}$	$76.3_{\pm 0.19}$	$36.3_{\pm 0.77}$	$60.1_{\pm 0.66}$	$55.3_{\pm 0.48}$	$83.5_{\pm 0.28}$	$42.4_{\pm 0.45}$	$78.1_{\pm 0.14}$	$48.2_{\pm 0.45}$	$95.0_{\pm 0.04}$
ELR	$10.4_{\pm 0.24}$	$31.7_{\pm 0.44}$	$10.0_{\pm 0.64}$	$30.1_{\pm 0.88}$	$10.8_{\pm 0.21}$	$32.7_{\pm 0.53}$	$10.3_{\pm 0.39}$	$30.8_{\pm 0.35}$	$40.3_{\pm 0.39}$	$93.0_{\pm 0.24}$
FCE	$56.9_{\pm 0.58}$	$79.2_{\pm 0.14}$	$43.7_{\pm 0.15}$	$66.2_{\pm 0.19}$	$55.3_{\pm 0.54}$	$83.5_{\pm 0.24}$	$41.4_{\pm 0.55}$	$77.3_{\pm 0.75}$	$39.0_{\pm 0.05}$	$67.8_{\pm 0.47}$
FCE+B	$56.1_{\pm 2.22}$	$81.8_{\pm 1.37}$	$50.2_{\pm 0.02}$	$77.2_{\pm 0.19}$	$54.2_{\pm 0.44}$	$83.3_{\pm 0.43}$	$43.8_{\pm 0.02}$	$77.5_{\pm 0.13}$	$40.0_{\pm 0.35}$	$73.2_{\pm 0.08}$
FCE+B*	$56.1_{\pm 2.22}$	$82.2_{\pm 0.39}$	$50.2_{\pm 0.02}$	$77.2_{\pm 0.19}$	$54.2_{\pm 0.44}$	$83.4_{\pm 0.24}$	$45.1_{\pm 0.37}$	$79.9_{\pm 0.24}$	$43.1_{\pm 0.40}$	$79.4_{\pm 0.12}$
GCE	$60.0_{\pm 0.13}$	$82.6_{\pm 0.63}$	$44.9_{\pm 0.07}$	$67.2_{\pm 0.34}$	$53.8_{\pm 0.55}$	$81.6_{\pm 0.14}$	$39.4_{\pm 0.44}$	$74.0_{\pm 0.36}$	$44.8_{\pm 0.62}$	$91.2_{\pm 0.70}$
GCE+B	$59.4_{\pm 0.02}$	$83.5_{\pm 0.24}$	$50.3_{\pm 0.11}$	$75.3_{\pm 0.64}$	$55.4_{\pm 0.55}$	$83.0_{\pm 0.35}$	$46.5_{\pm 1.44}$	$77.7_{\pm 0.35}$	$47.1_{\pm 0.20}$	$93.5_{\pm 0.43}$
GCE+B*	$61.0_{\pm 1.33}$	$83.9_{\pm 0.74}$	$50.3_{\pm 0.11}$	$75.3_{\pm 0.64}$	$56.6_{\pm 0.10}$	$83.8_{\pm 0.88}$	$47.7_{\pm 0.35}$	$77.9_{\pm 0.03}$	$47.1_{\pm 0.20}$	$93.5_{\pm 0.43}$
SCE	$55.9_{\pm 0.53}$	$76.5_{\pm 0.15}$	$38.7_{\pm 0.60}$	$60.9_{\pm 0.41}$	$57.5_{\pm 0.19}$	$83.7_{\pm 0.17}$	$43.3_{\pm 0.87}$	$77.5_{\pm 0.75}$	$47.2_{\pm 0.33}$	$92.5_{\pm 0.01}$
SCE+B	$55.5_{\pm 0.90}$	$77.4_{\pm 0.84}$	$47.1_{\pm 1.32}$	$69.2_{\pm 1.18}$	$57.9_{\pm 0.83}$	$83.7_{\pm 0.41}$	$50.0_{\pm 1.62}$	$80.4_{\pm 0.65}$	$47.9_{\pm 0.80}$	$93.8_{\pm 0.05}$
SCE+B*	$56.6_{\pm 1.07}$	$78.5_{\pm 0.88}$	$47.3_{\pm 1.16}$	$69.6_{\pm 0.90}$	$57.9_{\pm 0.83}$	$83.7_{\pm 0.41}$	$50.0_{\pm 1.62}$	$80.4_{\pm 0.65}$	$47.9_{\pm 0.80}$	$93.8_{\pm 0.05}$
CE	$52.3_{\pm 1.35}$	$75.6_{\pm 0.93}$	$35.3_{\pm 1.14}$	$59.3_{\pm 0.81}$	$54.9_{\pm 0.12}$	$83.3_{\pm 0.25}$	$42.4_{\pm 0.16}$	$78.9_{\pm 0.56}$	$48.6_{\pm 0.11}$	$95.3_{\pm 0.10}$
CE+B	$50.9_{\pm 1.01}$	$76.5_{\pm 0.86}$	$39.9_{\pm 1.02}$	$65.8_{\pm 1.19}$	$52.9_{\pm 1.86}$	$83.2_{\pm 0.88}$	$34.7_{\pm 2.51}$	$73.4_{\pm 1.50}$	$45.5_{\pm 5.11}$	$93.0_{\pm 0.16}$
CE+B*	$50.9_{\pm 1.01}$	$78.2_{\pm 1.16}$	$39.9_{\pm 1.02}$	$68.1_{\pm 0.63}$	$53.3_{\pm 0.89}$	$83.2_{\pm 0.88}$	$45.9_{\pm 0.40}$	$79.7_{\pm 0.29}$	$50.2_{\pm 0.35}$	$\mathbf{95.9}_{\pm 0.14}$
CEP	$58.8_{\pm 0.87}$	$78.6_{\pm 0.38}$	$43.5_{\pm 0.24}$	$65.1_{\pm 1.27}$	$59.4_{\pm 0.08}$	$82.2_{\pm 0.03}$	$46.5_{\pm 0.17}$	$76.4_{\pm 0.25}$	$48.2_{\pm 0.05}$	$95.4_{\pm 0.07}$
CEP+B	$\mathbf{62.3}_{\pm 0.87}$	$\mathbf{85.1}_{\pm 0.46}$	$54.3_{\pm 0.86}$	$79.2_{\pm 0.93}$	$63.0_{\pm 0.92}$	$\mathbf{87.5}_{\pm 0.32}$	$53.0_{\pm 0.28}$	$82.8_{\pm 0.13}$	$45.0_{\pm 0.48}$	$95.0_{\pm 0.08}$
CEP+B*	$62.9_{\pm 0.79}$	$85.1_{\pm 0.46}$	$\mathbf{55.3}_{\pm 0.37}$	$\mathbf{79.8}_{\pm 0.08}$	$\mathbf{63.0}_{\pm 0.14}$	$87.5_{\pm 0.32}$	$\mathbf{55.6}_{\pm 0.66}$	$\mathbf{83.8}_{\pm 0.11}$	$47.7_{\pm 0.19}$	$95.9_{\pm 0.23}$

Table 1. Test accuracies for different losses on the noisy CIFAR100/Asym-CIFAR100/Non-Uniform EMNIST datasets. Losses implementing the Noise-Corrected Budget shaded in grey. When using this budget provides benefit, the corresponding value is boxed. Overall top values in yellow.

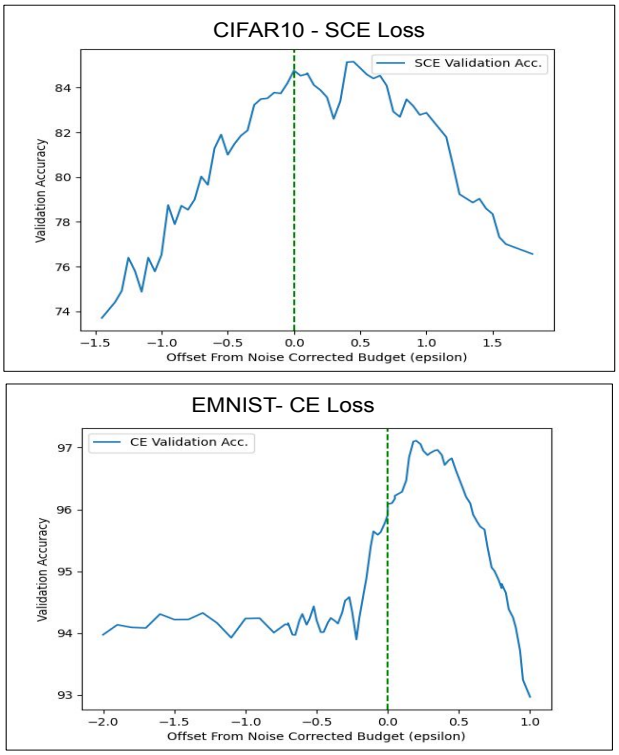

Figure 1. We plot the test accuracy as a function of the budget for noisy Cifar10/EMNIST using SCE/CE losses respectively. The $x-$axis is normalised so that our noise-corrected budget (Defn 4) is centred at zero. This is highlighted by the green dotted line. Both graphs show a bump with a peak near our chosen budget value.

followed by a decline in performance as the model starts to underfit the noisy data. The presence of a prominent bump and its proximity to our noise-corrected budget provides empirical evidence supporting our theoretical framework. Notably, the optimal budget seems to lie at $\epsilon > 0$, which we attribute to the intrinsic entropy present in the dataset. This deviation originates from our simplifying assumption, which modelled the underlying distributions as deterministic (Defn 4.2).

