Appendix

A. Relation of Reconstructibility to Other Assumptions in the Literature

In this appendix, we compare Assumption (b), i.e., the left-invertibility of the label transition matrix T, with other learnability/invertibility assumptions in the literature of weakly supervised learning.

A.1. Identifiability from Zhang et al. (2019)

Consider a probability distribution $P(X|\theta)$ that is parameterized by a set of parameters $\theta \in \Theta$. This parametric family of distributions satisfies the identifiability condition if

$$\forall \theta_1, \theta_2 \in \Theta, \theta_1 \neq \theta_2 \Longrightarrow P(X|\theta_1) \neq P(X|\theta_2). \tag{19}$$

In other words, if $P(X|\theta)$ is perfectly known, then the parameter θ can be uniquely identified.

Zhang et al. (2019) proved the consistency of their algorithm under several assumptions, among which the most fundamental is identifiability of the posterior probability distributions of weak and true labels. More precisely, they assumed that for any input pattern $x \in \mathcal{X}$, the posterior probability distribution of true labels, $P(Z|x)$, belongs to a parametric family of identifiable probability distributions. Let $P(Z|\theta)$ be a distribution in that family and Θ be a set of parameters. In addition, they also assumed that the posterior probabilities of weak labels are also identifiable; that is,

$$\forall \theta_1, \theta_2 \in \Theta, \theta_1 \neq \theta_2 \Longrightarrow P(Y|\theta_1) \neq P(Y|\theta_2), \tag{20}$$

where $P(Y = y|\theta) \equiv \sum_{z \in \mathcal{Z}} T_{yz} P(Z = z|\theta)$. Note that θ is different from model parameters such as weights in neural networks. Here, a model is a function $f_w : \mathcal{X} \to \Theta$ that is parameterized by a set of network weights w. For a given input x, it predicts a parameter $\theta = f_w(x)$ and in turn the posterior probabilities $P(Z|\theta)$.

The identifiability of true label distributions is automatically satisfied by careful implementation. For example, if the categorical posterior probabilities are expressed by using the softmax function, the choice $\Theta = \mathbf{1}_{\mathcal{Z}}^{\perp}$ guarantees identifiability. Therefore, we use this assumption in the discussion below.

Zhang et al. (2019) claimed that they successfully avoided relying on the existence of a left-inverse of T by resorting to the identifiability assumptions. However, without any prior knowledge on the true posterior probability distributions, identifiability implies the left-invertibility of T. Specifically, we can prove the following proposition.

Proposition 16. *Let T be a label transition matrix. Assume that for any $x \in \mathcal{X}$, a posterior probability distribution $P(Z|x)$ of true labels belongs to a parametric family $\{P(Z|\theta) \mid \theta \in \Theta\}$ of identifiable distributions. Then, the left-invertibility of T implies the identifiability of $P(Y|x)$, the posterior probability distribution of weak labels. Moreover, the converse also holds if $\{P(Z|\theta) \mid \theta \in \Theta\} = \mathcal{P}(\mathcal{Z})$.*

Proof. Suppose that T is left-invertible. Then, it holds that

$$P(Z = z|\theta) = \sum_{y \in \mathcal{Y}} R_{zy} P(Y = y|\theta), \tag{21}$$

where R is a left-inverse of T. This implies that if $P(Y|\theta_1) = P(Y|\theta_2)$, then $P(Z|\theta_1) = P(Z|\theta_2)$, from which it follows that $\theta_1 = \theta_2$ because of the identifiability of $P(Z|\theta)$. Therefore, Eq. (20) holds.

Conversely, suppose that $P(Y|\theta)$ is identifiable and also that $\{P(Z|\theta) \mid \theta \in \Theta\} = \mathcal{P}(\mathcal{Z})$. Let θ_1 and θ_2 be parameters in Θ such that $\theta_1 \neq \theta_2$, let $\boldsymbol{\Delta}_Z(\theta_1, \theta_2)$ be a vector in $\mathbb{R}^{\mathcal{Z}}$ with components $[\boldsymbol{\Delta}_Z(\theta_1, \theta_2)]_z = P(Z = z|\theta_1) - P(Z = z|\theta_2)$, and let $\boldsymbol{\Delta}_Y(\theta_1, \theta_2)$ be a vector in $\mathbb{R}^{\mathcal{Y}}$ with components $[\boldsymbol{\Delta}_Y(\theta_1, \theta_2)]_y = P(Y = y|\theta_1) - P(Y = y|\theta_2)$. By the assumption that $\{P(Z|\theta) \mid \theta \in \Theta\} = \mathcal{P}(\mathcal{Z})$, we have that

$$\{t\boldsymbol{\Delta}_Z(\theta_1, \theta_2) \mid t \in \mathbb{R}, \theta_1 \in \Theta, \theta_2 \in \Theta\} = \mathbf{1}_{\mathcal{Z}}^{\perp}. \tag{22}$$

On the other hand, by the assumption that $P(Y|\theta)$ is identifiable, for any $\theta_1, \theta_2 \in \Theta$ such that $\theta_1 \neq \theta_2$, it holds that

$$\boldsymbol{\Delta}_Y(\theta_1, \theta_2) = T\boldsymbol{\Delta}_Z(\theta_1, \theta_2) \neq \mathbf{0} \quad \text{and} \quad \boldsymbol{\Delta}_Z(\theta_1, \theta_2) \neq \mathbf{0}. \tag{23}$$

These equations imply that $\boldsymbol{\Delta}_Z(\theta_1, \theta_2)$ is a nonzero vector that does not belong to the kernel of T. They also imply that any vector in the kernel of T is perpendicular to $\mathbf{1}_{\mathcal{Z}}^{\perp}$. Moreover, $\mathbf{1}_{\mathcal{Z}}$ is not in the kernel of T: if that were the case, the uniform distribution of true labels would be mapped to the zero vector, which does not correspond to any weak-label distribution. Therefore, the kernel of T is $\{\mathbf{0}\}$, which means that T is left-invertible. □

This proposition suggests that the identifiability assumption is equivalent to the left-invertibility of T in cases with $\{P(Z|\theta) \mid \theta \in \Theta\} = \mathcal{P}(\mathcal{Z})$. Indeed, this is what usually happens in practice: we do not know *a priori* in which subset of $\mathcal{P}(\mathcal{Z})$ the true posterior probabilities reside, and thus, it is customary to take $\{P(Z|\theta) \mid \theta \in \Theta\}$ to be $\mathcal{P}(\mathcal{Z})$ itself. We can see from the proof above that if T is not left-invertible, then $\boldsymbol{\Delta}_Z(\theta_1, \theta_2)$ must lie outside the kernel of T for any $\theta_1, \theta_2 \in \Theta$ in order for $P(Y|\theta)$ to be identifiable. This constraint implies that $\{P(Z|\theta) \mid \theta \in \Theta\}$ has strictly lower dimensions than $\mathcal{P}(\mathcal{Z})$ does, which essentially means that we can exclude some of the labels in $\mathcal{Z}$ at the modeling step.

A.2. Non-Ambiguity Condition in Partial-Label Learning

In theoretical analyses of partial-label learning, the so-called non-ambiguity condition has been used (Cour et al., 2011; Cabannnes et al., 2020). In this section, we discuss the relation between the non-ambiguity and the left-invertibility of T.

