Lower-Bounded Proper Losses for Weakly Supervised Classification

Shuhei M. Yoshida [1 2] **Takashi Takenouchi** [2 3] **Masashi Sugiyama** [2 4]

Abstract

This paper discusses the problem of weakly supervised classification, in which instances are given weak labels that are produced by some label-corruption process. The goal is to derive conditions under which loss functions for weak-label learning are proper and lower-bounded—two essential requirements for the losses used in class-probability estimation. To this end, we derive a representation theorem for proper losses in supervised learning, which dualizes the Savage representation. We use this theorem to characterize proper weak-label losses and find a condition for them to be lower-bounded. From these theoretical findings, we derive a novel regularization scheme called generalized logit squeezing, which makes any proper weak-label loss bounded from below, without losing properness. Furthermore, we experimentally demonstrate the effectiveness of our proposed approach, as compared to improper or unbounded losses. The results highlight the importance of properness and lower-boundedness.

1. Introduction

Recent machine learning techniques have achieved state-of-the-art performance on many prediction tasks, but they usually require massive training data with clean annotations. One approach to reduce the costs of data preparation is so-called weakly supervised learning: each instance is annotated with a weak label that is cheaper to obtain but less informative than a true label. For classification, many types of weak supervision have been proposed. For example, in learning with noisy labels (Angluin & Laird, 1988; Natarajan et al., 2013; Patrini et al., 2017), one observes an instance with a label that may be corrupted. Positive-unlabeled (PU) learning of binary classification uses positive and unlabeled data, but not labeled negative data (Elkan & Noto, 2008; du Plessis et al., 2015). Another example is learning from partial labels, which are collections of candidate labels among which only one is true (Cour et al., 2011). Many of these approaches are understood as learning from weak labels that are produced by label-corruption processes, and some authors have taken unified approaches to tackle these problems (van Rooyen & Williamson, 2018; Zhang et al., 2019).

A fundamental theoretical question is under what conditions learning from weak labels is possible. To address this question, analysis of loss functions plays a central role. Among loss functions, proper losses are a particularly important class of losses that can correctly estimate class posterior probabilities (Winkler & Murphy, 1968; Buja et al., 2005; Gneiting & Raftery, 2007). Two major classes of proper weak-label losses have been proposed in the literature. One class derives from unbiased risk estimation, or backward loss correction (Patrini et al., 2017), in which a label-corruption process is inverted to estimate an expected risk with respect to the distribution of true labels. This approach has been taken, for example, in partial-label learning (Cid-sueiro, 2012), noisy-label learning (Natarajan et al., 2013; Patrini et al., 2017), PU learning (du Plessis et al., 2015), and complementary-label learning (Ishida et al., 2019). For a general label-corruption process, a recent work showed how to construct a proper weak-label loss from a loss for supervised learning (van Rooyen & Williamson, 2018). The other class of losses follows from forward loss correction (Patrini et al., 2017), in which proper loss functions are used for estimating the posterior distribution of weak labels. This approach has been applied to noisy-label learning (Patrini et al., 2017) and complementary-label learning (Yu et al., 2018). Moreover, Zhang et al. (2019) applied a forward-corrected loss to more general problems of learning from weak labels, although their discussion focused on the negative log-likelihood loss.

In addition to properness, lower-boundedness is another important requirement for loss functions so that learning can succeed. Losses that are not bounded from below are problematic, as they cause the objective to diverge to negative infinity, especially when using complex models like deep

[1]Biometrics Research Laboratories, NEC Corporation, Kawasaki, Kanagawa, Japan [2]RIKEN Center for Advanced Intelligence Project, Chuo-ku, Tokyo, Japan [3]National Graduate Institute for Policy Studies, Minato-ku, Tokyo, Japan [4]Department of Complexity Science and Engineering, The University of Tokyo, Kashiwa, Chiba, Japan. Correspondence to: Shuhei M. Yoshida <s_m_yoshida@nec.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

neural networks (Kiryo et al., 2017). Forward-corrected losses are known to be proper and lower-bounded (Patrini et al., 2017; Yu et al., 2018). On the other hand, backward-corrected losses are generally not guaranteed to be bounded from below (Natarajan et al., 2013; Cid-Sueiro et al., 2014; du Plessis et al., 2015; Kiryo et al., 2017; Patrini et al., 2017; Ishida et al., 2019; van Rooyen & Williamson, 2018). From a practical viewpoint, implementation tricks proposed by Kiryo et al. (2017) cause a training objective to be positive and work reasonably well, but they also result in an improper loss. Those tricks have also been applied to complementary-label learning (Ishida et al., 2019) and unlabeled-unlabeled learning (Lu et al., 2020). Chou et al. (2020) proposed a novel class of surrogate losses that are bounded from below, but these losses are not guaranteed to be proper. To the best of our knowledge, conditions under which proper weak-label losses are bounded from below have yet to be addressed.

Our contributions This paper discusses proper losses for weakly supervised learning of class posterior probability estimation. In particular, we obtain conditions under which proper weak-label losses are bounded from below. To do so, we derive the dual representation of proper losses for supervised learning. This representation is a dualized version of the Savage representation (Savage, 1971; Cid-Sueiro et al., 1999; Gneiting & Raftery, 2007), which characterizes a proper loss in terms of a Bayes risk. By using a theorem that we obtain, we characterize proper weak-label losses and derive a sufficient condition under which the resulting losses are bounded from below. The derived condition is not necessary but covers a large class of losses that are parameterized by convex functions constrained by a single inequality.

From these results, we derive a novel regularization scheme called generalized logit squeezing (gLS), which makes any proper weak-label loss bounded from below, without losing its properness. We also experimentally demonstrate the effectiveness of our proposed approach as compared to unbounded or improper losses. We show that gLS yields superior or competitive results as compared to baseline methods, regardless of the precise values of the hyperparameters that are specific to gLS, as long as those parameters are in the regime in which gLS gives T-proper and bounded losses.

2. Formulation

In this section, we introduce notations and basic notions, which we adopted from previous studies (Winkler & Murphy, 1968; Buja et al., 2005; Gneiting & Raftery, 2007; Cid-sueiro, 2012; van Rooyen & Williamson, 2018). We begin by summarizing the mathematical notations in Section 2.1. Then, the two key notions of weak labels and proper losses are described in Sections 2.2 and 2.3, respectively.

2.1. Notations

Boldface and calligraphic letters respectively denote vectors and sets. The sets of real numbers and extended real numbers are denoted by $\mathbb{R}$ and $\overline{\mathbb{R}} \equiv \mathbb{R} \cup \{-\infty, \infty\}$, respectively. Let $\mathcal{X}$ be a discrete set and $|\mathcal{X}|$ be its cardinality. The set $\mathbb{R}^{\mathcal{X}}$ is the $|\mathcal{X}|$-dimensional vector space whose dimensions are indexed with $x \in \mathcal{X}$. A matrix $I_{\mathcal{X}}$ is the identity matrix on $\mathbb{R}^{\mathcal{X}}$, $\mathbf{1}_{\mathcal{X}}$ is a vector in $\mathbb{R}^{\mathcal{X}}$ such that $(\mathbf{1}_{\mathcal{X}})_x = 1$ for all $x \in \mathcal{X}$, and $\mathbf{1}_{\mathcal{X}}^{\perp}$ is the orthogonal complement of $\mathbf{1}_{\mathcal{X}}$. The set of probability distributions over $\mathcal{X}$ is identified with the probability simplex $\mathcal{P}(\mathcal{X}) \equiv \{\boldsymbol{p} \in \mathbb{R}^{\mathcal{X}} | \sum_{x\in\mathcal{X}} p_x = 1, p_x \geq 0 \text{ for all } x \in \mathcal{X}\}$.