6. Conclusion, Limitations and Further Work

In this work, we have looked at mitigating the impact of label noise by training subject to a budget. We motivated this by noting that label noise implies a minimum achievable risk. We then concretely constructed these lower bounds for various losses. Later we empirically showed that the derived budgets indeed substantially improved the performance.

While our method proves successful in the settings we looked at, its applicability has a few limitations. One limitation is that our method requires an approximation for the aggregate noise rate. A second limitation is that the derived bounds are reliable if the dataset is well-approximated as deterministic. For datasets with high inherent randomness, such as in the medical domain, the proposed budgets might not suffice. An avenue for future work is extending our results to a more noisy data environment.

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N., and McGuinness, K. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pp. 312–321. PMLR, 2019.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

Chou, A. Truncated losses. https://github.com/AlanChou/Truncated-Loss/, 2019.

Englesson, E. and Azizpour, H. Generalized Jensen-Shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021.

Feng, C., Tzimiropoulos, G., and Patras, I. S3: supervised self-supervised learning under label noise. *CoRR*, abs/2111.11288, 2021. URL https://arxiv.org/abs/2111.11288.

Ghosh, A. and Kumar, H. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Goldberger, J. and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer. 2016.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018.

Ishida, T., Yamane, I., Sakai, T., Niu, G., and Sugiyama, M. Do we need zero training loss after achieving zero training error? In *International Conference on Machine Learning*, pp. 4604–4614. PMLR, 2020.

Janocha, K. and Czarnecki, W. M. On loss functions for deep neural networks in classification. *Schedae Informaticae*, 25:49, 2016.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pp. 2304–2313. PMLR, 2018.

Kim, T., Ko, J., Choi, J., Yun, S.-Y., et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24137–24149, 2021.

Larsen, J., Nonboe, L., Hintz-Madsen, M., and Hansen, L. Design of robust neural network classifiers. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 2, pp. 1205–1208 vol.2, 1998. doi: 10.1109/ICASSP.1998.675487.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Li, X.-C., Xia, X., Zhu, F., Liu, T., yao Zhang, X., and lin Liu, C. Dynamic loss for learning with label noise, 2023. URL https://openreview.net/forum?id=J_kUIC1DNHJ.

Liu, S. Truncated-loss. https://github.com/shengliu66/ELR, 2020.

Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Ma, X., Huang, H., Wang, Y., Erfani, S. R. S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Malach, E. and Shalev-Shwartz, S. Decoupling" when to update" from" how to update". *Advances in neural information processing systems*, 30, 2017.

Mnih, V. and Hinton, G. E. Learning to label aerial images from noisy data. In *Proceedings of the 29th International conference on machine learning (ICML-12)*, pp. 567–574, 2012.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Reed, S. E., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. 2014.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Savage, L. J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Song, H., Kim, M., and Lee, J.-G. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pp. 5907–5915. PMLR, 2019.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5552–5560, 2018.

Thiel, C. Classification on soft labels is robust against label noise. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 65–73. Springer, 2008.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 322–330, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00041. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00041.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv e-prints*, pp. arXiv–1710, 2017.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Zhou, T., Wang, S., and Bilmes, J. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*, 2020.

A. Problem Formulation

A.1. Classification via Empirical Risk Minimisation – Formalisation and Notation

Given a dataset of data-label pairs, the goal of discriminative classification is to learn a mapping from data to label space which predicts the labels of unseen datapoints with high accuracy. One way in which this classifier can be learned is through a process known as Empirical Risk Mimimisation (ERM) on a surrogate loss function. Here one defines a (parametric) family of models and selects the parameter which minimises a quantity called the 'empirical risk' defined using the dataset.

Let $\mathcal{X} \subset \mathbb{R}^d$ be some data domain and let $\mathcal{Y} = \{1,2,3,...c\}$ be some label domain. Suppose there exists some latent distribution $p(x,y)$ over data-label space and we have a dataset $\{(x_i,y_i)\}_{i=1}^N$ of samples drawn independent and identically distributed (iid) from it.

Definition A.1 (Probability Estimator)**.** Let Δ^{c-1} denote the **probability simplex** of $c-$dimensional non-negative vectors which sum to one. A **probability estimator** $\vec{q}: \mathcal{X} \to \Delta^{c-1}$ is a model which takes a point in dataspace and outputs a distribution over labels. We will denote this as $\vec{q}(x) := (q(y=1|x), q(y=2|x), ..., q(y=c|x))$.

Definition A.2 (Elementwise/Batchwise Losses)**.** An **elementwise** loss function is a function which takes a predicted distribution over labels and evaluates it against an observed label. That is $\mathcal{L}: \Delta^{c-1} \times \mathcal{Y} \to \mathbb{R}$. A **batchwise** loss evaluates a batch of predictions against a corresponding batch of labels $\mathcal{L}: (\Delta^{c-1} \times \mathcal{Y})^N \to \mathbb{R}$.

Definition A.3 (Proper)**.** Let $\mathcal{L}$ be an elementwise loss. We define $\mathcal{L}$ as **proper** if, for any given $\vec{p}$, the expected loss $L_{\mathcal{L}}(\vec{q},\vec{p}) := \sum_{i=1}^c p_i \mathcal{L}(\vec{q}, y=i)$ is minimised by setting $\vec{q} = \vec{p}$. If $\vec{p}$ is the unique minimising point, we refer to $\mathcal{L}$ as **strictly proper**.

Definition A.4 (Pointwise Risk)**.** Given some parametric probability estimator $\vec{q}_\theta(x)$, a distribution $p(x, y)$ and an elementwise loss $\mathcal{L}$, we define the **Pointwise $\mathcal{L}-$Risk** of q_θ at x to be the expected loss of $\vec{q}_\theta(x)$ under the label distribution at $p(y|x)$. Formally, $R_{\mathcal{L}}(q_\theta)(x) := \mathbb{E}_{y \sim p(y|x)}[\mathcal{L}(q_\theta(x), y)])$

We then define the **Generalised $\mathcal{L}$-Risk** of q_θ with respect to $p(x,y)$ as the expectation of the pointwise risk with respect to $p(x)$, $R_{\mathcal{L}}(q_\theta) := \mathbb{E}_{x \sim p(x)}[R_{\mathcal{L}}(q_\theta)(x)]$. The **Empirical $\mathcal{L}$-Risk** is defined by approximating this expression on a dataset of samples drawn from p. We distinguish this from the former using a hat $\hat{R}_{\mathcal{L}}(q_\theta) := \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\vec{q}_\theta(x_i), y_i)$

Our aim, with a given parametric family of probability estimators, is to select θ^* that minimises the expected misclassification rate on a randomly sampled test set. Typically, *surrogate losses* are minimised during training, in part to overcome the challenge of optimising $0-1$ loss [1]. By far, the most commonly used surrogate is the cross-entropy loss (CE).

A.2. Label Noise

The standard classification approaches are practical on datasets with clean labels. Unfortunately, datasets are frequently contaminated with labelling errors (also referred to as label noise) in real-world scenarios, which can significantly degrade performance. Formally,

Definition A.5 (Label Noise)**.** Label noise refers to any process that randomly modifies the labels of samples drawn from a distribution over data-label space, denoted as $p(x,y)$. In particular, we consider label noise that can be modelled using a noising distribution $p(\tilde{y}|y,x)$, which specifies the transition probabilities between the clean label y and the noisy label $\tilde{y}$ for each data point x. The resulting noisy distribution is denoted as $\tilde{p}(x,\tilde{y}) = \sum_y p(\tilde{y}|y,x)p(x,y)$.

For each data point x, the noising distribution can be represented by a transition matrix $T(x)$, where $T(x)_{ij} := p(\tilde{y}=j|y=i,x)$. When the noising distribution is independent of x, we refer to the label noise as **uniform**. We say that the label noise is **symmetric** at x if $T(x)_{ii} = 1 - \eta(x)$ and $T_{ij} = \frac{\eta(x)}{c-1}, i \neq j$, where $\eta(x)$ is the noise rate at x and c is the number of distinct labels. More generally, the noise rate at x is defined as the probability that a label is corrupted at that point: $\eta(x) := p(\tilde{y} \neq y|x) = 1 - \sum_{i=1}^c T_{ii}(x) p(y=i|x)$.

We use the term **noisy risk** when our risks (Defn A.4) are evaluated with respect to the noisy distribution $\tilde{p}(x,\tilde{y})$. We denote these as $R^\eta, \hat{R}^\eta$ respectively. This distinguishes them from the **clean risks** evaluated with respect to the un-noised distribution $p(x,y)$.

B. Proofs

In this section we give proofs for statements given in the main paper and additional statements from Appendix C

[1] The $0-1$ loss is 1 if a sample is misclassified and 0 otherwise

Lemma 3.2 The GCE, SCE and forward-corrected CE (FCE) losses are all strictly $f-$proper where $f_{GCE}(\vec{p})_i = \frac{p_i^{\frac{1}{1-a}}}{\sum_{i=1}^c p_i^{\frac{1}{1-a}}}$, $f_{SCE}(\vec{p})_i = \frac{p_i}{\lambda - Ap_i}$ and $f_{FCE}(\vec{p}) = T^{-1}\vec{p}$. Here T is the invertible stochastic matrix used to define the correction, and λ is a constant selected to ensure the correct normalisation.