A partial label $y \in \mathcal{Y}$ is a candidate set of labels, only one of which is correct. Obviously, $\mathcal{Y} \subset 2^{\mathcal{Z}} \setminus \{\emptyset\}$, where $2^{\mathcal{Z}}$ is the power set of $\mathcal{Z}$. The empty set $\emptyset$ is not in $\mathcal{Y}$ because a partial label always contains a correct label.

Definition 17 (Non-ambiguity condition)**.** Let $P(z' \in Y \setminus \{z\} \mid Z = z)$ be the probability that a partial label contains an incorrect label z', given a true label z. Then, the ambiguity degree ϵ is defined as follows[3]:

$$\epsilon \equiv \sup_{z, z' \in \mathcal{Z}, P(Z=z)>0} P(z' \in Y \setminus \{z\} \mid Z = z). \tag{24}$$

Partial labels are said to satisfy the non-ambiguity condition if $\epsilon < 1$.

The ambiguity degree is the maximum probability of co-occurrence of an incorrect label z' with a correct label z. To gain some intuition into the ambiguity degree and the non-ambiguity condition, let us consider two extreme cases: $\epsilon = 0$ and $\epsilon = 1$. The equality $\epsilon = 0$ implies that a weak label y is always a singleton $\{z\}$ if the correct label is z. That is, every instance is given only the correct label, and therefore, this is equivalent to supervised learning. On the other hand, when $\epsilon = 1$ and the non-ambiguity condition is not satisfied, there is a pair of labels z and z' in $\mathcal{Z}$ such that if a true label of an instance is z, an incorrect label z' is always given to that instance as well.

There is a simple example that does not satisfy the non-ambiguity condition but has a left-invertible label transition matrix T. Consider a binary classification problem $\mathcal{Z} = \{1, 2\}$ with a partial label set $\mathcal{Y} = \{\{1\}, \{1, 2\}\}$. If we identify 1 with the positive label and 2 with the negative label, this problem is often referred to as positive-unlabeled (PU) learning or learning with totally asymmetric label noise. The label transition matrix T has the following form:

$$T = \begin{pmatrix} r & 0 \\ 1 - r & 1 \end{pmatrix}, \tag{25}$$

where r $(0 < r < 1)$ is the proportion of positively labeled instances in truly positive instances. This T is left-invertible and yet breaks the non-ambiguity condition, because examples with the correct label 2 always have a partial label $\{1, 2\}$ including the incorrect label 1.

We can further show the following proposition.

Proposition 18. *Suppose that partial labels satisfy the non-ambiguity condition. If $\|\mathcal{Z}\| = 2$ or 3, then the label transition matrix T is left-invertible. On the other hand, if $\|\mathcal{Z}\| > 3$, then the label transition matrix is not necessarily left-invertible.*

Proof. We prove the case with $\|\mathcal{Z}\| \leq 3$ by proving its contraposition. Suppose that T is not left-invertible. Then, the column vectors $\boldsymbol{t}_z$ of T $(z \in \mathcal{Z})$ are not linearly independent; that is, there exists $\{a_z\}_{z \in \mathcal{Z}}$ such that

$$\sum_{z \in \mathcal{Z}} a_z \boldsymbol{t}_z = \mathbf{0} \tag{26}$$

[3] In the original paper (Cour et al., 2011), the ambiguity degree was defined with the probability conditioned on an input pattern x as well. We omit that conditioning for brevity because we assume that the distribution of weak labels does not depend on x.

and at least one of a_z is nonzero. Because $\boldsymbol{t}_z \in \mathcal{P}(\mathcal{Y})$, it follows that $\sum_{z\in\mathcal{Z}} a_z = 0$. Therefore, without loss of generality, we can assume that one of the following two equations holds:

$$\boldsymbol{t}_{z_1} = \boldsymbol{t}_{z_2}, \tag{27}$$
$$\boldsymbol{t}_{z_3} = a_1\boldsymbol{t}_{z_1} + a_2\boldsymbol{t}_{z_2} \quad (a_1 > 0, a_2 > 0). \tag{28}$$

By noting that $(\boldsymbol{t}_z)_y = P(Y = y|Z = z)$ and that $z \in y$ if $P(Y = y|Z = z) > 0$, we can see that the former implies

$$P(Y = \{z_1, z_2\}|Z = z_1) = P(Y = \{z_1, z_2\}|Z = z_2) = 1, \tag{29}$$

while the latter implies

$$P(z_3 \in Y \setminus \{z_1\}|Z = z_1) = P(z_3 \in Y \setminus \{z_2\}|Z = z_2) = 1. \tag{30}$$

In either case, we have $\epsilon = 1$, and therefore, the non-ambiguity condition is broken.

If $\|\mathcal{Z}\| = 4$, we can find an example T that is left-invertible but non-ambiguous. One such example is

$$T = \begin{pmatrix} 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \end{pmatrix}. \tag{31}$$

If $\mathcal{Z} = \{1, 2, 3, 4\}$ and T's columns from left to right correspond to 1 to 4, then the rows from top to bottom represent partial labels $(1, 3)$, $(1, 4)$, $(2, 3)$, and $(2, 4)$. We can see that this is not left-invertible by noting that a nonzero vector $(1, 1, -1, -1)^\top$ is in the kernel of T. On the other hand, the ambiguity degree ϵ is 0.5, and therefore, the non-ambiguity condition is satisfied.

We can also construct an example for an arbitrary $\|\mathcal{Z}\| > 4$ by using Eq. (31). For instance, the following block diagonal form of T is non-invertible and non-ambiguous:

$$T = \begin{pmatrix} T_4 & 0 \\ 0 & I \end{pmatrix}, \tag{32}$$

where T_4 is the label transition matrix for the first four labels given by Eq. (31), and I is an identity matrix with an appropriate size. □

A.3. Weak Noise Condition in Learning from Noisy Labels

Theoretical analyses of noisy-label learning have assumed that the noise rate is smaller than some threshold, which often coincides with the point at which T is not left-invertible (Angluin & Laird, 1988; Natarajan et al., 2013). For example, (Natarajan et al., 2013) considered label noise that has the following transition matrix:

$$T = \begin{pmatrix} 1 - r_+ & r_- \\ r_+ & 1 - r_- \end{pmatrix}. \tag{33}$$

They assumed that $r_+ + r_- < 1$, and T ceases to be left-invertible at the boundary $r_+ + r_- = 1$. However, T recovers left-invertibility for a noise rate above the threshold (e.g., complementary label learning, which can be seen as the extreme case in which labels are flipped with probability 1). There, our framework is still applicable.

B. Multiple Weak-Label Datasets

The arguments in the main text deal with scenarios with only one weak-label set $\mathcal{Y}$ and an associated transition matrix T. In this appendix, we show that without making formal changes, we can extend the formulation to scenarios with multiple samples having different noise characteristics.