The theory of convex functions has offered useful tools for analyzing proper losses (Gneiting & Raftery, 2007; Dawid, 2007). A function $f : \mathcal{C} \to \mathbb{R}$ is convex if $f((1-\lambda)\boldsymbol{x}_0 + \lambda\boldsymbol{x}_1) \leq (1-\lambda)f(\boldsymbol{x}_0) + \lambda f(\boldsymbol{x}_1)$ for all $\lambda \in (0, 1)$ and $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathcal{C}$. It is strictly convex if the equality holds only when $\boldsymbol{x}_0 = \boldsymbol{x}_1$. A convex function f is said to be closed if its epigraph $\{(\boldsymbol{x}, t) \in \mathcal{C} \times \mathbb{R} | t \geq f(\boldsymbol{x})\}$ is a closed set. A vector $\underline{\nabla} f(\boldsymbol{x})$ is a subgradient of f at a point $\boldsymbol{x} \in \mathcal{C}$ if it satisfies $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \underline{\nabla} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle$ for all $\boldsymbol{y} \in \mathcal{C}$. In general, subgradients may not be unique at a given point. The set of all the subgradients of f at $\boldsymbol{x} \in \mathcal{C}$ is called the subdifferential and is denoted by $\partial f(\boldsymbol{x})$. The convex conjugate of a convex function $f : \mathcal{C} \to \mathbb{R}$ is denoted by f^* and is defined as $f^*(\boldsymbol{v}) = \sup_{x\in\mathcal{C}} [\langle \boldsymbol{v}, \boldsymbol{x}\rangle - f(\boldsymbol{x})]$.

2.2. Weak Labels in Classification Learning

Let $\mathcal{X}$ be a space of instances, $\mathcal{Z} = \{z_1, z_2, \ldots, z_c\}$ be a set of true (latent) labels, and $\mathcal{Y} = \{y_1, y_2, \ldots, y_{c_W}\}$ be a set of weak (observed) labels. In weakly supervised learning of classification, an algorithm is given a training set sampled from $\mathcal{X} \times \mathcal{Y}$ in accordance with an unknown data distribution, and it learns to classify an instance $x \in \mathcal{X}$ into a true class $z \in \mathcal{Z}$. The true labels for training instances are not available to the learner.

We focus on a setting in which weak labels are characterized by a conditional distribution $p(y|x, z)$, or a label transition matrix $T(x)$, whose matrix element $T_{yz}(x)$ is $p(y|x, z)$. In this paper, we assume that (a) $T(x) \equiv T$, which means that weak labels are independent of input data x, and that (b) T has a left inverse R such that $RT = I_{\mathcal{Z}}$. In principle, Assumption (a) can be lifted by replacing T with $T(x)$ and R with $R(x)$ in the following analysis, even though such scenarios are more challenging to deal with in practice, because they require knowing $T(x)$ for all $x \in \mathcal{X}$. Assumption (b) requires that weak labels be informative enough for a learner to infer a distribution over the true labels. Concretely, we can reconstruct true-label posterior probabilities from weak-label posterior probabilities by using the following identity:

$$p(z|x) = \sum_{z' \in \mathcal{Z}} (RT)_{zz'} p(z'|x) = \sum_{y \in \mathcal{Y}} R_{zy} p(y|x). \quad (1)$$

A label transition matrix T satisfying Assumption (b) is said to be reconstructible, and R is called a reconstruction matrix of T. In particular, T is reconstructible only if $|\mathcal{Z}| \leq |\mathcal{Y}|$.

Solving weakly supervised classification always requires some assumption like Assumption (b) that constrains the form of T. See Appendix A for a comparison of Assumption (b) with other assumptions that have been made in previous works.

The following are illustrative examples with $\mathcal{Z} = \{z_1, z_2, z_3\}$.

Example 1 (Learning with label noise, Natarajan et al. (2013))**.** If instances are equipped with noisy labels, then the weak-label set $\mathcal{Y}$ is identical to $\mathcal{Z}$. For a three-class setting with symmetric noise, T is

$$T = \begin{pmatrix} 1-p & p/2 & p/2 \\ p/2 & 1-p & p/2 \\ p/2 & p/2 & 1-p \end{pmatrix}, \quad (2)$$

and its reconstruction matrix is

$$R = \frac{1}{2-3p} \begin{pmatrix} 2-p & -p & -p \\ -p & 2-p & -p \\ -p & -p & 2-p \end{pmatrix}, \quad (3)$$

where $p \in (0, 1)$ is the mislabeled probability. Note that T is not reconstructible if $p = \frac{2}{3}$, in which case the weak labels become independent of the true labels.

Example 2 (Partial labels, Cour et al. (2011))**.** Consider three-class classification with $\mathcal{Y} = \{(1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}$. A label $y \in \mathcal{Y}$ is called a partial label. For example, $(1,1,0)$ indicates that the true label is either z_1 or z_2, but not z_3. In a scenario in which a spurious label is added with probability p, the label transition matrix T is

$$T = \begin{pmatrix} T_1^\mathsf{T} & T_2^\mathsf{T} & T_3^\mathsf{T} \end{pmatrix}^\mathsf{T}, \quad (4)$$

where

$$T_1 = (1-p)^2 I_3, \quad (5)$$

$$T_2 = \begin{pmatrix} (1-p)p & (1-p)p & 0 \\ (1-p)p & 0 & (1-p)p \\ 0 & (1-p)p & (1-p)p \end{pmatrix}, \quad (6)$$

$$T_3 = \begin{pmatrix} p^2 & p^2 & p^2 \end{pmatrix}. \quad (7)$$

This T is left-invertible unless $p = 1$. The left-inverse is not unique.

So far, we have assumed that there is only one weak-label set $\mathcal{Y}$ and a label transition matrix T, and the arguments in the rest of this paper are made for such a scenario. Note, however, that the arguments here can also be applied to scenarios in which two or more data sources with different noise characteristics are available. Importantly, this can be done without changing any formal aspect of our theory. See Appendix B for the details of this point.

2.3. Proper Losses for Weak-Label Learning

A common strategy for classification is to estimate the class posterior probabilities. To this end, an expected loss should preferably be minimized when an estimator gives the true posterior probabilities:

$$\mathbb{E}_{(x,z)\sim p(x,z)}[l(q(z|x), z)] \geq \mathbb{E}_{(x,z)\sim p(x,z)}[l(p(z|x), z)], \quad (8)$$

where $p(x, z) \in \mathcal{P}(\mathcal{X} \times \mathcal{Z})$ is a sample distribution, $q(z|x) \in \mathcal{P}(\mathcal{Z})$ denotes the estimated posterior probabilities for a given instance $x \in \mathcal{X}$, and $l : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \overline{\mathbb{R}}$ is a loss function. Because the inequalities at different points in $\mathcal{X}$ are mutually independent, we focus on the conditional risk at a fixed x, omit the conditioning variable x, and simply use a vector notation like $\boldsymbol{p} \in \mathcal{P}(\mathcal{Z})$ for the class posterior probabilities in the rest of the paper. Loss functions satisfying Eq. (8) are said to be proper (Winkler & Murphy, 1968). A loss function is said to be strictly proper when the equality in Eq. (8) holds only if $\boldsymbol{p} = \boldsymbol{q}$ (Gneiting & Raftery, 2007). Strict properness is often more desirable than properness itself, because it leads to a Fisher-consistent estimator $\arg\min_{\boldsymbol{q}} \mathbb{E}_{z\sim\boldsymbol{p}}[l(\boldsymbol{q}, z)]$ for the class posterior probabilities. It also guarantees that the minima of the empirical and expected losses are unique, which thereby renders the loss minimization problem well-posed.

In weak-label learning, we use a loss function defined on a pair of predicted posterior probabilities $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$ and a weak label $y \in \mathcal{Y}$; we refer to this function as a weak-label loss. The notion of properness can be extended to weak-label losses (Cid-sueiro, 2012).