Proof. We begin by introducing the following notation: Let $\mathcal{L}$ be an elementwise loss and let $\vec{p},\vec{q}$ be two distributions, we denote the expected loss of $\vec{q}$ with respect to $\vec{p}$ to be $L_{\mathcal{L}}(\vec{q},\vec{p}) := \sum_{i=1}^c p_i \mathcal{L}(\vec{q},i)$.

Let us begin by considering GCE. The expected loss may be written $L_{GCE}(\vec{q},\vec{p}) := \sum_{i=1}^c p_i \mathcal{L}_{GCE}(\vec{q},i) := \sum_{i=1}^c p_i \frac{1-q_i^a}{a}$. We find the minima by constructing the Langrangian $A(\vec{q},\lambda) := \sum_{i=1}^c p_i \frac{1-q_i^a}{a} + \lambda(\sum_{i=1}^c q_i - 1)$. By taking partials and equating to zero, we obtain $q_i^{1-a} = \frac{ap_i}{\lambda}, \forall i$. Using the fact that $\sum_{i=1}^c q_i = 1$ one may find the value of λ. Specifically, $\lambda = a(\sum_{i=1}^c p_i^{\frac{1}{1-a}})^{1-a}$. Thus overall one has $q_i^* = (\frac{ap_i}{\lambda})^{\frac{1}{1-a}} = \frac{p_i^{\frac{1}{1-a}}}{\sum_{i=1}^c p_i^{\frac{1}{1-a}}}$. Let us repeat this for the SCE loss. The expected loss may be written $L_{SCE}(\vec{q},\vec{p}) := \sum_{i=1}^c p_i \mathcal{L}_{SCE}(\vec{q},i) := \sum_{i=1}^c p_i (A(1-q_i) - log(q_i))$. As before, we construct the relevant Lagrangian and find the stationary points: $B(\vec{q},\lambda) := \sum_{i=1}^c p_i (A(1-q_i) - log(q_i)) + \lambda(\sum_{i=1}^c q_i - 1)$. Taking partials and equating to zero we obtain $p_i(A + \frac{1}{q_i}) = \lambda \implies q_i^* = \frac{p_i}{\lambda - Ap_i}$. Here the value of the normalisation constant λ cannot be found in closed form for high values of c and must be computed numerically. Finally, we consider the forward-corrected CE loss. We assume that the loss is corrected by some invertible stochastic matrix T. $L_{FCorr}(\vec{q},\vec{p}) := \sum_{i=1}^c p_i \mathcal{L}_{FCorr}(\vec{q},i) := \sum_{i=1}^c -p_i log((T\vec{q})_i)$. We remark that since CE is proper that this is minimised on the simplex by $\vec{p} = T\vec{q}^* \iff \vec{q}^* = T^{-1}\vec{p}$. For each loss, the function f obtained is bijective as desired. □

Lemma 4.1 Let $\mathcal{L}$ be an elementwise, strictly $f-$proper loss. For any probability estimator $\vec{q}(x)$, we may derive a bound for the noisy pointwise risk that is tight and obtained when $\vec{q}(x) = \tilde{p}(\tilde{y}|x)$. Specifically, for all $x \in supp(p(x))$ we have;

$$R_{\mathcal{L}}^{\eta}(\vec{q})(x) \geq R_{\mathcal{L}}^{\eta}(f(\tilde{p}(\tilde{y}|x))) \tag{5}$$

Proof. Recollect that since $\mathcal{L}$ is a strictly $f-$proper loss, this implies that there exists a strictly proper loss $\tilde{\mathcal{L}}$ such that, for all $\vec{q}$, $\mathcal{L}(f(\vec{q}),i) = \tilde{\mathcal{L}}(\vec{q},i)$. Now, let x be some arbitrary point in the support of $p(x)$ and let $\vec{q}(x)$ be some probability estimator. The pointwise noisy risk of $\vec{q}$ at x may be written as $R_{\mathcal{L}}^{\eta}(\vec{q})(x) := \sum_{i=1}^c \tilde{p}(\tilde{y}=i|x)\mathcal{L}(\vec{q}(x),i) = \sum_{i=1}^c \tilde{p}(\tilde{y}=i|x)\tilde{\mathcal{L}}(f^{-1}(\vec{q}(x)),i) \geq \sum_{i=1}^c \tilde{p}(\tilde{y}=i|x)\tilde{\mathcal{L}}(\tilde{p}(\tilde{y}|x),i) = \sum_{i=1}^c \tilde{p}(\tilde{y}=i|x)\mathcal{L}(f(\tilde{p}(\tilde{y}|x)),i) =: R_{\mathcal{L}}^{\eta}(f(\tilde{p}(\tilde{y}|x)))$. The inequality follows from the definition of the properness of $\tilde{\mathcal{L}}$ while the following equality follows from the definition of $f-$properness. It remains to show that this is attained uniquely by $\vec{q}(x) = f(\tilde{p}(\tilde{y}|x))$. Since $\tilde{\mathcal{L}}$ is strictly proper, we know that the inequality is only obtained by $f^{-1}(\vec{q}(x)) = \tilde{p}(\tilde{y}=i|x)$. The injectivity of f (as specified in the definition of $f-$proper) means this occurs uniquely at $\vec{q}(x) = f(\tilde{p}(\tilde{y}=i|x))$ as desired. □

Lemma 4.5 Let $p(x,y)$ be a distribution where $p(y|x)$ is deterministic, and let $\tilde{p}(x,\tilde{y})$ be a noisy distribution obtained by applying label noise to $p(x,y)$. Assume that $\mathcal{L}$ is a symmetric (strictly) f-proper loss and let J denote the entropy function of its base loss. For any probability estimator q, we can derive a bound for the noisy risk when the label noise is symmetric that is achieved (uniquely) if $q(x) = \tilde{p}(\tilde{y}|x)$. Specifically, for each x, we have the following lower bound, where $\eta(x)$ denotes the noise rate at x:

$$R_{\mathcal{L}}^{\eta}(q) \geq \mathbb{E}_{x\sim p(x)}[J\Big((1-\eta(x), \frac{\eta(x)}{c-1}, \frac{\eta(x)}{c-1}, ..., \frac{\eta(x)}{c-1})\Big)]$$

Proof. Our proof follows similar lines to Lemma 4.1 with the additional application of Theorem 4.3.

Recollect that since $\mathcal{L}$ is a (strictly) $f-$proper loss, this implies that there exists a (strictly) proper loss $\tilde{\mathcal{L}}$ such that, for all $\vec{q}$, $\mathcal{L}(f(\vec{q}),i) = \tilde{\mathcal{L}}(\vec{q},i)$. Now, let x be some arbitrary point in the support of $p(x)$ and let $\vec{q}(x)$ be some probability estimator.

The pointwise noisy risk of $\vec{q}$ at x may be written:

$$R^{\eta}_{\mathcal{L}}(\vec{q})(x) := \sum_{i=1}^{c} \tilde{p}(\tilde{y}=i|x)\mathcal{L}(\vec{q}(x),i) \tag{6}$$

$$= \sum_{i=1}^{c} \tilde{p}(\tilde{y}=i|x)\tilde{\mathcal{L}}(f^{-1}(\vec{q}(x)),i) \tag{7}$$

$$\geq \sum_{i=1}^{c} \tilde{p}(\tilde{y}=i|x)\tilde{\mathcal{L}}(\tilde{p}(\tilde{y}|x),i) \tag{8}$$

$$= \sum_{i=1}^{c} \tilde{p}(\tilde{y}=i|x)\mathcal{L}(f(\tilde{p}(\tilde{y}|x)),i) =: R^{\eta}_{\mathcal{L}}(f(\tilde{p}(\tilde{y}|x))) \tag{9}$$

The inequality (Eqn 8) follows from the definition of the properness of $\tilde{\mathcal{L}}$ while the following equality (Eqn. 9) follows from the definition of $f-$properness. Note that Equation 9 is the definition of the entropy of $\tilde{p}(\tilde{y}|x)$ with respect to the entropy function J associated with $\tilde{\mathcal{L}}$. Thus for all $x \in \text{supp}(p(x))$, we have $R^{\eta}_{\mathcal{L}}(\vec{q})(x) \geq J(\tilde{p}(\tilde{y}|x))$. As in the proof of Lemma 4.1, note that when $\mathcal{L}$ is (*strictly*) $f-$proper, this is attained (uniquely) by setting $\vec{q}(x) = f(\tilde{p}(\tilde{y}=i|x))$. It remains to show that $J(\tilde{p}(\tilde{y}|x)) = J((1-\eta(x), \frac{\eta(x)}{c-1}, \frac{\eta(x)}{c-1}, ..., \frac{\eta(x)}{c-1}))$ when the noise is symmetric. This follows from the determinism assumption.