Let N be the number of training sets. They all have the same true-label set $\mathcal{Z} = \{z_1, z_2, \ldots, z_C\}$ and base distribution $p(x, z)$, but each has its own weak-label set $\mathcal{Y}^{(d)} = \{y_1^{(d)}, y_2^{(d)}, \ldots, y_{C_d}^{(d)}\}$ and label transition matrix $T^{(d)}$. We show that

this problem can be mapped to a problem with a single weak-label set $\mathcal{Y} = \cup_{d=1}^{N} \mathcal{Y}^{(d)}$ and a label transition matrix. A partial risk on the dth set is defined as $R_d[q(z|x)] \equiv \mathbb{E}_{(x,y)\sim p^{(d)}(x,y)}[l_{\mathrm{W}}(q(z|x), y)]$, where $p^{(d)}(x, y) \equiv \sum_{z\in\mathcal{Z}} T^{(d)}_{yz} p(x, z)$. The total risk is defined as a convex combination of the partial risks:

$$R[q(z|x)] \equiv \sum_{d=1}^{N} \alpha_d R_d[q(z|x)] \tag{34}$$

$$\equiv \mathbb{E}_{(x,y)\sim\sum_{z\in\mathcal{Z}} T_{yz} p(x,z)}[l_{\mathrm{W}}(q(z|x), y)], \tag{35}$$

where the coefficients α_d are positive real numbers satisfying $\sum_{d=1}^{N} \alpha_d = 1$, and the total label transition matrix from $\mathcal{Z}$ to $\mathcal{Y}$ is defined as $T = (\alpha_1 T^{(1)\mathsf{T}}, \alpha_2 T^{(2)\mathsf{T}}, \ldots, \alpha_N T^{(N)\mathsf{T}})^{\mathsf{T}}$. In fact, the α_d may be absorbed in a weak-label loss, and we may simply set $\alpha_d = 1/N$ for all d. The equality $T^{\mathsf{T}}\mathbf{1}_{\mathcal{Y}} = \mathbf{1}_{\mathcal{Z}}$ can be verified by using $T^{(d)\mathsf{T}}\mathbf{1}_{\mathcal{Y}^{(d)}} = \mathbf{1}_{\mathcal{Z}}$ for all $d = 1, 2, \ldots, N$, and therefore, T is formally qualified as a transition matrix.

As in the discussion in the main text, we assume that T is left-invertible. This assumption is weaker than requiring that all of the $T^{(d)}$ be left-invertible. By using this T as a label transition matrix, we can formally treat the multiple-source scenario exactly the same as the single-source case. In the training phase, we need to calculate the empirical risk. This can be done by first calculating the empirical partial risks from respective training sets with a partial-label set $\mathcal{Y}^{(d)}$ and then aggregating the results.

Example 19. Consider three-class classification from two weakly labeled datasets. One set is labeled by an annotator who distinguishes Class 1 from the other classes, and the other set, by another annotator who distinguishes Class 2 from the other classes. Such a scenario is represented by the following transition matrices:

$$T^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad T^{(2)} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}^{\mathsf{T}}. \tag{36}$$

Here, T is reconstructible, while $T^{(1)}$ and $T^{(2)}$ are not. An example of R is

$$R = \begin{pmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix}. \tag{37}$$

C. Proofs Omitted in Main Text

C.1. Theorem 5

We first prove the following lemma, which relates Condition 1 of the theorem to the finiteness of the convex conjugate of $F(\boldsymbol{v})$.

Lemma 20. *Let $F : C \subset \mathbf{1}_{\mathcal{Z}}^{\perp} \to \mathbb{R}$ be a closed convex function. Then its convex conjugate $F^*(\boldsymbol{q})$ is finite for all $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$ if and only if $\sup_{\boldsymbol{v}\in C} [\max_{z\in\mathcal{Z}} v_z - F(\boldsymbol{v})] < \infty$.*

Proof. Without loss of generality, the condition that $F^*(\boldsymbol{q}) < \infty$ for all $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$ can be replaced with the finiteness at $\boldsymbol{q} = \boldsymbol{e}_y$ for all $y \in \mathcal{Y}$, where $\boldsymbol{e}_y \in \mathcal{P}(\mathcal{Z})$ is a standard unit vector. This is because of Jensen's inequality and the fact that $\mathcal{P}(\mathcal{Z})$ is the convex hull of a set of the standard unit vectors. Then, the lemma can be seen as a special case of Proposition 14 with $\mathcal{Y} = \mathcal{Z}$ and $R = I_{\mathcal{Z}}$. □

Proof of Theorem 5. Suppose that l is a regular proper loss. From Theorem 4, there exists a closed convex function $S : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ and its subgradient function $\underline{\nabla} S : \mathcal{P}(\mathcal{Z}) \to \overline{\mathbb{R}}^{\mathcal{Z}}$ such that

$$l(\boldsymbol{q}, z) = -[\underline{\nabla} S(\boldsymbol{q})]_z + \langle \boldsymbol{q}, \underline{\nabla} S(\boldsymbol{q})\rangle - S(\boldsymbol{q}). \tag{38}$$

Let $S^* : \tilde{\mathcal{C}} \to \mathbb{R}$ be the convex conjugate of S. From the Savage representation and the identity $S^*(\underline{\nabla} S(\boldsymbol{q})) + S(\boldsymbol{q}) = \langle \boldsymbol{q}, \underline{\nabla} S(\boldsymbol{q})\rangle$, which follows from the equality condition of the Fenchel-Young inequality, we have $l(\boldsymbol{q}, z) = \lambda_{S^*}(\underline{\nabla} S(\boldsymbol{q}), z)$.

Now we need to show that the restriction of S^* to $\mathcal{C} \equiv \tilde{\mathcal{C}} \cap \mathbf{1}_{\mathcal{Z}}^{\perp}$, denoted as $F : \mathcal{C} \to \mathbb{R}$, satisfies the two conditions of the theorem. Suppose that $\boldsymbol{v} = v_{\parallel}\mathbf{1}_{\mathcal{Z}} + \boldsymbol{v}_{\perp}$, where $v_{\parallel} \in \mathbb{R}$ and $\boldsymbol{v}_{\perp} \in \mathbf{1}_{\mathcal{Z}}^{\perp}$. Because $\langle \boldsymbol{q}, \boldsymbol{v} \rangle = v_{\parallel} + \langle \boldsymbol{q}, \boldsymbol{v}_{\perp} \rangle$ for $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$, it holds that for all $\boldsymbol{v} \in \tilde{\mathcal{C}}$,

$$S^*(\boldsymbol{v}) \equiv \sup_{\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})} [\langle \boldsymbol{q}, \boldsymbol{v} \rangle - S(\boldsymbol{q})] \tag{39}$$

$$= v_{\parallel} + F(\boldsymbol{v}_{\perp}). \tag{40}$$

This implies that

$$\langle \boldsymbol{q}, \boldsymbol{v} \rangle - S^*(\boldsymbol{v}) = \langle \boldsymbol{q}, \boldsymbol{v}_{\perp} \rangle - F(\boldsymbol{v}_{\perp}) \tag{41}$$

for all $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$ and $\boldsymbol{v} \in \tilde{\mathcal{C}}$. By taking the supremum of this equality over $\boldsymbol{v} \in \tilde{\mathcal{C}}$, we conclude that $F^*(\boldsymbol{q}) = S(\boldsymbol{q})$ for all $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$, where F^* is the convex conjugate of F. Because $S(\boldsymbol{q})$ is finite for all $\boldsymbol{q} \in \mathcal{P}(\mathcal{Z})$, $F^*(\boldsymbol{q})$ is also finite in $\mathcal{P}(\mathcal{Z})$. By Lemma 20, this implies Condition 1 of the theorem.