Definition 3. Let T be a label transition matrix. A weak-label loss $l_{\mathrm{W}} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ is called T-proper if, for all $\boldsymbol{p}$ and $\boldsymbol{q}$ in $\mathcal{P}(\mathcal{Z})$,

$$\mathbb{E}_{y\sim T\boldsymbol{p}}\left[l_{\mathrm{W}}(\boldsymbol{q}, y)\right] \geq \mathbb{E}_{y\sim T\boldsymbol{p}}\left[l_{\mathrm{W}}(\boldsymbol{p}, y)\right], \quad (9)$$

where the vector $T\boldsymbol{p}$ is a point in the probability simplex $\mathcal{P}(\mathcal{Y})$ and represents a probability distribution over $\mathcal{Y}$. The weak-label loss is said to be strictly proper when the equality in Eq. (9) holds only if $\boldsymbol{p} = \boldsymbol{q}$.

3. Dual Representation of Proper Losses

In this section, we derive a representation of proper loss functions for supervised learning, which we call a dual representation. It is closely related to the so-called Savage representation (Savage, 1971; Gneiting & Raftery, 2007). The Savage representation expresses a proper loss in terms of its Bayes risk, whereas our representation uses a convex function that is related to the convex conjugate of the Bayes risk. The dual representation will be useful for our later discussion of the lower-unboundedness of proper weak-label losses.

We start by reviewing the Savage representation, which requires a mild regularity condition (Gneiting & Raftery, 2007). In general, losses can be positive infinity for some $(\boldsymbol{q}, z) \in \mathcal{P}(\mathcal{Z}) \times \mathcal{Z}$. A loss function is said to be regular if it is finite for any $(\boldsymbol{q}, z) \in \mathcal{P}(\mathcal{Z}) \times \mathcal{Z}$ except possibly that $l(\boldsymbol{q}, z) = \infty$ when $q_z = 0$. Regular proper losses for class posterior probability estimation are known to have the following representation (Cid-Sueiro et al., 1999; Gneiting & Raftery, 2007).

Theorem 4 (Savage representation)**.** *A regular loss function $l : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \overline{\mathbb{R}}$ is (strictly) proper if and only if there exists a closed (strictly) convex function $S : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ such that for $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$ and $z \in \mathcal{Z}$,*

$$l(\boldsymbol{q}, z) = -\left[\underline{\nabla} S(\boldsymbol{q})\right]_z + \langle \boldsymbol{q}, \underline{\nabla} S(\boldsymbol{q})\rangle - S(\boldsymbol{q}), \tag{10}$$

where $\underline{\nabla} S(\boldsymbol{q}) \in \overline{\mathbb{R}}^{\mathcal{Z}}$ is a subgradient of S at a point $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$.

By using the definition of the subgradient, we can easily verify that the convex function S in the theorem is the negative Bayes risk; that is,

$$S(\boldsymbol{p}) = -\min_{\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})} \mathbb{E}_{z \sim \boldsymbol{p}}\left[l(\boldsymbol{q}, z)\right] \equiv -\underline{L}(\boldsymbol{p}), \tag{11}$$

where $\underline{L}(\boldsymbol{p})$ is the Bayes risk. Thus, Theorem 4 shows that a proper loss function is determined by its Bayes risk, up to the choice of $\underline{\nabla} S(\boldsymbol{q}) \in \partial S(\boldsymbol{q})$ at points where S is not differentiable (Williamson et al., 2016).

Importantly, the sum of the second and third terms in Eq. (10) is the convex conjugate $S^*(\underline{\nabla} S(\boldsymbol{q}))$ of S (Reid et al., 2015). This fact leads to the "dual" of the Savage representation. For a closed convex function F whose domain is a convex subset $\mathcal{C}$ of $\mathbf{1}_{\mathcal{Z}}^{\perp}$, we define a function $\lambda_F : \mathcal{C} \times \mathcal{Z} \to \mathbb{R}$ as

$$\lambda_F(\boldsymbol{v}, z) = -v_z + F(\boldsymbol{v}). \tag{12}$$

The following theorem shows that under a certain condition on F, λ_F is essentially a proper loss for which $\mathcal{C}$ parameterizes the probability simplex $\mathcal{P}(\mathcal{Z})$.

Theorem 5. *Let $l : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \overline{\mathbb{R}}$ be a regular loss. Then, it is proper if and only if there exists a closed convex function $F : \mathcal{C} \subset \mathbf{1}_{\mathcal{Z}}^{\perp} \to \mathbb{R}$ that satisfies the following conditions:*

1. *$F(\boldsymbol{v}) - \max_{z \in \mathcal{Z}} v_z$ is bounded from below.*
2. *With $F^*(\boldsymbol{p})$ the convex conjugate of $F(\boldsymbol{v})$, it holds that $l(\boldsymbol{q}, z) = \lambda_F(\underline{\nabla} F^*(\boldsymbol{q}), z)$, where $\underline{\nabla} F^*(\boldsymbol{p})$ is an appropriately chosen subgradient function.*

Furthermore, $F^*(\boldsymbol{p})$ at a point $\boldsymbol{p} \in \mathcal{P}(\mathcal{Z})$ is a negative Bayes risk for this loss.

A full proof of this theorem is presented in Appendix C.1. In Appendix D, we also derive conditions on F under which the associated proper loss is strictly proper; however, we do not use them in the following discussion. Theorem 5 elucidates that F in the proved representation is closely related to the convex conjugate of the negative Bayes risk $-\underline{L}$. Therefore, in the rest of the paper, the representation of a proper loss given in Condition 2 is called the dual representation.

Here, we contrast our Theorem 5 with related results. Indeed, a representation of proper losses that uses $\lambda_F(\boldsymbol{v}, z)$ is not new. Reid et al. (2015) showed that proper losses can be written with $\underline{L}^*$. van Rooyen & Williamson (2018) also showed with different proof techniques that any proper loss has the form of Condition 2 in Theorem 5. In a more general context, Nowak-Vila et al. (2019) and Blondel et al. (2020) discussed loss functions for structured prediction and arrived at the same representation. There is another line of research on the related notions of matching losses (Kivinen & Warmuth, 1997) and the Bregman divergence (Bregman, 1967; Banerjee et al., 2005), which are the special case of proper losses that have strictly convex and continuously differentiable Bayes risks. In particular, Amid et al. (2019) proved that matching losses have the dual representation. However, none of those previous studies obtained Condition 1 in Theorem 5, and therefore, they only succeeded in proving the necessity of the dual representation. In contrast, Theorem 5 gives necessary and sufficient conditions for a loss to be proper, which is made possible by constraining the convex functions by Condition 1. The theorem is also applicable to general proper losses that may possibly have non-smooth or not strictly convex Bayes risks.

Consider a proper loss $l(\boldsymbol{p}, z) = \lambda_F(\underline{\nabla} F^*(\boldsymbol{p}), z)$. If $\underline{\nabla} F^*(\boldsymbol{p})$ is invertible on $\mathcal{P}(\mathcal{Z})$, then λ_F can be regarded as a composite proper loss with a link function $\underline{\nabla} F^*(\boldsymbol{p})$ (Williamson et al., 2016). In this case, we can use a model that outputs a value on $\mathcal{C}$ instead of the class posterior probabilities.

This approach has practical advantages. Given $\boldsymbol{v} \in \mathcal{C}$, a loss is just $\lambda_F(\boldsymbol{v}, z)$ and is always guaranteed to be convex as a function of $\boldsymbol{v} \in \mathcal{C}$. This may facilitate optimization. In addition, once the best prediction $\hat{\boldsymbol{v}}$ is obtained, it can be converted into class probabilities by using $\boldsymbol{p} \in \partial F(\hat{\boldsymbol{v}})$. That is, we can completely circumvent calculation of the convex conjugate F^*, which may not be straightforward in general. Because a subdifferential map ∂F^* of a strictly convex function F^* is injective (see Appendix D), it follows that $\underline{\nabla} F^*(\boldsymbol{p})$ is invertible if $F^*(\boldsymbol{p})$ is strictly convex, or equivalently, if $l(\boldsymbol{p}, z)$ is strictly proper.