Suppose that, for all $x \in \text{supp}(p(x))$ that the label noise is symmetric with rate $\eta(x)$. Let $T(x)$ denote the noising transition matrix at x, that is $T_{ij} := p(\tilde{y} = j | y = i, x)$. By the deterministic assumption, we have some k such that $p(y=k|x) = 1$ and $p(y=i|x) = 0$ otherwise. Thus $\tilde{p}(\tilde{y}|x) = \sum_{y=1}^{c} \tilde{p}(\tilde{y}|y,x)p(y|x) = \tilde{p}(\tilde{y}|y=k,x) = (T_{1k}, T_{2k}, ..., T_{ck})$. The noise rate is defined to be the probability that the label is altered by our noise and may thus be expressed as $\eta(x) = 1 - T_{kk}$. By the definition of symmetric noise (Definition A.5), we have $T_{ik} = \frac{\eta(x)}{1-c}$ for $i \neq k$. Hence $J(\tilde{p}(\tilde{y}|x)) = J((T_{1k}, T_{2k}, ..., T_{ck})) = J((1-\eta(x), \frac{\eta(x)}{c-1}, \frac{\eta(x)}{c-1}, ..., \frac{\eta(x)}{c-1}))$ as desired. □

Corollary 4.6 When the label noise is uniform and symmetric with rate η, the following bound on the risk of any probability estimator is tight and achieved only when $q(x) = \tilde{p}(\tilde{y}\,|\,x)$.

$$R^{\eta}_{\mathcal{L}}(q) \geq J\Big((1-\eta, \frac{\eta}{c-1}, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1})\Big) \tag{10}$$

Proof. Uniform label noise means that for all x, $\eta(x) = \eta$. Thus from Lemma 4.5 we have $R^{\eta}_{\mathcal{L}}(q) \geq \mathbb{E}_{x\sim p(x)}[J\Big((1-\eta(x), \frac{\eta(x)}{c-1}, \frac{\eta(x)}{c-1}, ..., \frac{\eta(x)}{c-1})\Big)] = \mathbb{E}_{x\sim p(x)}[J\Big((1-\eta, \frac{\eta}{c-1}, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1})\Big)] = J\Big((1-\eta, \frac{\eta}{c-1}, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1})\Big)$. This bound is obtained by setting $\vec{q}(x) = \tilde{p}(\tilde{y}|x)$ for all $x \in \text{supp}(p(x))$. This is unique if $\mathcal{L}$ is *strictly* $f-$proper. □

Corollary 4.7 Let $\mathcal{L}$ be a symmetric (strictly) f-proper loss and let $\vec{u}_c(\eta) := (1-\eta, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1})$. We have the following bound on the noisy risk if the label noise is symmetric and uniform, which is obtained (uniquely) by setting $\vec{q}(x) = \tilde{p}(\tilde{y}|x)$ for all $x \in \text{supp}(p(x))$.

$$R^{\eta}_{\mathcal{L}}(q) \geq (1-\eta)\mathcal{L}(f(\vec{u}_c(\eta)),1) + \eta\mathcal{L}(f(\vec{u}_c(\eta)), i \neq 1)$$

Proof. Lemma 4.1 states that, for all $x \in supp(p(x))$ we have, $R^{\eta}_{\mathcal{L}}(\vec{q})(x) \geq R^{\eta}_{\mathcal{L}}(f(\tilde{p}(\tilde{y}|x)))$. Since $p(y|x)$ is deterministic it follows we have some k such that $p(y=k|x) = 1$ and $p(y=i|x) = 0$ otherwise. Since $\mathcal{L}$ is symmetric, we may, without loss of generality, let $k = 1$. Thus $\tilde{p}(\tilde{y}|x) = \sum_{y=1}^{c} \tilde{p}(\tilde{y}|y,x)p(y|x) = \tilde{p}(\tilde{y}|y=k,x) = (1-\eta, \frac{\eta}{c-1}, \frac{\eta}{c-1}, ..., \frac{\eta}{c-1}) =: \vec{u}_c(\eta)$. The second-to-last equality follows from our assumption that our noise is symmetric and uniform. Putting these together we have, for all $x \in supp(p(x))$ we have, $R^{\eta}_{\mathcal{L}}(\vec{q})(x) \geq R^{\eta}_{\mathcal{L}}(f(\vec{u}_c(\eta))) := (1-\eta)\mathcal{L}(f(\vec{u}_c(\eta)),1) + \eta\mathcal{L}(f(\vec{u}_c(\eta)), i \neq 1)$. Our result follows by taking expectations with respect to $p(x)$ on both sides. By Lemma 4.1, equality is obtained by setting $\vec{q}(x) = \tilde{p}(\tilde{y}|x)$ for all $x \in \text{supp}(p(x))$. This is unique if $\mathcal{L}$ is *strictly* $f-$proper.

□

Lemma B.1. *Let $p(x,y)$ be a distribution where $p(y|x)$ is deterministic, and let $\tilde{p}(x,\tilde{y})$ be a noisy distribution obtained by applying label noise to $p(x,y)$. Assume that $\mathcal{L}$ is a symmetric (strictly) f-proper loss and let J denote the entropy function of its base loss. For any probability estimator q, we may lower bound the generalised noisy risk of $\vec{q}$ in terms of a quantity $A(\eta(x),c)$ where $\eta(x)$ denotes the noise rate at x and c the number of classes. This bound is achieved (uniquely) by $q(x)=\tilde{p}(\tilde{y}|x)$. Specifically,*

$$R^{\eta}_{\mathcal{L}}(q) \geq \mathbb{E}_{x\sim p(x)}[A(\eta(x),c)]$$

Where $A(\eta(x),c)$ lies in the following interval:

$$A(\eta(x),c) \in \big[J((1-\eta(x),\eta(x),0,0,...,0)), J\Big((1-\eta(x),\frac{\eta(x)}{c-1},\frac{\eta(x)}{c-1},...,\frac{\eta(x)}{c-1})\Big)\big]$$

Proof. Let $\vec{q}(x)$ be a probability estimator and let x be some point in the support of $p(x)$. We established in the proof of Lemma 4.5 that $R^{\eta}_{\mathcal{L}}(\vec{q})(x) \geq J(\tilde{p}(\tilde{y}|x))$. We have equality (uniquely) when $\vec{q}(x) = \tilde{p}(\tilde{y}|x)$. Let $T(x)$ denote the noising transition matrix at x, that is $T_{ij}(x) := p(\tilde{y} = j|y = i, x)$. By the determinism assumption, we have some k such that $p(y = k|x) = 1$ and $p(y = i|x) = 0$ otherwise. Thus $\tilde{p}(\tilde{y}|x) = \sum_{y=1}^{c} \tilde{p}(\tilde{y}|y, x)p(y|x) = \tilde{p}(\tilde{y}|y = k, x) = (T_{1k}(x), T_{2k}(x), \dots, T_{ck}(x))$. Let $A(\eta(x), c) := J(T_{1k}(x), T_{2k}(x), \dots, T_{ck}(x))$ where $\eta(x) := 1 - T_{kk}$ is the noise rate at x. The symmetry of J means that, without loss of generality, we may let $k = 1$. It remains to show that $A(\eta(x),c) \in \big[J((1-\eta(x),\eta(x),0,0,...,0)), J\Big((1-\eta(x),\frac{\eta(x)}{c-1},\frac{\eta(x)}{c-1},...,\frac{\eta(x)}{c-1})\Big)\big]$.

Upper Limit: We begin by demonstrating that $A(\eta(x),c)$ is upper bounded by $J((1-\eta(x),\frac{\eta(x)}{c-1},\frac{\eta(x)}{c-1},...,\frac{\eta(x)}{c-1}))$. Let $\Delta(\eta(x))$ denote the set of non-negative vectors $(a_1,a_2,...,a_{c-1})$ such that $a_i \leq 1$ and $\sum_{i=1}^{c-1} a_i = \eta(x)$. We wish to show the supremum of $J((1-\eta(x),a_1,a_2,...,a_{c-1}))$ is attained on $\Delta(\eta(x))$ by setting $a_i = \frac{\eta(x)}{c-1}$ for all i. This corresponds to the label noise being symmetric at x. By Theorem 4.3, J is a (strictly) concave function. Moreover, the symmetry assumption implies that J is a symmetric function of its variables. Define the function $g(a_1,a_2,...,a_{c-1}) := J(1-\eta(x),a_1,a_2,...,a_{c-1})$. We wish to show that g attains its maximum on the relevant domain when $a_i = a_j$ for all i,j. We begin by noting that the (strict) concavity of J implies the (strict) concavity of g. To see this consider two arbitrary vectors $\vec{x} = (x_1,x_2,...x_{c-1}), \vec{y} = (y_1,y_2,...y_{c-1})$. Now $g(\lambda\vec{x}+(1-\lambda)\vec{y}) = J(\lambda\vec{x}'+(1-\lambda)\vec{y}')$ where $\vec{x}' := (1-\eta(x),x_1,x_2,...,x_{c-1})$ and $\vec{y}' := (1-\eta(x),y_1,y_2,...,y_{c-1})$. Thus the concavity of J implies $g(\lambda\vec{x}+(1-\lambda)\vec{y}) := J(\lambda\vec{x}'+(1-\lambda)\vec{y}') \geq \lambda J(\vec{x}')+(1-\lambda)J(\vec{y}') = \lambda g(\vec{x})+(1-\lambda)g(\vec{y})$ as desired. Thus g is a symmetric (strictly) concave function of its variables.