To show Condition 2, we need to relate the subgradients of $S(\boldsymbol{q})$ with those of $F^*(\boldsymbol{q})$. We first note that they have the same projections of the subgradients onto $\mathbf{1}_{\mathcal{Z}}^{\perp}$, because $F^*(\boldsymbol{q}) = S(\boldsymbol{q})$ in $\mathcal{P}(\mathcal{Z})$. Regarding the component of the subgradients that is parallel to $\mathbf{1}_{\mathcal{Z}}$, it holds that $\langle \mathbf{1}_{\mathcal{Z}}, \underline{\nabla} F^*(\boldsymbol{q}) \rangle = 0$ because $F^*(\boldsymbol{q})$ is independent of $\langle \boldsymbol{q}, \mathbf{1}_{\mathcal{Z}} \rangle$; that is, for all $t \in \mathbb{R}$,

$$F^*(\boldsymbol{q} + t\mathbf{1}_{\mathcal{Z}}) = \sup_{\boldsymbol{v} \in \mathcal{C}} [\langle \boldsymbol{q} + t\mathbf{1}_{\mathcal{Z}}, \boldsymbol{v} \rangle - F(\boldsymbol{v})] \tag{42}$$

$$= \sup_{\boldsymbol{v} \in \mathcal{C}} [\langle \boldsymbol{q}, \boldsymbol{v} \rangle - F(\boldsymbol{v})] \tag{43}$$

$$= F^*(\boldsymbol{q}). \tag{44}$$

On the other hand, because S is defined on $\mathcal{P}(\mathcal{Z})$, it holds that $\boldsymbol{v} + t\mathbf{1}_{\mathcal{Z}} \in \partial S(\boldsymbol{q})$ for all $\boldsymbol{v} \in \partial S(\boldsymbol{q})$ and $t \in \mathbb{R}$. The choice of this t does not affect the loss function's value. Therefore, we can always choose $\underline{\nabla} S(\boldsymbol{q})$ such that $\underline{\nabla} S(\boldsymbol{q}) = \underline{\nabla} F^*(\boldsymbol{q})$, which implies Condition 2.

Conversely, suppose that there exists a closed convex function $F : C \subset \mathbf{1}_{\mathcal{Z}}^{\perp} \to \mathbb{R}$ that satisfies the two conditions. Its convex conjugate F^* is finite at all $\boldsymbol{p} \in \mathcal{P}(\mathcal{Z})$ by Lemma 20. Let $S : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be a restriction of F^* on $\mathcal{P}(\mathcal{Z})$. In general, a subdifferential of a function is not larger as a set than a subdifferential of its restriction; that is, $\partial F^*(\boldsymbol{q}) \subset \partial S(\boldsymbol{q})$. This implies that $\underline{\nabla} F^*(\boldsymbol{q})$ can be seen as a subgradient function of S. From this fact and the identity $F(\underline{\nabla} F^*(\boldsymbol{q})) + F^*(\boldsymbol{q}) = \langle \boldsymbol{q}, \underline{\nabla} F^*(\boldsymbol{q}) \rangle$, which follows from the equality condition of the Fenchel-Young inequality (Rockafellar, 1996), the loss $l(\boldsymbol{q}, z) = \lambda_F(\underline{\nabla} F^*(\boldsymbol{q}), z)$ conforms to the Savage representation and is proper. □

C.2. Proposition 6

Let $\boldsymbol{v}_0$ be a minimizer of $\mathbb{E}_{z \sim \boldsymbol{p}} [\lambda_F(\boldsymbol{v}, z)]$. Then it holds that

$$-\langle \boldsymbol{v}_0, \boldsymbol{p} \rangle + F(\boldsymbol{v}_0) = \min_{\boldsymbol{v} \in \mathcal{C}} [-\langle \boldsymbol{v}, \boldsymbol{p} \rangle + F(\boldsymbol{v})] \tag{45}$$

$$= -F^*(\boldsymbol{p}), \tag{46}$$

where $\boldsymbol{v}_0 \in \partial F^*(\boldsymbol{p})$. This proves the claim because $\boldsymbol{p}$ is always a member of $\mathcal{P}(\mathcal{Z})$.

C.3. Theorem 7

Our proof of this theorem relies on the following lemma that gives a general relation between T-properness and properness (Cid-sueiro, 2012).

Lemma 21. *A weak-label loss $l_{\mathrm{W}} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ is (strictly) T-proper if and only if a loss function $l : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \overline{\mathbb{R}}$, defined as $l(\boldsymbol{q}, z) = \sum_{y \in \mathcal{Y}} T_{yz} l_{\mathrm{W}}(\boldsymbol{q}, y)$, is (strictly) proper.*

This lemma follows from the identity $\mathbb{E}_{y \sim T\boldsymbol{p}}[l_{\mathrm{W}}(\boldsymbol{q}, y)] = \mathbb{E}_{z \sim \boldsymbol{p}}[\sum_{y \in \mathcal{Y}} T_{yz} l_{\mathrm{W}}(\boldsymbol{q}, y)]$. The left-hand (right-hand) side is minimized by $\boldsymbol{q} = \boldsymbol{p}$ if and only if the loss in the expected value is T-proper (proper). The lemma indicates that having corrupted labels and a weak-label loss is equivalent to having clean labels and a mixed weak-label loss as a supervised-learning loss.

Proof of Theorem 7. From Lemma 21, the weak-label loss l_{W} is T-proper if and only if a loss defined as $l(\boldsymbol{q}, z) \equiv \sum_{y\in\mathcal{Y}} T_{yz} l_{\mathrm{W}}(\boldsymbol{q}, y)$ is proper. Then, from Theorem 5, there exists a closed convex function F defined on a subset of $\mathbf{1}_{\mathcal{Z}}^{\perp}$ that satisfies Condition 1 of the theorem and the following equation:

$$-[\underline{\nabla} F^*(\boldsymbol{q})]_z + F(\underline{\nabla} F^*(\boldsymbol{q})) = \sum_{y\in\mathcal{Y}} T_{yz} l_{\mathrm{W}}(\boldsymbol{q}, y), \tag{47}$$

where F^* is the convex conjugate of F and $\underline{\nabla} F^*(\boldsymbol{q})$ is a subgradient of F^* at a point $\boldsymbol{q}$. By using the identity $T^{\mathsf{T}}\mathbf{1}_{\mathcal{Y}} = \mathbf{1}_{\mathcal{Z}}$, we find that

$$-[\underline{\nabla} F^*(\boldsymbol{q})]_z = \sum_{y\in\mathcal{Y}} T_{yz}[l_{\mathrm{W}}(\boldsymbol{q}, y) - F(\underline{\nabla} F^*(\boldsymbol{q}))]. \tag{48}$$

Note that the right-hand side is a product of a matrix T^{T} and a vector in $\mathbb{R}^{\mathcal{Y}}$, whose yth component is $[l_{\mathrm{W}}(\boldsymbol{q}, y) - F(\underline{\nabla} F^*(\boldsymbol{q}))]$. By the left-invertibility of T, we can invert this equation up to possibly nonzero $\boldsymbol{\Delta}(\boldsymbol{q}) \in \operatorname{coker} T$. □

C.4. Lemma 11

Let T be a label transition matrix corresponding to a reconstruction matrix R. Then, $\boldsymbol{v} = T^{\mathsf{T}}(R^{\mathsf{T}}\boldsymbol{v})$ for $\boldsymbol{v} \in \mathbf{1}_{\mathcal{Z}}^{\perp}$. Because all the elements of T are nonnegative and $T^{\mathsf{T}}\mathbf{1}_{\mathcal{Y}} = \mathbf{1}_{\mathcal{Z}}$, a component of $\boldsymbol{v}$ is a convex combination of $(R^{\mathsf{T}}\boldsymbol{v})_y$. Therefore, $v_z \le \max_{y\in\mathcal{Y}}(R^{\mathsf{T}}\boldsymbol{v})_y$ for all $z \in \mathcal{Z}$.