Note that F^* and $-\underline{L}$ are different in a subtle way, though they are closely related: the Bayes risk is defined only on $\mathcal{P}(\mathcal{Z})$, while F^* has a larger domain. For example, F^* might be finite at points in $\operatorname{aff}\mathcal{P}(\mathcal{Z}) \setminus \mathcal{P}(\mathcal{Z})$, where $\operatorname{aff}\mathcal{P}(\mathcal{Z})$ represents the affine hull of $\mathcal{P}(\mathcal{Z})$. It also holds that $F^*(\boldsymbol{p}) = F^*(\boldsymbol{p} + t\mathbf{1}_{\mathcal{Z}})$ for all $p \in \mathcal{P}(\mathcal{Z})$ and $t \in \mathbb{R}$, but $\boldsymbol{p} + t\mathbf{1}_{\mathcal{Z}}$ is not in $\mathcal{P}(\mathcal{Z})$ if $t \neq 0$.

This might lead us to suspect that minimizing $\lambda_F(\boldsymbol{v}, z)$ can result in a solution that does not correspond to posterior probabilities in $\mathcal{P}(\mathcal{Z})$. Indeed, even if F satisfies the conditions in Theorem 5, there might be a point $\boldsymbol{v} \in \mathcal{C}$ for which any solution of $\underline{\nabla} F^*(\boldsymbol{p}) = \boldsymbol{v}$ does not belong to $\mathcal{P}(\mathcal{Z})$. This is because the theorem guarantees the convex conjugate of F to be well-defined in $\mathcal{P}(\mathcal{Z})$ but also allows it to exist outside $\mathcal{P}(\mathcal{Z})$. However, the following proposition, which is proved in Appendix C.2, guarantees that minimizers of the loss always correspond to some point in $\mathcal{P}(\mathcal{Z})$.

Proposition 6. *Let $F : \mathcal{C} \to \mathbb{R}$ be a convex function that satisfies the conditions in Theorem 5. Then, any minimizer $\boldsymbol{v}$ of $\mathbb{E}_{z\sim\boldsymbol{p}}[\lambda_F(\boldsymbol{v}, z)] = \sum_{z\in\mathcal{Z}} p_z \lambda_F(\boldsymbol{v}, z)$, if one exists, satisfies $\underline{\nabla} F^*(\boldsymbol{q}) = \boldsymbol{v}$ for some $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$, where F^* is the convex conjugate of F.*

4. Characterization of T-Proper Losses

In this section, we characterize T-proper losses, which may possibly be lower-unbounded. Our main theorem here is closely related to backward correction in that it involves inversion of a label-corruption process. However, because our result gives necessary and sufficient conditions for T-properness, it also holds for forward-corrected losses and any other T-proper losses.

For a closed convex function $F : \mathcal{C} \subset \mathbf{1}_{\mathcal{Z}}^{\perp} \to \mathbb{R}$ and a reconstruction matrix R for weak labels $\mathcal{Y}$, we define a function $\lambda_{F,R} : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$ as

$$\lambda_{F,R}(\boldsymbol{v}, y) = -(R^{\mathsf{T}}\boldsymbol{v})_y + F(\boldsymbol{v}). \tag{13}$$

Then we can state the main theorem of this section as follows:

Theorem 7. *Let T be a label transition matrix for weak labels $\mathcal{Y}$, and let $l_{\mathrm{W}} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a weak-label loss. Then, l_{W} is T-proper if and only if there exist a closed convex function $F : \mathcal{C} \subset \mathbf{1}_{\mathcal{Z}}^{\perp} \to \mathbb{R}$, a reconstruction matrix R of T, and a function $\boldsymbol{\Delta}(\boldsymbol{q})$ taking values on the cokernel[1] of T, which satisfy the following conditions:*

1. *$F(\boldsymbol{v}) - \max_{z\in\mathcal{Z}} v_z$ is bounded from below.*
2. *It holds that $l_{\mathrm{W}}(\boldsymbol{q}, y) = \lambda_{F,R}(\underline{\nabla} F^*(\boldsymbol{q}), y) + \Delta_y(\boldsymbol{q})$, where $\underline{\nabla} F^*(\boldsymbol{q})$ is an appropriately chosen subgradient function.*

See Appendix C.3 for a proof.

Because of the assumption of reconstructibility, we have that $\boldsymbol{\Delta}(\boldsymbol{q}) \equiv \mathbf{0}$ if $|\mathcal{Z}| = |\mathcal{Y}|$. On the other hand, if $|\mathcal{Z}| < |\mathcal{Y}|$, a label transition matrix T has a cokernel of nonzero dimension, and therefore, $\boldsymbol{\Delta}(\boldsymbol{q})$ might take finite values. However, even if $\boldsymbol{\Delta}(\boldsymbol{q}) \neq 0$ for some $\boldsymbol{q}$, by the definition of $\operatorname{coker} T$, we have that $\langle T\boldsymbol{p}, \boldsymbol{\Delta}(\boldsymbol{q})\rangle = 0$ for all $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$, which leads to the following proposition:

Proposition 8. *The function $\boldsymbol{\Delta}(\boldsymbol{q})$ in Theorem 7 does not contribute to the expected loss; that is, $\mathbb{E}_{y\sim T\boldsymbol{p}}[\Delta_y(\boldsymbol{q})] = 0$ for all $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$. In particular, it holds that $\boldsymbol{\Delta}(\boldsymbol{q}) \equiv \mathbf{0}$ if $|\mathcal{Z}| = |\mathcal{Y}|$.*

Two well-known classes of T-proper losses are forward and backward correction losses. Because Theorem 7 is applicable to any T-proper loss, the loss functions of these classes also conform to it. We demonstrate this in the following two examples.

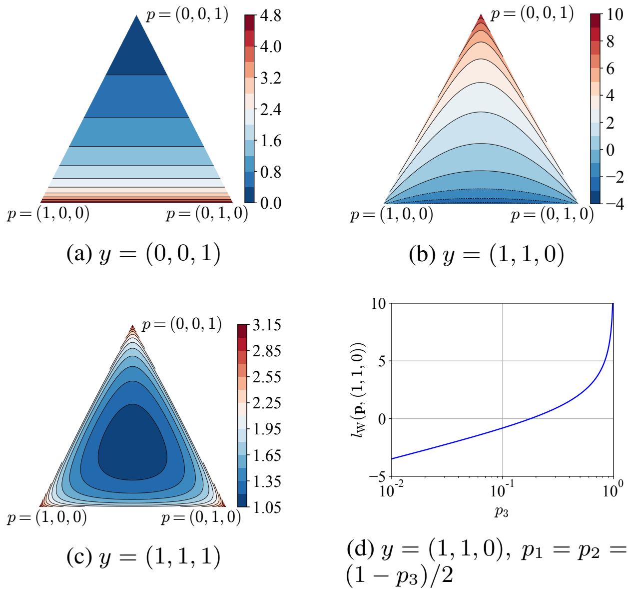
Example 9 (Forward correction)**.** Let $l_{\mathcal{Y}} : \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a proper loss for estimating weak-label posterior probabilities. Note the difference from a weak-label loss $l_{\mathrm{W}} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ and a proper loss $l : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \overline{\mathbb{R}}$ for supervised learning. A weak-label loss $l_{\mathrm{W}} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ is called the forward correction of $l_{\mathcal{Y}}$ if $l_{\mathrm{W}}(\boldsymbol{q}, y) = l_{\mathcal{Y}}(T\boldsymbol{q}, y)$. Its T-properness is a consequence of the properness of $l_{\mathcal{Y}}$ and the reconstructibility of T. In Appendix E, we prove that forward correction losses conform to Theorem 7. □