Let $\vec{a}^*$ denote a maxima of g on $\Delta(\eta(x))$. Let σ denote the cyclic permutation of the components of $\vec{a}$. That is $\sigma(a_1,a_2,...,a_{c-1}) := (a_{c-1},a_1,a_2,...,a_{c-2})$. By the symmetry of g, we know that if $\vec{a}^*$ is a maxima then so is $\sigma^i(\vec{a}^*)$ for all i. Hence by the (strict) concavity of g, we have:

$$\begin{aligned}
g\big(\frac{\eta(x)}{c-1},\frac{\eta(x)}{c-1},...,\frac{\eta(x)}{c-1}\big) &= g\big(\frac{1}{c-1}(\vec{a}^*+\sigma(\vec{a}^*)+\sigma^2(\vec{a}^*)+...+\sigma^{c-2}(\vec{a}^*))\big) && (11)\\
&\geq \frac{1}{c-1}g\big(\vec{a}^*\big)+\frac{1}{c-1}g\big(\sigma(\vec{a}^*)\big)+...+\frac{1}{c-1}g\big(\sigma^{c-1}(\vec{a}^*)\big) && (12)\\
&= g(\vec{a}^*) && (13)
\end{aligned}$$

Hence g is maximised by setting $a_i = \frac{\eta(x)}{c-1}$ for all i as desired. This is the unique maxima when $\mathcal{L}$ is strictly $f-$proper.

Lower Limit: It now remains to show that the lower bound on $A(\eta(x),c)$ holds. The (strict) concavity means that g attains it minima on the vertices of $\Delta(\eta(x))$ (eg $(\eta(x),0,...,0)$. To see this let $\vec{a}^* = (a_1^*,a_2^*,...,a_{c-1}^*)$ denote a minima of g on $\Delta(\eta(x))$. Then we have,

$$\begin{aligned}
g(a_1^*,a_2^*,...,a_{c-1}^*) &= g(a_1^*\vec{e_1}+a_2^*\vec{e_2}+...+a_{c-1}^*\vec{e}_{c-1}) && (14)\\
&\geq \sum_{i=1}^{c-1}\frac{a_i^*}{\eta(x)}g(\eta(x)\vec{e_i}) && (15)\\
&= g(\eta(x),0,...,0) && (16)\\
&= J((1-\eta(x),\eta(x),0,0,...,0)) && (17)
\end{aligned}$$

$\vec{e}_i$ denotes the coordinate vector with 1 in the ith position and zeros elsewhere. Equation 16 holds by the symmetry of g and since $\sum a_i^* = \eta(x)$. Thus we have shown that g is lower bounded by $J((1-\eta(x),\eta(x),0,0,...,0))$ as desired. Moreover, this infimum is obtained on the vertices of $\Delta(\eta(x))$. □

C. Additional Theory and Discussion

C.1. Sensitivity of Bounds

The Noised-Corrected Budget defined in Definition 4.8 was established on the basis of Lemma 4.5 and Corollary 4.6, which assume noise is symmetric and/or uniform. When we deviate from these noise conditions, we generally find that this budget is too high in that an optimal probability estimator could achieve a (noisy) risk lower than this value without overfitting. Since we use this budget in all noise conditions, it is essential to get an idea of the size of the gap between our budget and the minimum achievable risk. Ideally we want this gap to be small. In this section, we look briefly at this topic, noting that this gap is smaller for GCE and SCE than CE. This implies that the noise-corrected budget is more suitably used with SCE and GCE than with CE when noise deviates from idealised assumptions.

In Lemma 4.5 we derived a lower bound on the pointwise noisy risk of a probability estimator when $\mathcal{L}$ is $f-$proper. This bound holds when the label noise is symmetric. When label noise deviates from this condition, our bounds no longer necessarily hold. More generally, we have the following bound:

Lemma C.1. *Let $p(x,y)$ be a distribution where $p(y|x)$ is deterministic, and let $\tilde{p}(x,\tilde{y})$ be a noisy distribution obtained by applying label noise to $p(x,y)$. Assume that $\mathcal{L}$ is a symmetric (strictly) f-proper loss and let J denote the entropy function of its base loss. For any probability estimator q, we may lower bound the generalised noisy risk of $\vec{q}$ in terms of a quantity $A(\eta(x),c)$ where $\eta(x)$ denotes the noise rate at x and c the number of classes. This bound is achieved (uniquely) by $q(x)=\tilde{p}(\tilde{y}|x)$. Specifically,*

$$R^{\eta}_{\mathcal{L}}(q) \geq \mathbb{E}_{x\sim p(x)}[A(\eta(x),c)]$$

Where $A(\eta(x),c)$ lies in the following interval:

$$A(\eta(x),c) \in \left[J((1-\eta(x),\eta(x),0,0,...,0)), J\left((1-\eta(x),\frac{\eta(x)}{c-1},\frac{\eta(x)}{c-1},...,\frac{\eta(x)}{c-1})\right)\right]$$

The proof is given at the end of Appendix B.

Lemma 4.1 tells us that $A(\eta(x),c)$ attains this upper limit of $J((1-\eta(x),\frac{\eta(x)}{c-1},\frac{\eta(x)}{c-1},...,\frac{\eta(x)}{c-1}))$ only if our label noise is symmetric at x. Conversely, as indicated in our proof (Appendix B), the lower limit is obtained when the label noise flips labels to only one other class.

Corollary C.2. *When the label noise is uniform with rate η, one may lower bound the noisy risk of any probability estimator $\vec{q}$ in terms of a quantity $A(\eta,c)$. Moreover, this bound is tight and achieved only when $q(x)=\tilde{p}(\tilde{y}\,|\,x)$.*

$$R^{\eta}_{\mathcal{L}}(q) \geq A(\eta,c)$$

Where $A(\eta,c)$ lies in the following interval, achieving the upper limit only if the label noise is symmetric for all $x \in \text{supp}(p(x))$:

$$A(\eta,c) \in \left[J((1-\eta,\eta,0,0,...,0)), J\left((1-\eta,\frac{\eta}{c-1},\frac{\eta}{c-1},...,\frac{\eta}{c-1})\right)\right] \tag{18}$$

Proof. This follows immediately from Lemma C.1 when $\eta(x)$ has no dependence on x ($\eta(x)=\eta$). □

Corollary C.2 indicates that when noise is uniform but not symmetric, our Noise-Corrected Budget (Definition 4.8) of $J((1-\eta,\frac{\eta}{c-1},\frac{\eta}{c-1},...,\frac{\eta}{c-1}))$ is too high since the true minimum achievable risk is lower than this budget. In other words, there exists a probability estimator which attains a risk lower than our budget. This non-optimality is the cost we incur as a result of requiring a simple, easily computable budget. Importantly, this corollary gives us a rough way to quantify this non-optimality, using the difference between the upper and lower limits of the interval $\left[J((1-\eta(x),\eta(x),0,0,...,0)), J\left((1-\eta(x),\frac{\eta(x)}{c-1},\frac{\eta(x)}{c-1},...,\frac{\eta(x)}{c-1})\right)\right]$. When this difference is large, one could construct two types of label noise with the same rate η, such that the difference in the minimum achievable risks between these noise types is significant. Conversely, when this gap is small, the minimum

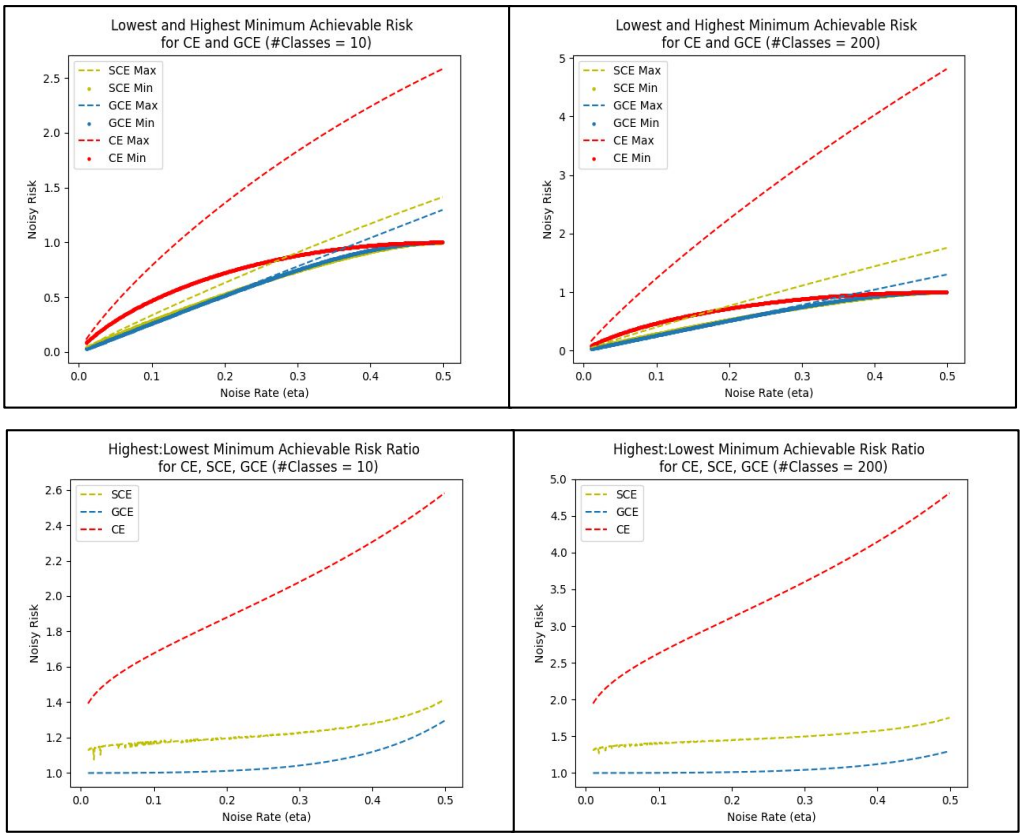

Figure 2. On the top row, we plot the upper and lower limits of $A(\eta,c)$ for $\eta \in (0,0.5]$ from Corollary C.2 for the CE (red), SCE (yellow) and GCE (blue) losses for 10 classes (left) and 200 classes (right). On the bottom row, we plot a ratio of these upper and lower limits instead. We observe that the difference between these upper and lower limits is far greater for CE than the other losses. This is more pronounced for more classes.

achievable risk for any type of label noise at a fixed rate η is similar. This is a desirable property and suggests that simply setting our budget to our noise-corrected budget is probably suitable regardless of the specifics of the noising process.