C.5. Proposition 14

Suppose that $F^*(R\boldsymbol{e}_y) < \infty$ for all $y \in \mathcal{Y}$. By the definition of the convex conjugate, it follows that

$$\begin{aligned} F^*(R\boldsymbol{e}_y) &\equiv \sup_{\boldsymbol{v}\in\mathcal{C}}[\langle R\boldsymbol{e}_y, \boldsymbol{v}\rangle - F(\boldsymbol{v})] && (49)\\ &= \sup_{\boldsymbol{v}\in\mathcal{C}}[\langle \boldsymbol{e}_y, R^{\mathsf{T}}\boldsymbol{v}\rangle - F(\boldsymbol{v})] && (50)\\ &= \sup_{\boldsymbol{v}\in\mathcal{C}}[(R^{\mathsf{T}}\boldsymbol{v})_y - F(\boldsymbol{v})] && (51)\\ &< \infty. && (52) \end{aligned}$$

By maximizing both sides over $y \in \mathcal{Y}$, we obtain

$$\begin{aligned} \max_{y\in\mathcal{Y}} F^*(R\boldsymbol{e}_y) &= \max_{y\in\mathcal{Y}} \sup_{\boldsymbol{v}\in\mathcal{C}} \left[(R^{\mathsf{T}}\boldsymbol{v})_y - F(\boldsymbol{v})\right] && (53)\\ &= \sup_{\boldsymbol{v}\in\mathcal{C}} \left[\max_{y\in\mathcal{Y}}(R^{\mathsf{T}}\boldsymbol{v})_y - F(\boldsymbol{v})\right] && (54)\\ &< \infty. && (55) \end{aligned}$$

Conversely, suppose that $\sup_{\boldsymbol{v}\in C}\left[\max_{y\in\mathcal{Y}}(R^{\mathsf{T}}\boldsymbol{v})_y - F(\boldsymbol{v})\right] < \infty$. Because $\langle \boldsymbol{q}, R^{\mathsf{T}}\boldsymbol{v}\rangle \le \max_{y\in\mathcal{Y}}(R^{\mathsf{T}}\boldsymbol{v})_y$ for all $\boldsymbol{q} \in \mathcal{P}(\mathcal{Y})$, it follows that

$$\begin{aligned} F^*(R\boldsymbol{q}) &\equiv \sup_{\boldsymbol{v}\in C} [\langle R\boldsymbol{q}, \boldsymbol{v}\rangle - F(\boldsymbol{v})] && (56)\\ &\equiv \sup_{\boldsymbol{v}\in C} \left[\langle \boldsymbol{q}, R^{\mathsf{T}}\boldsymbol{v}\rangle - F(\boldsymbol{v})\right] && (57)\\ &\le \sup_{\boldsymbol{v}\in C} \left[\max_{y\in\mathcal{Y}}(R^{\mathsf{T}}\boldsymbol{v})_y - F(\boldsymbol{v})\right] && (58)\\ &< \infty. && (59) \end{aligned}$$

The proposition follows as the special case with $\boldsymbol{q} = \boldsymbol{e}_y$.

D. Condition for Strict Properness

In this appendix, we prove the conditions for a dual representation to give a strictly proper loss.

As already stated in Theorem 4, a proper loss is strictly proper if and only if the associated negative Bayes risk $S(\boldsymbol{p})$ is strictly convex. To dualize this condition, we introduce some notations. Let $\mathcal{Z}' = \{z'_1, z'_2, \ldots, z'_{K'}\}$ be a subset of the label set $\mathcal{Z} = \{z_1, z_2, \ldots, z_K\}$. Then, a mapping $\pi_{\mathcal{Z}'} : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}^{\mathcal{Z}'}$ denotes a natural projection from $\mathbb{R}^{\mathcal{Z}}$ onto $\mathbb{R}^{\mathcal{Z}'}$, and $\rho_{\mathcal{Z}'} : \mathbb{R}^{\mathcal{Z}'} \to \mathbf{1}^{\perp}_{\mathcal{Z}'}$ denotes an orthogonal projection from $\mathbb{R}^{\mathcal{Z}'}$ onto its subspace $\mathbf{1}^{\perp}_{\mathcal{Z}'}$. We also need a "projection onto the bottom," $\sigma_{\mathcal{Z}'} : \mathbf{1}^{\perp}_{\mathcal{Z}'} \to \mathbb{R}^{\mathcal{Z}'-\{z'_1\}}$, which is a natural projection from $\mathbf{1}^{\perp}_{\mathcal{Z}'}$, as a subspace of $\mathbb{R}^{\mathcal{Z}'}$, onto $\mathbb{R}^{\mathcal{Z}'-\{z'_1\}}$. Let $F : \mathcal{C} \to \mathbb{R}$ be a function whose domain is a convex subset $\mathcal{C}$ of $\mathbf{1}^{\perp}_{\mathcal{Z}}$. Then, we define a function $F_{\mathcal{Z}'} : \mathcal{C}_{\mathcal{Z}'} \to \mathbb{R}$ as the closure of

$$\tilde{F}_{\mathcal{Z}'}(\boldsymbol{v}) \equiv \inf_{\boldsymbol{v}' \in \mathcal{C} \cap \pi^{-1}_{\mathcal{Z}'} \circ \rho^{-1}_{\mathcal{Z}'}(\boldsymbol{v})} \left[F(\boldsymbol{v}') - \langle \frac{1}{|\mathcal{Z}'|} \mathbf{1}_{\mathcal{Z}'}, \pi_{\mathcal{Z}'}(\boldsymbol{v}') \rangle \right], \tag{60}$$

where $\mathcal{C}_{\mathcal{Z}'} \equiv \rho_{\mathcal{Z}'} \circ \pi_{\mathcal{Z}'}(\mathcal{C})$.

Theorem 22. *A proper loss in the dual representation associated with a closed convex function F is strictly proper if and only if for all subsets $\mathcal{Z}'$ of $\mathcal{Z}$, a function $F_{\mathcal{Z}'} \circ \sigma^{-1}_{\mathcal{Z}'}$ is differentiable on some subset $\mathcal{D}_{\mathcal{Z}'}$ of $\sigma_{\mathcal{Z}'}(\mathcal{C}_{\mathcal{Z}'})$ and the range of $\partial F_{\mathcal{Z}'}$ on $\sigma^{-1}_{\mathcal{Z}'}(\mathcal{D}_{\mathcal{Z}'})$ contains the relative interior of $\mathcal{P}(\mathcal{Z}')$.*

D.1. Proof

Discussing strict convexity on a closed set $\mathcal{P}(\mathcal{Z})$ is not straightforward. Instead, the following lemma allows us to decompose it into strict convexity on open subsets of $\mathcal{P}(\mathcal{Z})$.