Example 10 (Backward correction)**.** Let $l : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \overline{\mathbb{R}}$ be a proper loss for fully supervised learning. A backward-corrected loss $l_{\mathrm{W}} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ associated with l is defined as $l_{\mathrm{W}}(\boldsymbol{q}, y) = \sum_{z\in\mathcal{Z}} R_{zy} l(\boldsymbol{q}, z)$. By applying Theorem 5 to $l(\boldsymbol{q}, z)$, we find that $l_{\mathrm{W}}(\boldsymbol{q}, y) = -[R^{\mathsf{T}}\underline{\nabla} F^*(\boldsymbol{q})]_y + F(\underline{\nabla} F^*(\boldsymbol{q}))(R^{\mathsf{T}}\mathbf{1}_{\mathcal{Z}})_y$. It can be shown that $R^{\mathsf{T}}\mathbf{1}_{\mathcal{Z}} - \mathbf{1}_{\mathcal{Y}} \in \operatorname{coker} T$ (see Appendix F for a proof). Therefore, the backward-corrected loss has the form given in Theorem 7 with $\boldsymbol{\Delta}(\boldsymbol{q}) = F(\underline{\nabla} F^*(\boldsymbol{q}))(R^{\mathsf{T}}\mathbf{1}_{\mathcal{Z}} - \mathbf{1}_{\mathcal{Y}})$. For any label transition matrix T, we can choose a reconstruction matrix R such that $R^{\mathsf{T}}\mathbf{1}_{\mathcal{Z}} = \mathbf{1}_{\mathcal{Y}}$ (see Appendix C.6 in van Rooyen & Williamson (2018) and Appendix F in this paper); therefore, we can always make $\boldsymbol{\Delta}(\boldsymbol{q})$ zero for a backward-corrected loss by using an appropriate R. □

5. Lower-Boundedness of T-Proper Losses

T-proper losses as constructed in Theorem 7 may not be bounded from below. Indeed, there is a gap between the boundedness criteria for proper losses and T-proper losses. In Section 5.1, we see this for an example of the softmax cross-entropy loss. In Section 5.2, we give a sufficient

[1] The cokernel of T is the kernel of T^{T}, i.e., a set of vectors $\boldsymbol{v}$ in $\mathbb{R}^{\mathcal{Y}}$ such that $T^{\mathsf{T}}\boldsymbol{v} = \mathbf{0}$.



(a) $y = (0, 0, 1)$

(b) $y = (1, 1, 0)$

(c) $y = (1, 1, 1)$

(d) $y = (1, 1, 0),\ p_1 = p_2 = (1 - p_3)/2$

Figure 1. Weak-label loss $l_{\mathrm{W}}(\boldsymbol{p}, y)$ with $F(\boldsymbol{v}) = \sum_{z \in \mathcal{Z}} \log(e^{v_z})$ for partial labels (Example 12). (a–c) $l_{\mathrm{W}}(\boldsymbol{p}, y)$ for all $\boldsymbol{p} \in \mathcal{P}(\mathcal{Z})$. By symmetry, the plots for weak labels y that are not shown here can be obtained by rotating one of these plots. (d) $l_{\mathrm{W}}(\boldsymbol{p}, (1, 1, 0))$ for $\boldsymbol{p} \in \mathcal{P}(\mathcal{Z})$ such that $p_1 = p_2 = (1 - p_3)/2$.

condition under which a T-proper loss is bounded from below.

5.1. T-Proper Loss May Not Be Bounded from Below

Consider a T-proper weak-label loss $l_{\mathrm{W}}(\boldsymbol{q}, y) = \lambda_{F,R}(\underline{\nabla} F^*(\boldsymbol{q}), y)$ with $\boldsymbol{\Delta}(\boldsymbol{q}) = \boldsymbol{0}$. To see if $\lambda_{F,R}(\boldsymbol{v}, y)$ is bounded from below, we need to compare $F(\boldsymbol{v})$ with $R^{\mathsf{T}}\boldsymbol{v}$. On the other hand, any regular proper loss is bounded from below, because the definition of regularity requires that the loss must not be negative infinity on the compact probability simplex. This is also reflected in Condition 1 of Theorem 7, which suffices to ensure the lower-boundedness of a loss of the form $-v_z + F(\boldsymbol{v})$. The following lemma implies that the boundedness of T-proper losses imposes a stronger restriction on $F(\boldsymbol{v})$ than that of proper losses.

Lemma 11. *Let R be a reconstruction matrix. Then $\max_{y \in \mathcal{Y}} (R^{\mathsf{T}}\boldsymbol{v})_y \geq \max_{z \in \mathcal{Z}} v_z$ for any vector $\boldsymbol{v} \in \mathbf{1}_{\mathcal{Z}}^{\perp}$.*

See Appendix C.4 for a proof.

Example 12. Consider $F(\boldsymbol{v}) = \log(\sum_{z \in \mathcal{Z}} e^{v_z})$, which corresponds to the softmax cross-entropy loss and satisfies $F(\boldsymbol{v}) > \max_{z \in \mathcal{Z}} v_z$ for all $\boldsymbol{v} \in \mathcal{Z}$. Also, up to an exponentially small correction, it holds that $F(t\boldsymbol{v}) \simeq t \max_{z \in \mathcal{Z}} v_z$ for $\boldsymbol{v} \neq \boldsymbol{0}$ and large positive t. This fact and Lemma 11 imply that $\lambda_{F,R}(t\boldsymbol{v}, y) \simeq t[-(R^{\mathsf{T}}\boldsymbol{v})_y + \max_{z \in \mathcal{Z}} v_z] \leq 0$ for $y \in \arg\max_{y \in \mathcal{Y}} (R^{\mathsf{T}}\boldsymbol{v})_y$. If we choose $\boldsymbol{v}$ such that this inequality is strict, this diverges to negative infinity as $t \to \infty$. Therefore, the weak-label loss constructed by applying Theorem 7 to this $F(\boldsymbol{v})$ with $\boldsymbol{\Delta}(\boldsymbol{q}) = \boldsymbol{0}$ is not bounded from below. To provide a concrete example, we examine partial labels as described in Example 2. Here, we take a reconstruction matrix

$$R = \begin{pmatrix} 1 & 0 & 0 & \frac{3-2p}{3(1-p)} & \frac{3-2p}{3(1-p)} & -\frac{3-p}{3(1-p)} & \frac{1}{3} \\ 0 & 1 & 0 & \frac{3-2p}{3(1-p)} & -\frac{3-p}{3(1-p)} & \frac{3-2p}{3(1-p)} & \frac{1}{3} \\ 0 & 0 & 1 & -\frac{3-p}{3(1-p)} & \frac{3-2p}{3(1-p)} & \frac{3-2p}{3(1-p)} & \frac{1}{3} \end{pmatrix}, \tag{14}$$

and we set $p = 0.1$. Figures 1(a–c) show contour plots of $l_{\mathrm{W}}(\boldsymbol{p}, y)$ for all $\boldsymbol{p} \in \mathcal{P}(\mathcal{Z})$ and $y = (0, 0, 1), (1, 1, 0)$, and $(1, 1, 1)$. We can just rotate these plots to find the plots for the other weak labels. Among these, $l_{\mathrm{W}}(\boldsymbol{p}, (1, 1, 0))$ is not bounded from below. To make the divergence clearer, Fig. 1(d) shows the same function on the line satisfying $p_1 = p_2$. The plot suggests that the loss indeed diverges logarithmically to negative infinity, or equivalently, it diverges linearly in the logit, which is consistent with the above discussion. □

5.2. Sufficient Condition for Lower-Boundedness

Now, we are ready to state a sufficient condition for T-proper losses to be bounded from below. Lemma 11 implies that if $F(\boldsymbol{v}) - \max_{y \in \mathcal{Y}} (R^{\mathsf{T}}\boldsymbol{v})_y$ has a lower bound on $\mathcal{C}$, then Condition 1 in Theorem 7 is automatically satisfied. Therefore, we have the following theorem.