On the top row of Figure 2, we give a plot of the upper and lower limits of $A(\eta,c)$ (Corollary C.2) for $\eta \in (0,0.5]$ for $c=10$ (left) and $c=200$ (right) for GCE, SCE and CE. The upper limit is given by a dotted line, while the lower limit is given by a filled line in the same colour. Each loss is scaled so they may be more easily compared. Similarly, in the row below, we plot the ratios of the upper and lower limits of $A(\eta,c)$ for each loss. These graphs show that the difference between the upper and lower limits is much greater for CE than for SCE and GCE. This difference is more pronounced when the number of classes is greater. The result is that on non-symmetric noise, our noise-corrected budget (Definition 4.8) will generally be less suitable when used in conjunction with CE than when used with GCE or SCE.

C.2. $f-$Proper Losses: A Discussion

Recall Definition 3.1 below.

Definition 3.1 Let $\mathcal{L}$ be an elementwise loss and $f:\Delta^{c-1}\to\Delta^{c-1}$ be an bijective function. We define $\mathcal{L}$ as (strictly) f**-proper** if there exists a (strictly) proper loss $\mathcal{L}_{\text{proper}}$ such that for all $\vec{q}\in\Delta^{c-1}$, $\mathcal{L}(f(\vec{q}),i)=\mathcal{L}_{\text{proper}}(\vec{q},i)$. We refer to $\mathcal{L}_{\text{proper}}$ as the **base loss**.

It is instructive to consider how broad this family of losses is since the results in Section 4 hold for all losses which satisfy this definition. We mentioned in Section 3 that this definition trivially contains all proper losses by letting $f=id.$. Below we give a sufficient condition for a loss to be $f-$proper.

Proposition C.3. *Let $\mathcal{L}$ be some elementwise loss function. Let $L_{\mathcal{L}}:\Delta^{c-1}\times\Delta^{c-1}\to\mathbb{R}$ denote its expected loss function*

$L_{\mathcal{L}}(\vec{q},\vec{p}) := \sum_{i=1}^{c} p_i \mathcal{L}(\vec{q},i)$. *Define* $g(\vec{p}) := \operatorname{argmin}_{\vec{q}} L_{\mathcal{L}}(\vec{q},\vec{p})$, *if* g *is surjective then* $\mathcal{L}$ *is* $f-$*proper.*

Proof. Let $\mathcal{L}$ be some elementwise loss such that g (as defined above) is surjective for some loss $\mathcal{L}$. By the definition of g we have $L_{\mathcal{L}}(\vec{q},\vec{p}) \geq L_{\mathcal{L}}(g(\vec{p}),\vec{p}) := \sum_{i=1}^{c} p_i \mathcal{L}(g(\vec{p}),i)$. Now, define the elementwise loss $\tilde{\mathcal{L}}(\vec{q},i) := \mathcal{L}(g(\vec{q}),i)$. We claim that $\tilde{\mathcal{L}}$ is proper. Let $\vec{p} \in \Delta^{c-1}$ then $L_{\tilde{\mathcal{L}}}(\vec{q},\vec{p}) := \sum_{i=1}^{c} p_i \tilde{\mathcal{L}}(\vec{q},i) := \sum_{i=1}^{c} p_i \mathcal{L}(g(\vec{q}),i) = L_{\mathcal{L}}(g(\vec{q}),\vec{p})$. This is minimised by setting $g(\vec{q}) = g(\vec{p})$ which occurs at $\vec{p} = \vec{q}$. (If g is injective this occurs uniquely when $\vec{p} = \vec{q}$ although this is not required). Thus it follows that $\tilde{\mathcal{L}}$ is proper. Since g is surjective then one may define an injective inverse function $f := g^{-1}$ on Δ^{c-1}. Thus we have $\mathcal{L}(\vec{q},i) = \tilde{\mathcal{L}}(f(\vec{q}),i)$ with $\tilde{\mathcal{L}}$ proper and f injective as desired. Hence $\mathcal{L}$ is $f-$proper. □

C.3. Noise Model Plots

In Lemma 3.2 we showed that the SCE, GCE and FCE losses are $f-$proper and derived the corresponding functions f. As discussed, these functions can be interpreted as denoising models i.e. $p(y|x) \approx f(\tilde{p}(\tilde{y}|x))$. In Figure 3, we give plots of f for SCE, GCE and FCE. The $x-$axis is the true probability p of an event occurring. On the $y-$axis we plot $f(p)$ against p. For proper losses, one sets $q = p$, which corresponds to no noise model. The graphs for GCE and SCE are remarkably similar. One can interpret their graphs as a noise model where labels which are intrinsically uncertain are more likely to incur label noise than those which are less ambiguous. FCE requires a noise model in order to be fully specified; we assume symmetric label noise at $\eta = 0.4$. Varying η will change the steepness of the respective f. Finally, we plot MAE. This loss function is not $f-$proper; however, it's useful as a reference. We see that the expected loss is minimised by letting $q = 0$ if $p < 0.5$ and $q = 1$ otherwise. The graphs of SCE and GCE lie between those of MAE and CE. By varying the parameters of these losses, we can interpolate between them.

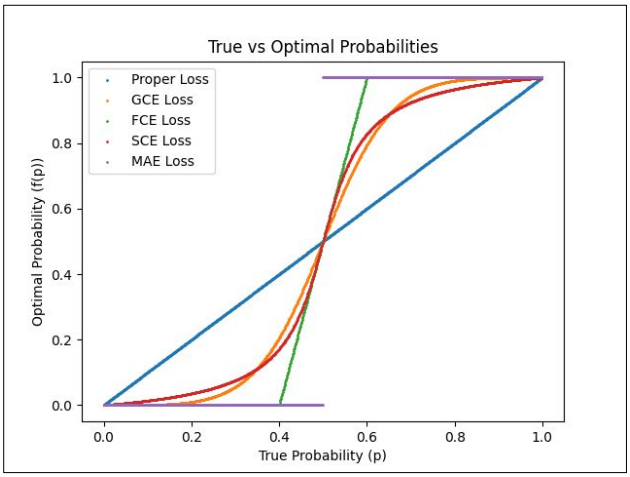

Figure 3. Plot of $f(p)$ for SCE ($A = 8$), GCE ($a = 0.7$), FCE ($\eta = 0.4$), CE and MAE in the binary case. We have the true probability p on the x-axis and the choice of q, which minimises the expected loss on the y-axis.

C.4. Explicit Budgets

From Lemma 3.2 and Corollary 4.7, we can produce the noise-corrected budgets (Definition 4.8) for GCE and SCE. The budget for GCE is given below.

$$B_{GCE}(\eta,c) := \frac{(1-\eta)}{a}\left(1-\left(\frac{(1-\eta)^{\frac{1}{1-a}}}{(1-\eta)^{\frac{1}{1-a}}+(c-1)(\frac{\eta}{c-1})^{\frac{1}{1-a}}}\right)^{a}\right)+$$
$$\frac{\eta}{a}\left(1-\left(\frac{\frac{\eta}{c-1}^{\frac{1}{1-a}}}{(1-\eta)^{\frac{1}{1-a}}+(c-1)(\frac{\eta}{c-1})^{\frac{1}{1-a}}}\right)^{a}\right)$$

The noise-corrected budget for SCE is

$$B_{SCE}(\eta,c) := (1-\eta)\left(-log\left(\frac{1-\eta}{\lambda - A(1-\eta)}\right) + A\left(1-\frac{1-\eta}{\lambda - A(1-\eta)}\right)\right) \tag{19}$$

$$+\eta\left(-log\left(\frac{\eta}{\lambda(c-1)-A\eta}\right) + A\left(1-\frac{\eta}{\lambda(c-1)-A\eta}\right)\right) \tag{20}$$

Recollect that λ is chosen so that the resulting distribution normalises: $\frac{1-\eta}{\lambda - A(1-\eta)} + \frac{\eta(c-1)}{\lambda(c-1)-A\eta} = 1$ and may be computed numerically or by solving the resulting quadratic.

D. Further Experiments

D.1. Experimental Details

The number of training epochs was the same for each loss. For MNIST, FashionMNIST, TinyImageNet and Animals10N, we used 100 epochs; for all other datasets, we used 120 epochs. Each experiment in Tables 1 and 3 was run three times, and the mean and unbiased estimate of the standard deviation is given. We used a ResNet18 architecture for all experiments except TinyImageNet and Animals10N, where a ResNet34 was used. Each experiment is carried out on a single GeForce GTX Titan X. We used a batch size of 300 in all experiments except TinyImageNet and Animals10N, where this is reduced to 200. A learning rate of 0.0001 was used for all losses except MAE (lr = 0.001) and ELR where we used their recommended learning rate of 0.01. We use a learning rate scheduler which scales our learning rate by 0.6 at epoch 60. We use an SGD optimiser with weight decay parameter of 0.01 and momentum 0.5. Our implementation of the *Truncated Loss* comes from the github implementation of GCE (Chou, 2019). We use the official codebase for our implementation of ELR (Liu, 2020). Other losses are re-implementations based on details given in the respective papers. Our SCE loss used the recommended hyperparameter of $A = 8$. Our GCE loss used $a = 0.4$. FCE requires one to define a noise model. In each case, we assume noise is symmetric at the relevant rate. For Animals10N, this rate is set to 11%, which is the estimated noise rate.