Lemma 23. *A function S is strictly convex on $\mathcal{P}(\mathcal{Z})$ if and only if it is strictly convex on a convex, relatively open subset $P_{\mathcal{Z}}(\mathcal{Z}') \equiv \{\boldsymbol{p} \in \mathcal{P}(\mathcal{Z}) | p_z \neq 0 \ (z \in \mathcal{Z}'), p_z = 0 \ (z \notin \mathcal{Z}')\}$ for any $\mathcal{Z}' \subset \mathcal{Z}$.*

Proof. Suppose that S is strictly convex on $\mathcal{P}(\mathcal{Z})$. Clearly, it is strictly convex on any convex subset of $\mathcal{P}(\mathcal{Z})$. Conversely, suppose that S is strictly convex on $P_{\mathcal{Z}}(\mathcal{Z}')$ for any subset $\mathcal{Z}'$ of $\mathcal{Z}$. Let $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ be two different elements of $\mathcal{P}(\mathcal{Z})$, let l be a line segment connecting them, and let λ, λ_1, and λ_2 be real numbers such that $0 < \lambda_2 < \lambda < \lambda_1 < 1$. Also, define $\boldsymbol{p} = \lambda \boldsymbol{p}_1 + (1-\lambda)\boldsymbol{p}_2$ and $\boldsymbol{p}'_i = \lambda_i \boldsymbol{p}_1 + (1-\lambda_i)\boldsymbol{p}_2$ $(i = 1, 2)$. We can see that $\boldsymbol{p}$ is also a convex combination of $\boldsymbol{p}'_1$ and $\boldsymbol{p}'_2$:

$$\boldsymbol{p} = \lambda' \boldsymbol{p}'_1 + (1-\lambda')\boldsymbol{p}'_2, \quad \text{where} \quad \lambda' = \frac{\lambda_2 - \lambda}{\lambda_1 - \lambda_2}. \tag{61}$$

Because the relative interior of l is contained in only one of the $P_{\mathcal{Z}}(\mathcal{Z}')$, $\boldsymbol{p}$, $\boldsymbol{p}'_1$, and $\boldsymbol{p}'_2$ are all contained in that $P_{\mathcal{Z}}(\mathcal{Z}')$. Therefore, by assumption,

$$\begin{aligned} S(\boldsymbol{p}) &< \lambda' S(\boldsymbol{p}'_1) + (1-\lambda') S(\boldsymbol{p}'_2) && (62)\\ &\leq \lambda'[\lambda_1 S(\boldsymbol{p}_1) + (1-\lambda_1) S(\boldsymbol{p}_2)] + (1-\lambda')[\lambda_2 S(\boldsymbol{p}_1) + (1-\lambda_2) S(\boldsymbol{p}_2)] && (63)\\ &= \lambda S(\boldsymbol{p}_1) + (1-\lambda) S(\boldsymbol{p}_2). && (64) \end{aligned}$$

As this holds true for any $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in $\mathcal{P}(\mathcal{Z})$ and $\lambda \in (0, 1)$, $S(\boldsymbol{p})$ is strictly convex on $\mathcal{P}(\mathcal{Z})$. □

We now focus on a single $\mathcal{Z}' \subset \mathcal{Z}$ and S restricted on $P_{\mathcal{Z}}(\mathcal{Z}')$. Because $\langle \boldsymbol{p}, \boldsymbol{v} \rangle = \langle \pi_{\mathcal{Z}'}(\boldsymbol{p}), \pi_{\mathcal{Z}'}(\boldsymbol{v}) \rangle$ for any $\boldsymbol{p} \in P_{\mathcal{Z}}(\mathcal{Z}')$ and $\boldsymbol{v} \in \mathcal{C}$, we can define a function $S_{\mathcal{Z}'}$ as

$$S_{\mathcal{Z}'}(\boldsymbol{p}) \equiv \sup_{\boldsymbol{v} \in C} [\langle \boldsymbol{p}, \pi_{\mathcal{Z}'}(\boldsymbol{v}) \rangle - F(\boldsymbol{v})] \tag{65}$$

for $\boldsymbol{p} \in \pi_{\mathcal{Z}'}(P_{\mathcal{Z}}(\mathcal{Z}'))$, and this function is equal to $S(\boldsymbol{p})$ in $\pi_{\mathcal{Z}'}(P_{\mathcal{Z}}(\mathcal{Z}'))$. Clearly, $S_{\mathcal{Z}'}$ is strictly convex if and only if S is strictly convex on $P_{\mathcal{Z}}(\mathcal{Z}')$. Note that $\pi_{\mathcal{Z}'}(P_{\mathcal{Z}}(\mathcal{Z}'))$ is the relative interior of $\mathcal{P}(\mathcal{Z}')$, and that the above definition is applicable to points within the relative boundary of $P_{\mathcal{Z}}(\mathcal{Z}')$. Therefore, $S_{\mathcal{Z}'}$ can be extended to a function on $\mathcal{P}(\mathcal{Z}')$.

Lemma 24. *A convex function f is strictly convex on a convex, relatively open subset $\mathcal{C}$ of its domain if and only if $\partial f(\boldsymbol{p}_1) \cap \partial f(\boldsymbol{p}_2) = \emptyset$ for any pair of two different points $\boldsymbol{p}_1, \boldsymbol{p}_2$ in $\mathcal{C}$.*[4]

[4]The condition that $\mathcal{C}$ is relatively open can be removed, in which case $\partial f(\boldsymbol{p})$ can be empty for some $\boldsymbol{p}$. See, for example, Theorem 26.3 in Rockafellar (1996).

Proof. Suppose that f is not strictly convex on $\mathcal{C}$. Then, there exist $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \mathcal{C}$ and $\lambda \in (0,1)$ such that $f(\lambda \boldsymbol{p}_1 + (1-\lambda)\boldsymbol{p}_2) = \lambda f(\boldsymbol{p}_1) + (1-\lambda) f(\boldsymbol{p}_2)$. Take $\boldsymbol{v} \in \partial f(\boldsymbol{p})$, and let H be a graph of an affine function $h(\boldsymbol{q}) \equiv f(\boldsymbol{p}) + \langle \boldsymbol{q} - \boldsymbol{p}, \boldsymbol{v} \rangle$. Then, H is a supporting hyperplane of the epigraph of f at $(\boldsymbol{p}, f(\boldsymbol{p}))$. Because $(\boldsymbol{p}, f(\boldsymbol{p}))$ belongs to the relative interior of the line segment joining $(\boldsymbol{p}_1, f(\boldsymbol{p}_1))$ and $(\boldsymbol{p}_2, f(\boldsymbol{p}_2))$, these two points also lie in H. Therefore, $\boldsymbol{v} \in \partial f(\boldsymbol{p}_1)$ and $\boldsymbol{v} \in \partial f(\boldsymbol{p}_2)$, which implies that $\partial f(\boldsymbol{p}_1) \cap \partial f(\boldsymbol{p}_2) \neq \emptyset$.

Conversely, suppose that there exist two different points $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in $\mathcal{C}$ such that $\partial f(\boldsymbol{p}_1) \cap \partial f(\boldsymbol{p}_2) \neq \emptyset$. Let $\boldsymbol{v}$ be an element of $\partial f(\boldsymbol{p}_1) \cap \partial f(\boldsymbol{p}_2)$. Then for a certain constant k, a graph H of an affine function $h(\boldsymbol{q}) = \langle \boldsymbol{q} - \boldsymbol{p}, \boldsymbol{v} \rangle + k$ is a supporting hyperplane of the epigraph of f and contains $(\boldsymbol{p}_1, f(\boldsymbol{p}_1))$ and $(\boldsymbol{p}_2, f(\boldsymbol{p}_2))$. This implies that H contains the line segment joining $(\boldsymbol{p}_1, f(\boldsymbol{p}_1))$ and $(\boldsymbol{p}_2, f(\boldsymbol{p}_2))$. Thus, f cannot be strictly convex along the line segment connecting $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$. □