Theorem 13. *Let T be a label transition matrix for weak labels $\mathcal{Y}$, and let $F : \mathcal{C} \subset \mathbf{1}_{\mathcal{Z}}^{\perp} \to \mathbb{R}$ be a closed convex function. If $F(\boldsymbol{v}) - \max_{y \in \mathcal{Y}} (R^{\mathsf{T}}\boldsymbol{v})_y$ is bounded from below in $\mathcal{C}$, then a weak-label loss $l_{\mathrm{W}}(\boldsymbol{q}, y) = \lambda_{F,R}(\underline{\nabla} F^*(\boldsymbol{q}), y)$ is T-proper and lower-bounded, where R is a reconstruction matrix of T, $\underline{\nabla} F^*(\boldsymbol{q})$ is a subgradient function of the convex conjugate $F^*(\boldsymbol{q})$ of $F(\boldsymbol{v})$, and the function $\lambda_{F,R}$ is defined as $\lambda_{F,R}(\boldsymbol{v}, y) = -(R^{\mathsf{T}}\boldsymbol{v})_y + F(\boldsymbol{v})$.*

Theorem 13 gives a sufficient condition for a T-proper loss to have a lower bound, but it is not necessary. For example, a T-proper loss is not of the above form whenever it has a contribution of $\boldsymbol{\Delta}(\boldsymbol{q})$, as in Theorem 7, that cannot be absorbed in $\lambda_{F,R}(\boldsymbol{q}, y)$. Still, Theorem 13 gives a large class of lower-bounded T-proper losses that are parameterized by a convex function F that is constrained only by a single inequality.

We can also interpret the condition in Theorem 13 in terms of its dual, or the Bayes risk. Crudely speaking, the condition can be understood as a constraint to ensure that the Bayes risk is finite at "class probabilities given a weak label." More precisely, the following proposition paraphrases the condition imposed on F in Theorem 13 into a condition on F^*. See Appendix C.5 for a proof.

Proposition 14. *Let R be a reconstruction matrix for weak labels $\mathcal{Y}$, and let $F : \mathcal{C} \subset \mathbf{1}_{\mathcal{Z}}^{\perp} \to \mathbb{R}$ be a closed convex*

Algorithm 1 Training of the linear model with the backward-corrected cross entropy and generalized logit squeezing.

Input: training data $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, reconstruction matrix R, coefficient k, exponent α, batch size N, SGD-like algorithm $\mathcal{A}$.
Output: weight matrix W
Initialize weights W
repeat
  Sample minibatch $(X, \boldsymbol{y})$ from D
  $V \leftarrow XW^\top$
  $l_{\text{ce}} \leftarrow \frac{1}{N}\sum_{i=1}^{N}[-(VR)_{iy_i} + \log\sum_z \exp(v_{iz})]$
  $l_{\text{gLS}} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\sum_z \frac{k}{2}|v_{iz}|^\alpha$
  $l \leftarrow l_{\text{ce}} + l_{\text{gLS}}$
  Update W by using an algorithm $\mathcal{A}$
until a stopping criterion is met

function. Then, $F(\boldsymbol{v}) - \max_{y\in\mathcal{Y}}(R^\top\boldsymbol{v})_y$ is bounded from below in $\mathcal{C}$ if and only if $F^*(R\boldsymbol{e}_y) < \infty$ for all $y \in \mathcal{Y}$, where $\boldsymbol{e}_y \in \mathcal{P}(\mathcal{Y})$ is a distribution over weak labels that concentrates on a single weak label y.

The condition $F^*(R\boldsymbol{e}_y) < \infty$ can be informally paraphrased as $\underline{L}(R\boldsymbol{e}_y) > -\infty$, because F^* and $-\underline{L}$ are equal in $\mathcal{P}(\mathcal{Z})$. Note, however, that $R\boldsymbol{e}_y$ is not necessarily in $\mathcal{P}(\mathcal{Z})$ because of the negative components of R, and therefore, F^* and $-\underline{L}$ may not be equal at $R\boldsymbol{e}_y$.

A function $\lambda_{F,R}(\boldsymbol{v}, y)$ is convex as a function of $\boldsymbol{v} \in \mathcal{C}$, because it is a sum of the linear function $(R^\top\boldsymbol{v})_y$ and a convex function $F(\boldsymbol{v})$ (van Rooyen & Williamson, 2018). As with proper losses, therefore, we can obtain the benefits of the convexity of $\lambda_{F,R}(\boldsymbol{v}, y)$ by using $\mathcal{C}$-valued models.

5.3. Generalized Logit Squeezing

If we note that $\max_{y\in\mathcal{Y}}(R^\top\boldsymbol{v})_y$ is a positively homogeneous function of degree 1, then any convex function F that grows superlinearly satisfies the condition of Theorem 13. This fact leads to the following corollary of Theorem 13, which gives a useful way to regularize an unbounded loss.

Corollary 15. *Let $F : \mathcal{C} \subset \mathbf{1}^{\perp}_{\mathcal{Z}} \to \mathbb{R}$ be a convex function, let α be a real number that is greater than 1, and let k be a positive number. We define a convex function F' as*

$$F'(\boldsymbol{v}) = F(\boldsymbol{v}) + \frac{k}{2}\sum_{z\in\mathcal{Z}}|v_z|^\alpha \tag{15}$$

for $\boldsymbol{v} \in \mathcal{C}$. Then, a weak-label loss $l_{\mathrm{W}}(\boldsymbol{q}, y) = \lambda_{F',R}(\underline{\nabla}F'^*(\boldsymbol{q}), y)$ is T-proper and lower-bounded.

To facilitate the use of this corollary, we present pseudocode for training the linear model with the backward-corrected cross entropy loss in Algorithm 1.

Table 1. Comparison of three losses. BC, GA, and gLS respectively stand for backward correction, gradient ascent, and generalized logit squeezing.

	Proper	Bounded
BC	✓	✗
BC + GA	✗	✓
BC + gLS	✓	✓

The term $\sum_{z\in\mathcal{Z}}|v_z|^\alpha$ is convex if and only if $\alpha \geq 1$. The conclusion of the corollary for $\alpha = 1$ depends on the precise form of F and the value of k.

Corollary 15 indicates that if a T-proper loss associated with F is not bounded from below, then we can replace F with $F + \frac{k}{2}\sum_{z\in\mathcal{Z}}|v_z|^\alpha$ to make the loss bounded while keeping its T-properness. We refer to the proposed regularization scheme of Eq. (15) as generalized logit squeezing (gLS). The special case with $\alpha = 2$ has the same form as the regularization schemes called feature contraction (Li & Maki, 2018) and logit squeezing (Kannan et al., 2018). Those previous studies focused on the performance of supervised learning (Li & Maki, 2018) or adversarial robustness (Kannan et al., 2018), and they were mostly empirical. On the other hand, gLS has a solid theoretical foundation that guarantees its asymptotic success in weakly supervised learning.

Although gLS might appear similar to L^p regularization, they are different concepts. gLS penalizes a model's large output values, whereas normal L^p regularization pulls training trajectories toward smaller norms of the weights. They both restrict the model space but in different ways, and their actual effects on learning might be very different. On the other hand, if $\boldsymbol{v} \in \mathcal{C}$ is a linear function of the weights, then gLS is closely related to L^p regularization, because the gLS term is a positively homogeneous function of degree α on the weights. In this particular case, the two regularization schemes could be expected to work in a similar way.

6. Experiment

In this section, we experimentally compare three different losses, all of which derive from the cross-entropy loss, to demonstrate the effectiveness of lower-bounded proper losses[2]. Table 1 summarizes these losses. We take the backward correction (BC) of the softmax cross entropy as a baseline loss:

$$\lambda_{F,R}(\boldsymbol{v}, y) = -(R^\top\boldsymbol{v})_y + \log\sum_{z\in\mathcal{Z}} e^{v_z}, \tag{16}$$

which is proper but lower-unbounded. Here, $\boldsymbol{v}$ is so-called logits, which can be converted into class posterior probabili-

[2] The code is publicly available at https://github.com/yoshum/lower-bounded-proper-losses.

Table 2. Mean and sample standard deviation of the test accuracy. The best accuracy for each dataset and model is shown in boldface. BC: backward correction; GA: gradient ascent; gLS: generalized logit squeezing, with the exponent α fixed to 2.