Baseline Losses: We compare our results against those obtained by standard robust losses. These baselines include mean-squared error (MSE), mean absolute error (MAE), NCE-MAE (Ma et al., 2020), ELR (Liu et al., 2020), Curriculum loss (CL) (Zhou et al., 2020), Bootstrapping loss (Boot.) (Reed et al., 2014), Spherical loss (Sph.), Mix-up (Zhang et al., 2017), and a version of GCE that incorporates the additional tricks outlined by Zhang & Sabuncu (2018). To differentiate this version of GCE from our simplified GCE, we refer to it as 'Truncated loss' (Trunc.) due to its use of truncation.

The budgets we employ in each experiment are obtained by substituting the relevant number of classes c and the noise rate η into Eqn. 4. An exception is the case of Non-uniform EMNIST, where we use a class number of $c = 2$ to reflect that label is a mixture of the clean label and classifier labels.

Dataset Noise: Many of our experiments employ synthetic symmetric label noise. We construct symmetric label noise by taking the training set and randomly altering the labels of a proportion ($\eta \in \{0.2, 0.4, 0.6\}$) of samples. Labels are switched to a different class with equal probability. For 'Asym-Cifar100,' we introduce asymmetric noise. This is constructed by randomly transitioning labels within the 20 superclasses of CIFAR100. For example, within the superclass 'fish' (comprised of aquarium-fish, flatfish, ray, shark, and trout), we change training labels to other members of the set with a probability of $\eta \in 0.2, 0.4$ (e.g., flatfish $\rightarrow$ trout). For 'Non-Uniform EMNIST,' we investigate the impact of using non-uniform noise. We train a linear classifier on EMNIST and, with a probability of 0.6, modify the label of a data point in our training set to match the output of this classifier. Since the performance of the classifier varies across data-space, this creates noise with an x-dependence.

D.1.1. Cross-Entropy with Prior

One of the losses used in our experiments is cross-entropy with a 'prior' term (CEP). We give an explanation of the motivation for this additional loss term and details of how it's implemented.

In Section 4 we assumed that the un-noised distribution $p(y|x)$ is deterministic for each x (i.e. $p(k|x) = 1, p(i \neq k|x) = 0$). Thus, in the case of symmetric noise with a known noise rate η, the noisy label distribution $\tilde{p}(\tilde{y}|x)$ is of the form for each x:

$$\tilde{p}(\tilde{y}|x) = \left(\frac{\eta}{c-1}, \frac{\eta}{c-1}, \ldots, \underbrace{1-\eta}_{\text{kth position}}, \ldots, \frac{\eta}{c-1}\right) \tag{21}$$

We argue, therefore, that it is reasonable to introduce a term to penalise our model when its outputs deviate from this distribution. This is achieved through a regularisation term which measures the KL-divergence between our model probabilities and the desired distribution (Eqn 21). Let $\vec{p}_\eta := (p_1,p_2,...,p_c) := (1-\eta, \frac{\eta}{c-1},...,\frac{\eta}{c-1})$ and let $q_1,q_2,...,q_c$ denote the probabilities output by our model. We sort the q_i into descending order (which we denote as $q_{\sigma(i)}$) and define our prior term as:

$$\mathcal{L}_{prior}(\vec{q},\vec{p}_\eta) := -\sum_{i=1}^{c} p_i \log(q_{\sigma(i)}) \tag{22}$$

Thus, overall we have $\mathcal{L}_{CEP}(\vec{q},i) := \mathcal{L}_{CE}(\vec{q},i) + \mathcal{L}_{prior}(\vec{q},\vec{p}_\eta)$. Our results (Tables 1,4,3) show that this additional term generally results in additional improvement over using the noise-corrected budget alone. This prior acts as a method of feasible set reduction: There are many different probability estimators which achieve a training error equal to our noise-corrected budget. Therefore, by introducing a prior term (Eqn.22) we can further restrict the set of admissible models.

D.1.2. Optimal Budgets

In our experiment tables in Section 5, we give results using our noise-corrected budgets. We additionally give results where the budget is treated as a hyperparameter. We do not search over the entire space; rather, we do a grid search near the noise-corrected budget. For MNIST, FashionMNIST, EMNIST, CIFAR10 and CIFAR100, we search over $\{-0.2,-0.15,-0.1,...,0.15,0.2\}$ where e.g. 0.2 means that we add 0.2 onto our noise-corrected budget. For Asymmetric CIFAR100 (ACIFAR100) and Non-uniform EMNIST (NU-EMNIST), this range is broadened to $\{-0.6,-0.55,...0.55,0.6\}$. The budgets which give the best results are given in Table D.1.2. When the optimal budget is higher than the noise-corrected budget, this is highlighted in blue. Otherwise, the cell is indicated in red. In our original table, we have columns for Top1 and Top5 accuracy which often have slightly different optimal budgets. For brevity, we combine these by taking a mean of these values.

	MNIST		Fashion		EMNIST		CIFAR10		CIFAR100		ACIFAR100		NU-EMNIST
	0.4	0.6	0.2	0.4	0.2	0.4	0.2	0.4	0.2	0.4	0.2	0.4	0.6
FCE	-0.05	-0.05	0.0	-0.05	0.05	0.05	0.05	0.1	0.03	0.0	-0.1	-0.35	-0.2
GCE	0.0	0.0	0.05	0.0	0.03	0.05	0.05	0.05	0.05	0.0	0.05	0.05	0.0
SCE	0.0	0.05	0.0	0.2	0.2	0.1	0.0	0.2	0.2	0.2	0.0	0.0	0.0
CEB	0.0	0.0	0.0	0.0	0.05	0.05	0.05	0.05	0.1	0.1	0.2	0.0	-0.6
CEP	-0.15	-0.15	-0.15	-0.15	0.0	0.02	0.0	0.0	-0.08	-0.08	-0.08	-0.08	-0.1

Table 2. table giving the offset of the 'optimal' budget from the noise-corrected budget. Here a negative (blue) number means that the budget is greater than the noise-corrected budget. Positive (red) means the optimal budget is lower. Grey means that the optimal budget is zero, i.e no offset.