By applying this lemma to $S_{\mathcal{Z}'}$, its strict convexity becomes equivalent to the injectivity of the subdifferential map $\partial S_{\mathcal{Z}'}$. On the other hand, the inverse of a subdifferential map of a closed convex function is the subdifferential map of its conjugate function (Corollary 23.5.1 in Rockafellar (1996)). Indeed, it holds that

$$S_{\mathcal{Z}'}(\boldsymbol{p}) = \sup_{\boldsymbol{v} \in C} \left[\langle \boldsymbol{p}, \sigma_{\mathcal{Z}'} \circ \pi_{\mathcal{Z}'}(\boldsymbol{v}) \rangle + \langle \frac{1}{|\mathcal{Z}'|} \mathbf{1}_{\mathcal{Z}'}, \pi_{\mathcal{Z}'}(\boldsymbol{v}) \rangle - F(\boldsymbol{v}) \right] \tag{66}$$

$$= \sup_{\boldsymbol{v}' \in \sigma_{\mathcal{Z}'} \circ \pi_{\mathcal{Z}'}(C)} \left\{ \sup_{\boldsymbol{v} \in C \cap \pi_{\mathcal{Z}'}^{-1} \circ \sigma_{\mathcal{Z}'}^{-1}(\boldsymbol{v}')} \left[\langle \boldsymbol{p}, \boldsymbol{v}' \rangle + \langle \frac{1}{|\mathcal{Z}'|} \mathbf{1}_{\mathcal{Z}'}, \pi_{\mathcal{Z}'}(\boldsymbol{v}) \rangle - F(\boldsymbol{v}) \right] \right\} \tag{67}$$

$$= \sup_{\boldsymbol{v}' \in \sigma_{\mathcal{Z}'} \circ \pi_{\mathcal{Z}'}(C)} \left[\langle \boldsymbol{p}, \boldsymbol{v}' \rangle - F_{\mathcal{Z}'}(\boldsymbol{v}') \right], \tag{68}$$

and therefore, that $\partial S_{\mathcal{Z}'} = (\partial F_{\mathcal{Z}'})^{-1}$. This implies that the necessary and sufficient condition for $\partial S_{\mathcal{Z}'}(\boldsymbol{p}_1) \cap \partial S_{\mathcal{Z}'}(\boldsymbol{p}_2) = \emptyset$ for $\boldsymbol{p}_1 \neq \boldsymbol{p}_2$ is that $\{\boldsymbol{v}' | \{\boldsymbol{p}_1, \boldsymbol{p}_2\} \subset \partial F_{\mathcal{Z}'}(\boldsymbol{v}')\} = \emptyset$.

Lemma 25. *Suppose that $\boldsymbol{p} \in \partial F_{\mathcal{Z}'}(\boldsymbol{v})$ for some $\boldsymbol{p} \in \operatorname{int} \mathcal{P}(\mathcal{Z}')$ and $\boldsymbol{v}$ in the domain of $\partial F_{\mathcal{Z}'}$. Then, no other point in $\operatorname{int} \mathcal{P}(\mathcal{Z}')$ belongs to $\partial F_{\mathcal{Z}'}(\boldsymbol{v})$ if and only if $F_{\mathcal{Z}'} \circ \sigma_{\mathcal{Z}'}^{-1}$ is differentiable at $\sigma_{\mathcal{Z}'}(\boldsymbol{v})$.*

Proof. Let $\boldsymbol{v}_0$ be $\sigma_{\mathcal{Z}'}(\boldsymbol{v})$ and $\boldsymbol{p}_0$ be $\sigma_{\mathcal{Z}'} \circ \rho_{\mathcal{Z}'}(\boldsymbol{p})$. We can verify by direct calculation that $\partial(F_{\mathcal{Z}'} \circ \sigma_{\mathcal{Z}'}^{-1})(\boldsymbol{v}_0) = \sigma_{\mathcal{Z}'} \circ \rho_{\mathcal{Z}'}(\partial F_{\mathcal{Z}'}(\boldsymbol{v}))$. In addition, because $F_{\mathcal{Z}'}$ is defined on (a subset of) $\mathbf{1}_{\mathcal{Z}'}^{\perp}$, $\boldsymbol{q} + t\mathbf{1}_{\mathcal{Z}'} \in \partial F_{\mathcal{Z}'}(\boldsymbol{v})$ for any $\boldsymbol{q} \in \partial F_{\mathcal{Z}'}(\boldsymbol{v})$ and $t \in \mathbb{R}$. Therefore, $\{\boldsymbol{p}_0\} \in \partial(F_{\mathcal{Z}'} \circ \sigma_{\mathcal{Z}'}^{-1})(\boldsymbol{v}_0)$ is equivalent to $\partial F_{\mathcal{Z}'}(\boldsymbol{v}) = \{\boldsymbol{p}_0 + t\mathbf{1}_{\mathcal{Z}'} | t \in \mathbb{R}\}$, in which case $\partial F_{\mathcal{Z}'}(\boldsymbol{v})$ contains one and only one element of $\operatorname{int} \mathcal{P}(\mathcal{Z}')$. This implies the lemma because a convex function $F_{\mathcal{Z}'} \circ \sigma_{\mathcal{Z}'}^{-1}$ is differentiable at a point $\boldsymbol{v}_0$ if and only if it has a unique subgradient there. □

Finally, by combining these lemmas, we obtain Theorem 22.

E. Forward-Correction Loss

In this appendix, we verify that a forward-corrected loss l_{W} conforms to Theorem 7. 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 $l_{\mathrm{W}}(\boldsymbol{q}, y) = l_{\mathcal{Y}}(T\boldsymbol{q}, y)$, where $l_{\mathcal{Y}} : \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ is a proper loss for estimating weak-label posterior probabilities.

We first apply Theorem 5 to $l_{\mathcal{Y}}$ and find that $l_{\mathcal{Y}}(\boldsymbol{q}, y) = -[\underline{\nabla} F_{\mathcal{Y}}^*(\boldsymbol{q})]_y + F_{\mathcal{Y}}(\underline{\nabla} F_{\mathcal{Y}}^*(\boldsymbol{q}))$ for $\boldsymbol{q} \in \mathcal{P}(\mathcal{Y})$; here, $F_{\mathcal{Y}}^*(\boldsymbol{q})$ is the negative Bayes risk corresponding to $l_{\mathcal{Y}}$, and $F_{\mathcal{Y}}(\boldsymbol{v})$ is its convex conjugate. We also have the negative Bayes risk $S(\boldsymbol{q})$ and its conjugate $S^*(\boldsymbol{v})$ for the weak-label loss l_{W}. A key identity among these quantities is $S(\boldsymbol{q}) = F_{\mathcal{Y}}^*(T\boldsymbol{q})$, which further implies that $\underline{\nabla} S(\boldsymbol{q}) = T^{\mathsf{T}} \underline{\nabla} F_{\mathcal{Y}}^*(T\boldsymbol{q})$. The latter can be inverted to find that $\underline{\nabla} F_{\mathcal{Y}}^*(T\boldsymbol{q}) = R^{\mathsf{T}} \underline{\nabla} S(\boldsymbol{q}) - \boldsymbol{\Delta}(\boldsymbol{q})$ for some function $\boldsymbol{\Delta}(\boldsymbol{q})$ that takes values on $\operatorname{coker} T$. It also holds that

$$F_{\mathcal{Y}}(\underline{\nabla} F_{\mathcal{Y}}^*(T\boldsymbol{q})) + F_{\mathcal{Y}}^*(T\boldsymbol{q}) = \langle T\boldsymbol{q}, \underline{\nabla} F_{\mathcal{Y}}^*(T\boldsymbol{q}) \rangle \tag{69}$$