	Weight decay	MNIST, linear	MNIST, MLP	CIFAR-10, ResNet-20	CIFAR-10, WRN-28-2
BC	fixed	$81.52 \pm 1.44\,\%$	$83.09 \pm 0.67\,\%$	$28.86 \pm 2.06\,\%$	$29.57 \pm 1.58\,\%$
BC	tuned	$83.56 \pm 0.87\,\%$	$83.30 \pm 1.01\,\%$	$29.56 \pm 1.49\,\%$	$30.02 \pm 1.49\,\%$
BC + GA	fixed	$78.57 \pm 1.82\,\%$	$87.88 \pm 1.11\,\%$	$34.39 \pm 2.96\,\%$	$36.87 \pm 2.26\,\%$
BC + GA	tuned	$80.63 \pm 1.01\,\%$	$\mathbf{89.15} \pm 0.75\,\%$	$35.36 \pm 1.80\,\%$	$36.90 \pm 2.52\,\%$
BC + gLS	fixed	$\mathbf{83.77} \pm 0.55\,\%$	$88.63 \pm 0.38\,\%$	$\mathbf{49.71} \pm 3.04\,\%$	$\mathbf{49.98} \pm 2.59\,\%$

ties with the softmax function. In our experiments, this loss is made bounded from below in two different ways. The first way is to apply gLS to the backward-corrected cross entropy and use

$$\lambda_{F,R}(\boldsymbol{v}, y) = -(R^{\mathsf{T}}\boldsymbol{v})_y + \log \sum_{z\in\mathcal{Z}} e^{v_z} + \frac{k}{2} \sum_{z\in\mathcal{Z}} |v_z|^{\alpha}. \tag{17}$$

For brevity, we refer to this loss as BC + gLS. It is proper and lower-bounded if $\alpha > 1$, while it becomes improper and lower-unbounded if $\alpha < 1$. The properties for the boundary case of $\alpha = 1$ depend on the value of k. The other way to make the loss bounded from below is to use gradient ascent (GA) (Kiryo et al., 2017; Ishida et al., 2019; Lu et al., 2020), which updates a model in the ascending direction of the loss surface when the empirical class-conditional risk becomes negative. GA makes the training objective bounded from below but improper.

6.1. Setup

As a specific example of weak labels, we experimented with complementary labels (Ishida et al., 2017; Yu et al., 2018; Ishida et al., 2019). Let c be any category label. Then, a complementary label $\overline{c}$ put on an instance indicates that it belongs to a category other than c. For K-class classification, (unbiased) complementary labels are characterized by the following transition matrix:

$$T = \frac{1}{K-1}(1_K - I_K), \tag{18}$$

where 1_K is the $K \times K$ matrix with all elements 1, and I_K is the $K \times K$ identity matrix. This can be seen as an extreme case of noisy labels, where a label is corrupted with probability 1.

We evaluated the effectiveness of the losses on the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) datasets. To each instance in these datasets, we randomly assigned a complementary label with conditional probabilities given the ground-truth category, which is given by Eq. (18). For each dataset, we trained two models: a linear model and a feed-forward network with one hidden layer (multilayer perceptron; MLP) were used with MNIST, and ResNet-20 (He et al., 2016) and Wide-ResNet (WRN) 28-2 (Zagoruyko & Komodakis, 2016) were used with CIFAR-10. We used stochastic gradient descent with momentum to optimize the models. The momentum was fixed to 0.9, while the initial learning rates were chosen as those giving the best validation accuracy. The default value of the weight decay was 10^{-4}, but we also tuned it with BC and BC + GA to compare its effect with that of gLS. More details on the experimental procedure and the hyperparameters are given in Appendix G.

6.2. Results

In Table 2, we list the mean and the sample standard deviation of the test accuracy for 16 trials with the chosen hyperparameters. Here, we tuned the coefficient k and fixed the exponent $\alpha = 2$ for BC + gLS. If the weight decay was fixed to the default value for all the models, BC + gLS achieved the best test accuracy by a clear margin. It was particularly effective on the CIFAR-10 benchmark, which used more complex models. This is reasonable, because complex models are easier to fit to an unbounded training loss and are affected more severely than simple models; therefore, they are more sensitive to how the lower-unboundedness is prevented by regularization.

On the other hand, the effects of tuning the weight decay were not consistent among the different models. In the experiments with BC *without* GA, we found that tuning the weight decay brought a gain of approximately 2% to the linear model, which enabled BC to achieve performance comparable to that of the proposed method (BC + gLS), but the gains were insignificant for the other models. This is consistent with the observation that in the linear model, the two regularization schemes have similar functional forms, as explained in the previous section. By contrast, a larger weight decay seemed to penalize complex models too much and cause underfitting before it prevented loss divergence. The experiments with BC + GA showed a similar trend, except for the MLP model, which exhibited a gain of about 2% from tuning the weight decay. Overall, the weight decay could narrow or close the gap between the baselines and the proposed method for simpler models, but it did not have a significant effect for deeper models, which more severely

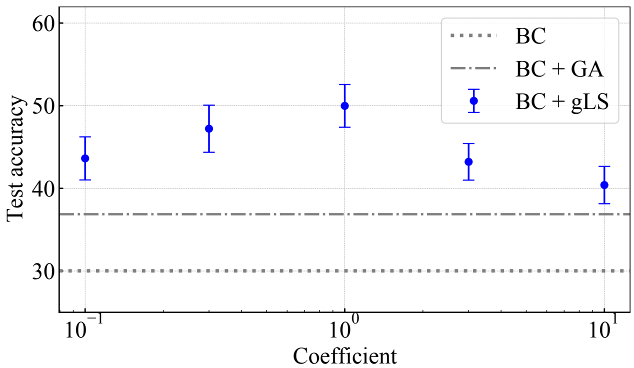


Figure 2. Sensitivity of the test accuracy to the coefficient of the squared logit term in Eq. (17). The bars represent the sample standard deviations. The horizontal lines indicate the test accuracies of BC (dotted) and BC + GA (dashed-dotted) with the weight decay coefficients that achieved the best validation accuracy.

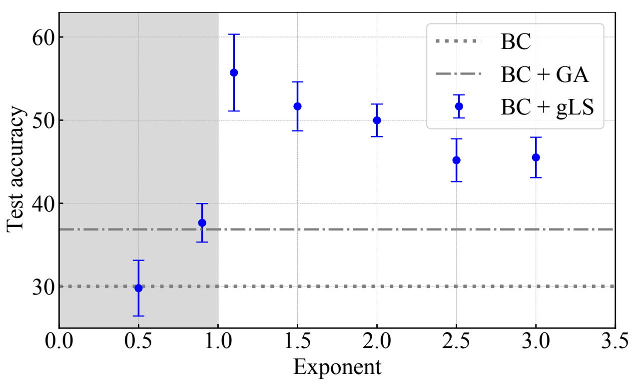


Figure 3. Performance of gLS for different exponents. The bars represent the sample standard deviations. The gray region ($\alpha < 1$) represents the regime in which gLS yields an improper, lower-unbounded loss. The horizontal lines indicate the test accuracies of BC (dotted) and BC + GA (dashed-dotted) with the weight decay coefficients that achieved the best validation accuracy.

suffer from overfitting due to a lower-unbounded loss.

We also examined how sensitive the accuracy is to the coefficient k of the gLS term in Eq. (17). Figure 2 shows the test accuracies for different k values on CIFAR-10 trained with WRN-28-2, and it indicates two findings. First, the test accuracy depended significantly on k, and it is thus important to choose an appropriate value of k to obtain the best results. In a sense, this is an obvious conclusion: both of the limits, $k \to 0$ and $k \to \infty$, are undesirable, because the former would converge to BC, while the latter would lead to a model that outputs zero for any input; therefore, there should be an optimal value of k. Second, however, the figure indicates that the results were not too sensitive to k and that BC + gLS yielded better results over two orders of magnitude of k as compared to the other methods.