Losses	MNIST 0.4	MNIST 0.6	FashionMNIST 0.2	FashionMNIST 0.4	EMNIST 0.2 Top 1	EMNIST 0.2 Top 5	EMNIST 0.4 Top 1	EMNIST 0.4 Top 5	CIFAR10 0.2	CIFAR10 0.4
MSE	$93.3_{\pm 0.47}$	$85.8_{\pm 0.95}$	$84.8_{\pm 0.22}$	$80.6_{\pm 0.84}$	$82.9_{\pm 0.29}$	$98.1_{\pm 0.04}$	$80.2_{\pm 0.19}$	$97.1_{\pm 0.07}$	$78.7_{\pm 1.51}$	$56.4_{\pm 0.11}$
MAE	$97.9_{\pm 0.08}$	$96.4_{\pm 0.08}$	$83.2_{\pm 0.10}$	$82.2_{\pm 0.37}$	$49.8_{\pm 2.83}$	$52.2_{\pm 0.10}$	$50.4_{\pm 1.14}$	$51.4_{\pm 0.96}$	$88.6_{\pm 1.34}$	$78.9_{\pm 5.95}$
NCE	$97.8_{\pm 0.06}$	$96.0_{\pm 0.25}$	$87.7_{\pm 0.26}$	$86.3_{\pm 0.14}$	$84.5_{\pm 0.25}$	$97.9_{\pm 0.05}$	$82.6_{\pm 0.81}$	$96.7_{\pm 0.03}$	$\mathbf{89.3}_{\pm 0.40}$	$\mathbf{86.0}_{\pm 0.81}$
MixUp	$95.8_{\pm 1.24}$	$86.8_{\pm 0.85}$	$86.9_{\pm 0.10}$	$82.3_{\pm 0.54}$	$84.3_{\pm 0.08}$	$98.1_{\pm 0.04}$	$81.6_{\pm 0.48}$	$97.1_{\pm 0.08}$	$86.0_{\pm 0.46}$	$77.9_{\pm 0.49}$
Sph.	$95.0_{\pm 0.41}$	$88.1_{\pm 0.82}$	$87.2_{\pm 0.04}$	$84.1_{\pm 0.75}$	$84.6_{\pm 0.12}$	$98.3_{\pm 0.05}$	$83.2_{\pm 0.29}$	$98.1_{\pm 0.58}$	$86.6_{\pm 0.01}$	$72.1_{\pm 0.80}$
Boot.	$86.6_{\pm 0.56}$	$71.2_{\pm 1.17}$	$82.0_{\pm 0.61}$	$73.4_{\pm 1.06}$	$80.5_{\pm 0.24}$	$96.7_{\pm 0.06}$	$77.3_{\pm 0.98}$	$95.0_{\pm 0.25}$	$77.0_{\pm 1.57}$	$58.2_{\pm 2.99}$
Trunc.	$97.1_{\pm 0.12}$	$94.2_{\pm 0.39}$	$87.8_{\pm 0.29}$	$85.3_{\pm 0.77}$	$84.1_{\pm 0.53}$	$97.4_{\pm 1.03}$	$83.1_{\pm 0.55}$	$97.2_{\pm 1.00}$	$88.3_{\pm 0.56}$	$84.2_{\pm 0.69}$
CL	$82.7_{\pm 0.57}$	$67.5_{\pm 1.83}$	$81.2_{\pm 0.34}$	$73.1_{\pm 0.66}$	$79.6_{\pm 0.17}$	$96.4_{\pm 0.05}$	$75.1_{\pm 0.67}$	$94.2_{\pm 0.24}$	$76.0_{\pm 2.16}$	$59.4_{\pm 4.20}$
ELR	$98.1_{\pm 0.04}$	$97.8_{\pm 0.07}$	$85.3_{\pm 0.23}$	$83.4_{\pm 0.02}$	$81.8_{\pm 0.26}$	$97.5_{\pm 0.21}$	$76.6_{\pm 0.10}$	$96.5_{\pm 0.11}$	$88.1_{\pm 0.82}$	$85.7_{\pm 0.06}$
FCE.	$95.4_{\pm 0.25}$	$92.3_{\pm 0.13}$	$83.6_{\pm 0.11}$	$79.9_{\pm 0.78}$	$83.1_{\pm 0.12}$	$98.4_{\pm 0.20}$	$80.6_{\pm 0.12}$	$98.0_{\pm 0.03}$	$84.7_{\pm 0.40}$	$75.1_{\pm 0.04}$
FCE+B	$95.7_{\pm 0.18}$	$92.7_{\pm 0.74}$	$84.8_{\pm 0.26}$	$81.7_{\pm 0.27}$	$83.4_{\pm 0.09}$	$98.5_{\pm 0.03}$	$81.6_{\pm 0.51}$	$98.1_{\pm 0.15}$	$86.7_{\pm 0.21}$	$82.2_{\pm 0.06}$
FCE+B*	$96.7_{\pm 0.17}$	$94.3_{\pm 0.50}$	$84.8_{\pm 0.26}$	$83.3_{\pm 0.22}$	$84.4_{\pm 0.06}$	$\mathbf{98.6}_{\pm 0.13}$	$83.1_{\pm 0.42}$	$98.1_{\pm 0.10}$	$87.2_{\pm 0.20}$	$82.2_{\pm 0.06}$
GCE	$94.4_{\pm 0.36}$	$83.8_{\pm 1.14}$	$86.4_{\pm 0.24}$	$81.6_{\pm 0.37}$	$84.3_{\pm 0.13}$	$98.4_{\pm 0.08}$	$82.7_{\pm 0.07}$	$97.9_{\pm 0.02}$	$81.1_{\pm 0.72}$	$60.0_{\pm 1.31}$
GCE+B	$96.6_{\pm 0.22}$	$94.0_{\pm 0.13}$	$86.5_{\pm 0.56}$	$85.5_{\pm 0.13}$	$84.1_{\pm 0.29}$	$98.4_{\pm 0.04}$	$82.8_{\pm 0.28}$	$98.0_{\pm 0.06}$	$86.1_{\pm 0.22}$	$79.0_{\pm 1.17}$
GCE+B*	$96.6_{\pm 0.22}$	$94.0_{\pm 0.13}$	$87.0_{\pm 0.04}$	$85.5_{\pm 0.13}$	$84.3_{\pm 0.09}$	$98.4_{\pm 0.06}$	$83.6_{\pm 0.25}$	$\mathbf{98.2}_{\pm 0.03}$	$86.7_{\pm 0.07}$	$80.2_{\pm 0.83}$
SCE	$89.5_{\pm 5.29}$	$70.2_{\pm 0.69}$	$82.7_{\pm 0.64}$	$74.4_{\pm 0.37}$	$82.1_{\pm 0.33}$	$96.8_{\pm 0.10}$	$79.6_{\pm 0.61}$	$95.4_{\pm 0.15}$	$78.2_{\pm 0.42}$	$59.0_{\pm 4.43}$
SCE+B	$97.0_{\pm 0.16}$	$93.4_{\pm 0.29}$	$87.5_{\pm 0.22}$	$85.2_{\pm 0.98}$	$83.5_{\pm 0.29}$	$97.3_{\pm 0.14}$	$81.8_{\pm 0.52}$	$96.4_{\pm 0.20}$	$88.9_{\pm 0.44}$	$84.7_{\pm 0.37}$
SCE+B*	$97.0_{\pm 0.16}$	$93.7_{\pm 0.52}$	$87.5_{\pm 0.22}$	$85.8_{\pm 0.67}$	$83.6_{\pm 0.03}$	$97.4_{\pm 0.02}$	$81.8_{\pm 0.52}$	$96.5_{\pm 0.26}$	$88.9_{\pm 0.44}$	$84.9_{\pm 0.20}$
CE	$80.8_{\pm 2.31}$	$67.3_{\pm 0.80}$	$80.9_{\pm 1.11}$	$72.1_{\pm 2.16}$	$79.9_{\pm 0.28}$	$96.4_{\pm 0.08}$	$75.6_{\pm 0.20}$	$94.2_{\pm 0.24}$	$76.9_{\pm 1.22}$	$59.9_{\pm 2.15}$
CE+B	$96.2_{\pm 0.32}$	$93.0_{\pm 0.09}$	$87.9_{\pm 0.10}$	$84.7_{\pm 0.37}$	$80.8_{\pm 0.08}$	$97.0_{\pm 0.04}$	$78.9_{\pm 0.12}$	$96.1_{\pm 0.26}$	$84.5_{\pm 0.73}$	$76.0_{\pm 1.13}$
CE+B*	$96.2_{\pm 0.32}$	$93.0_{\pm 0.09}$	$87.9_{\pm 0.10}$	$84.7_{\pm 0.37}$	$81.5_{\pm 0.11}$	$97.3_{\pm 0.02}$	$79.0_{\pm 0.09}$	$96.2_{\pm 0.01}$	$84.8_{\pm 0.55}$	$78.6_{\pm 1.28}$
CEP	$97.5_{\pm 0.08}$	$92.1_{\pm 0.44}$	$87.8_{\pm 0.12}$	$84.8_{\pm 0.23}$	$85.5_{\pm 0.10}$	$98.1_{\pm 0.07}$	$84.3_{\pm 0.22}$	$97.6_{\pm 0.14}$	$84.2_{\pm 0.51}$	$58.2_{\pm 2.94}$
CEP+B	$95.6_{\pm 0.32}$	$85.5_{\pm 0.77}$	$88.1_{\pm 0.31}$	$84.2_{\pm 0.33}$	$\mathbf{85.8}_{\pm 0.12}$	$98.3_{\pm 0.02}$	$\mathbf{84.8}_{\pm 0.10}$	$98.0_{\pm 0.04}$	$88.5_{\pm 0.32}$	$85.1_{\pm 0.20}$
CEP+B*	$\mathbf{98.5}_{\pm 0.05}$	$\mathbf{97.9}_{\pm 0.11}$	$\mathbf{88.4}_{\pm 0.04}$	$\mathbf{87.2}_{\pm 0.21}$	$85.8_{\pm 0.12}$	$98.3_{\pm 0.02}$	$84.8_{\pm 0.10}$	$98.0_{\pm 0.16}$	$88.5_{\pm 0.32}$	$85.1_{\pm 0.20}$

Table 3. Test accuracies obtained by using different losses on the noisy MNIST/ FashionMNIST/EMNIST/CIFAR10 datasets. Losses implementing the Noise-Corrected Budget shaded in grey. When using this budget provides benefit, the corresponding value is boxed. Overall top values in yellow.

Losses	TinyImageNet (0.2) Top 1	TinyImageNet (0.2) Top 5	TinyImageNet (0.4) Top 1	TinyImageNet (0.4) Top 5	Animals
L2 (MSE)	42.91	67.02	29.42	53.13	80.97
MAE	3.86	5.58	3.94	5.54	54.67
NCE-MAE	7.63	10.24	6.29	10.70	80.85
Mix-Up	47.13	70.08	31.05	58.96	83.76
Bootstrap	40.04	61.94	25.69	46.65	82.11
Truncated	43.35	63.67	38.14	59.99	81.69
Mix-Up	47.13	70.08	31.05	58.96	83.10
Curriculum	41.81	64.53	27.57	48.84	81.68
ELR	44.95	66.65	34.66	55.72	82.62
FCE	43.81	64.97	48.85	29.92	81.82
FCE+B	51.18	73.79	46.34	69.92	82.40
GCE	39.81	60.51	26.93	45.17	81.13
GCE+B	47.40	71.37	39.13	63.75	81.37
SCE	39.81	60.51	26.93	45.17	82.59
SCE+B	41.02	63.06	32.02	52.44	81.25
CE	39.34	61.82	25.84	46.08	81.45
CE+B	38.47	61.85	30.00	52.61	80.72
CEP	44.39	64.56	33.33	51.45	82.06
CEP+B	47.85	71.00	40.56	65.15	81.79

Table 4. Test accuracies obtained by using different losses on the noisy TinyImageNet and Animals10N datasets. Losses implementing the Noise-Corrected Budget are shaded in grey. When using this budget provides benefit, the corresponding value is boxed. Overall top values are in yellow.