$$= \langle \boldsymbol{q}, \underline{\nabla} S(\boldsymbol{q}) \rangle = S^*(\underline{\nabla} S(\boldsymbol{q})) + S(\boldsymbol{q}), \tag{70}$$

where the first and third equalities follow from the equality condition of the Fenchel-Young inequality (Rockafellar, 1996). By using $S(\boldsymbol{q}) = F_{\mathcal{Y}}^*(T\boldsymbol{q})$ in Eq. (70), we find that $F_{\mathcal{Y}}(\underline{\nabla} F_{\mathcal{Y}}^*(T\boldsymbol{q})) = S^*(\underline{\nabla} S(\boldsymbol{q}))$. Therefore, we confirm that

$$l_{\mathrm{W}}(\boldsymbol{q}, y) = l_{\mathcal{Y}}(T\boldsymbol{q}, y) \tag{71}$$

$$= -[R^{\mathsf{T}} \underline{\nabla} S(\boldsymbol{q})]_y + S^*(\underline{\nabla} S(\boldsymbol{q})) + \Delta_y(\boldsymbol{q}), \tag{72}$$

Table 3. Numbers of examples in the dataset splits. "Training (original)" refers to those that are originally defined as training splits in the datasets, while "Training (used)" indicates those that were actually used in training.

Dataset	Training (original)	Training (used)	Validation	Test
MNIST	60,000	54,000	6,000	10,000
CIFAR-10	50,000	45,000	5,000	10,000

Table 4. Initial learning rates with which the best validation accuracy was achieved for each setting.

	Weight decay	MNIST, linear	MNIST, MLP	CIFAR-10, ResNet-20	CIFAR-10, WRN-28-2
BC	fixed	0.003	0.0001	0.001	0.001
BC	tuned	0.0001	0.0001	0.0003	0.001
BC + GA	fixed	0.0001	0.01	0.01	0.003
BC + GA	tuned	0.001	0.01	0.003	0.003
BC + gLS	fixed	0.0003	0.0003	0.1	0.03

which conforms toTheorem 7.

F. Linear-Algebraic Properties of Reconstruction Matrix

In this appendix, we present a proof that for any reconstructible label transition matrix T, there exists a reconstruction matrix R such that $R^\top \mathbf{1}_\mathcal{Z} = \mathbf{1}_\mathcal{Y}$. Note that this further implies that $T^\top(R^\top \mathbf{1}_\mathcal{Z} - \mathbf{1}_\mathcal{Y}) = \mathbf{0}$, or that $R^\top \mathbf{1}_\mathcal{Z} - \mathbf{1}_\mathcal{Y} \in \operatorname{coker} T$.

A transition matrix T satisfies the identity $T^\top \mathbf{1}_\mathcal{Y} = \mathbf{1}_\mathcal{Z}$. This implies that for any $\boldsymbol{v} \in \mathbf{1}_\mathcal{Z}^\perp$,

$$\langle T\boldsymbol{v}, \mathbf{1}_\mathcal{Y} \rangle = 0, \tag{73}$$

and thus, that $T\mathbf{1}_\mathcal{Z}^\perp \subset \mathbf{1}_\mathcal{Y}^\perp$. Therefore, the restriction T' of T on $\mathbf{1}_\mathcal{Z}^\perp$ has a left-inverse R' defined on $\mathbf{1}_\mathcal{Y}^\perp$. A matrix $R = R' + k\mathbf{1}_\mathcal{Z}\mathbf{1}_\mathcal{Y}^\top$ is also a left-inverse of T'. Because $T\mathbf{1}_\mathcal{Z} \notin \mathbf{1}_\mathcal{Y}^\perp$, we can choose k such that R is a left-inverse of T. For such R, it holds that $R^\top \mathbf{1}_\mathcal{Z} \propto \mathbf{1}_\mathcal{Y}$, but because $T^\top R^\top \mathbf{1}_\mathcal{Z} = \mathbf{1}_\mathcal{Z}$ and $T^\top \mathbf{1}_\mathcal{Y} = \mathbf{1}_\mathcal{Z}$, we conclude that $R^\top \mathbf{1}_\mathcal{Z} = \mathbf{1}_\mathcal{Y}$.

G. Experimental Details

Datasets We used the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) datasets. Each dataset defines its training and test splits. In our experiment, we split the training split into two splits: one was used for training, and the other was used for validation. Table 3 lists the numbers of examples in the dataset splits.

Before starting the experiments, we converted the ground-truth labels in the training splits into complementary labels. A complementary label for an instance was randomly chosen with probabilities given by the transition matrix in Eq. (18).

Training procedure We used stochastic gradient descent with momentum to optimize the models. The momentum and the mini-batch size were fixed to 0.9 and 256, respectively. The initial learning rates were chosen from $\{0.1, 0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001\}$ as those giving the best validation accuracy. When a learning rate of 0.1 or 0.0001 achieved the best validation accuracy, we also tried two more values beyond the predefined range. In all such cases, we confirmed that the chosen values were at a peak or on a plateau of the validation accuracy. The chosen values are listed in Table 4. The default value of the weight decay coefficient is 10^{-4}, but when it is tuned, it is chosen from $\{0.1, 0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001\}$. The values of the weight decay coefficient that achieved the best validation

Table 5. Weight decay coefficient with which the best validation accuracy was achieved for BC and BC + GA.

	MNIST, linear	MNIST, MLP	CIFAR-10, ResNet-20	CIFAR-10, WRN-28-2
BC	0.0003	0.001	0.001	0.01
BC + GA	0.1	0.001	0.0003	0.0003

Table 6. Values of the coefficient k in Eq. (17) as chosen by the validation accuracy.

Dataset and model	k
MNIST, linear	0.03
MNIST, MLP	1.0
CIFAR-10, ResNet-20	1.0
CIFAR-10, WRN-28-2	1.0

Table 7. Numbers of epochs at which the best validation accuracy was achieved for each setting.

	MNIST, linear	MNIST, MLP	CIFAR-10, ResNet-20	CIFAR-10, WRN-28-2
BC	4.1	46.6	22.6	24.6
BC + GA	54.1	62.7	83.6	76.1
BC + gLS	19.5	59.1	47.9	39.7

accuracy are listed in Table 5.

Our proposed method, generalized logit squeezing (gLS), has two hyperparameters: the exponent α and the coefficient k. For a fixed α, we searched for the value of k that achieved the best validation accuracy. The candidate values were 10, 3, 1, 0.3, 0.1, 0.03, and 0.01. These results are listed in Table 6.

We adopted early stopping to determine the training time. Specifically, when the validation accuracy had not improved for the last 10 epochs, the learning rate was reduced by a factor of 10, and the third time the same condition was satisfied, the training was terminated. The test accuracy reported here is for the epochs with the best validation accuracy. Table 7 lists the numbers of epochs at which the best validation accuracy was achieved.

We used a simple grid search strategy for the hyperparameter search. The best hyperparameters (i.e., the learning rate and the gLS coefficient) were used in the evaluation step, in which a randomly initialized model was trained on the training split and evaluated on the test split. The training duration in the evaluation step was also determined by the early stopping strategy as described above.

Other details All the experiments were performed using on-premise computation servers equipped with NVIDIA's GeForce GTX 1080Ti and Tesla V100. The training duration varied significantly, depending on the methods and the model size, but the longest run took less than one hour on the Tesla V100. We used PyTorch (Paszke et al., 2019) to implement the experiments.