As Corollary 15 indicates, gLS yields lower-bounded T-proper losses as long as $\alpha > 1$. In Figure 3, we show the test accuracies on CIFAR-10 trained with WRN-28-2 by using various exponents. The results demonstrate that gLS gave superior results as compared to the baseline methods, even with $\alpha \neq 2$. Interestingly, the test accuracies improved as $\alpha \to 1$. As α became less than 1, however, the test accuracies immediately dropped to values similar to the baselines. This observation not only validates the effectiveness of our proposed approach but also underlines the importance of using T-proper and lower-bounded losses, which is the central premise that motivated our theoretical analysis.

7. Conclusion

In this paper, we have discussed proper losses for weakly supervised classification. We first derived the dual representation of proper losses for supervised learning. Instead of the Bayes risk, which plays a central role in the Savage representation, the derived theorem represents a loss with a function related to the convex conjugate of the Bayes risk. We then used this theorem to characterize T-proper losses and derived a sufficient condition for them to be bounded from below. These theoretical findings led to a novel regularization scheme called generalized logit squeezing (gLS), which prevents any proper weak-label loss from diverging to negative infinity, while keeping the properness of the original loss. We also experimentally demonstrated the effectiveness of our proposed approach. Remarkably, gLS yielded superior results as compared to the baseline methods regardless of the precise values of the hyperparameters that are specific to gLS, as long as those parameters were in the regime in which gLS gives T-proper and lower-bounded losses.

Acknowledgement

SMY gratefully acknowledges fruitful discussions with Akira Tanimoto and Makoto Terao. TT was partially supported by JSPS KAKENHI Grant Numbers 20K03753 and 19H04071.

References

Amid, E., Warmuth, M. K. K., Anil, R., and Koren, T. Robust Bi-Tempered Logistic Loss Based on Bregman Divergences. In Advances in Neural Information Processing Systems, pp. 15013–15022, 2019.

Angluin, D. and Laird, P. Learning From Noisy Examples. Machine Learning, 2(4):343–370, April 1988. ISSN 1573-0565. doi: 10.1023/A:1022873112823.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman Divergences. Journal of Machine Learning Research, 6(Oct):1705–1749, 2005. ISSN ISSN 1533-7928.

Blondel, M., Martins, A. F. T., and Niculae, V. Learning with Fenchel-Young losses. Journal of Machine Learning Research, 21(35):1–69, 2020. ISSN 1533-7928.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 7(3):200–217, January 1967. ISSN 0041-5553. doi: 10.1016/0041-5553(67)90040-7.

Buja, A., Stuetzle, W., and Shen, Y. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, November 2005.

Cabannnes, V., Rudi, A., and Bach, F. Structured prediction with partial labelling through the infimum loss. In III, H. D. and Singh, A. (eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119, pp. 1230–1239. PMLR, 13–18 Jul 2020.

Chou, Y.-T., Niu, G., Lin, H.-T., and Sugiyama, M. Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels. In International Conference on Machine Learning, pp. 1929–1938. PMLR, November 2020.

Cid-sueiro, J. Proper losses for learning from partial labels. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), Advances in Neural Information Processing Systems 25, pp. 1565–1573. Curran Associates, Inc., 2012.

Cid-Sueiro, J., Arribas, J., Urban-Munoz, S., and Figueiras-Vidal, A. Cost functions to estimate a posteriori probabilities in multiclass problems. IEEE Transactions on Neural Networks, 10(3):645–656, May 1999. ISSN 1941-0093. doi: 10.1109/72.761724.

Cid-Sueiro, J., García-García, D., and Santos-Rodríguez, R. Consistency of Losses for Learning from Weak Labels. In Calders, T., Esposito, F., Hüllermeier, E., and Meo, R. (eds.), Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, pp. 197–210, Berlin, Heidelberg, 2014. Springer. ISBN 978-3-662-44848-9. doi: 10.1007/978-3-662-44848-9_13.

Cour, T., Sapp, B., and Taskar, B. Learning from Partial Labels. The Journal of Machine Learning Research, 12 (null):1501–1536, July 2011. ISSN 1532-4435.

Dawid, A. P. The geometry of proper scoring rules. Annals of the Institute of Statistical Mathematics, 59(1): 77–93, March 2007. ISSN 1572-9052. doi: 10.1007/s10463-006-0099-8.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 213, Las Vegas, Nevada, USA, 2008. ACM Press. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401920.

Gneiting, T. and Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102(477):359–378, March 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

Ishida, T., Niu, G., Hu, W., and Sugiyama, M. Learning from Complementary Labels. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), Advances in Neural Information Processing Systems 30, pp. 5639–5649. Curran Associates, Inc., 2017.

Ishida, T., Niu, G., Menon, A., and Sugiyama, M. Complementary-Label Learning for Arbitrary Losses and Models. In International Conference on Machine Learning, pp. 2971–2980, May 2019.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial Logit Pairing. arXiv:1803.06373 [cs, stat], March 2018.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), Advances in Neural Information Processing Systems 30, pp. 1675–1685. Curran Associates, Inc., 2017.

Kivinen, J. and Warmuth, M. K. Relative loss bounds for multidimensional regression problems. In Proceedings of the 10th International Conference on Neural Information Processing Systems, NIPS'97, pp. 287–293, Cambridge, MA, USA, December 1997. MIT Press.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Cortes, C., and Burges, C. J. C. The MNIST database of handwritten digits, 1998.

Li, V. and Maki, A. Feature Contraction: New ConvNet Regularization in Image Classification. In Proceedings of the British Machine Vision Conference 2018, pp. 213, 2018.

Lu, N., Zhang, T., Niu, G., and Sugiyama, M. Mitigating Overfitting in Supervised Classification from Two Unlabeled Datasets: A Consistent Risk Correction Approach. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics 2020, March 2020.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with Noisy Labels. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), Advances in Neural Information Processing Systems 26, pp. 1196–1204. Curran Associates, Inc., 2013.

Nowak-Vila, A., Bach, F., and Rudi, A. A General Theory for Structured Prediction with Smooth Convex Surrogates. arXiv:1902.01958 [cs, stat], February 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc., 2019.

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.

du Plessis, M. C., Niu, G., and Sugiyama, M. Convex Formulation for Learning from Positive and Unlabeled Data. In International Conference on Machine Learning, pp. 1386–1394, June 2015.

Reid, M. D., Frongillo, R. M., Williamson, R. C., and Mehta, N. Generalized Mixability via Entropic Duality. In Conference on Learning Theory, pp. 1501–1522. PMLR, June 2015.

Rockafellar, R. T. Convex Analysis. Princeton University Press, Princeton, NJ, December 1996. ISBN 978-0-691-01586-6.

van Rooyen, B. and Williamson, R. C. A Theory of Learning with Corrupted Labels. Journal of Machine Learning Research, 18(228):1–50, 2018. ISSN 1533-7928.

Savage, L. J. Elicitation of Personal Probabilities and Expectations. Journal of the American Statistical Association, 66(336):783–801, 1971. ISSN 0162-1459. doi: 10.2307/2284229.

Williamson, R. C., Vernet, E., and Reid, M. D. Composite multiclass losses. The Journal of Machine Learning Research, 17(1):7860–7911, January 2016. ISSN 1532-4435.

Winkler, R. L. and Murphy, A. H. "Good" Probability Assessors. Journal of Applied Meteorology, 7(5):751–758, October 1968. ISSN 0021-8952. doi: 10.1175/1520-0450(1968)007⟨0751:PA⟩2.0.CO;2.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with Biased Complementary Labels. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), Computer Vision – ECCV 2018, Lecture Notes in Computer Science, pp. 69–85, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01246-5. doi: 10.1007/978-3-030-01246-5_5.

Zagoruyko, S. and Komodakis, N. Wide Residual Networks. In Proceedings of the British Machine Vision Conference 2016, 2016.

Zhang, Y., Charoenphakdee, N., and Sugiyama, M. Learning from Indirect Observations. arXiv:1910.04394 [cs, stat], October 2019.