Semi-Supervised Learning with Adversarially Missing Label Information — Supplement

Umar Syed **Ben Taskar**
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
{usyed,taskar}@cis.upenn.edu

1 Preliminaries

For ease of reference, we restate the key definitions and assumptions from the main paper.

Definition 1 (Uniform Convergence)**.** *Loss function L has* ϵ-uniform convergence *if with probability* $1-\delta$

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|E_{\mathcal{D}}[L(\boldsymbol{\theta},x,y)]-E_{\hat{\mathbf{x}},\hat{\mathbf{y}}}[L(\boldsymbol{\theta},x,y)]\right|\leq\epsilon(\delta,m)$$

where $(\hat{\mathbf{x}},\hat{\mathbf{y}})\sim\mathcal{D}^m$ and $\epsilon(\cdot,\cdot)$ is an expression bounding the rate of convergence.

We write h to denote a *labeling function* that maps examples $\mathcal{X}$ to labels $\mathcal{Y}$. Also, for any labeling function h and unlabeled training set $\mathbf{x}\in\mathcal{X}^m$, we let $\mathbf{h}(\mathbf{x})\in\mathcal{Y}^m$ denote the vector of labels whose ith component is $h(x_i)$.

We assume that the set of all possible examples $\mathcal{X}=\{\tilde{x}_1,\ldots,\tilde{x}_N\}$ is finite. Let $\mathbf{p}_{\mathbf{x}}$ be an N-length vector that represents unlabeled training set $\mathbf{x}$ as a distribution on $\mathcal{X}$, whose ith component is $\mathbf{p}_{\mathbf{x}}(i)\triangleq\frac{|\{j\,:\,x_j=\tilde{x}_i\}|}{m}$.

Assumption 1 (∞-Separability)**.** *For all labeled training sets $(\mathbf{x},\mathbf{y})$ and $R\in\mathcal{R}(\mathbf{x},\mathbf{y})$ there exists a collection of label sets $\{Y_{\tilde{x}}:\tilde{x}\in\mathcal{X}\}$ and real-valued function F such that*

$$R(\mathbf{q})=\sum_{i=1}^{m}\chi\{\mathrm{supp}(\mathbf{q}_i)\subseteq Y_{x_i}\}+F(\mathbf{q})$$

where the characteristic function $\chi\{\cdot\}$ is 0 when its argument is true and ∞ otherwise, and $F(\mathbf{q})<\infty$ for all $\mathbf{q}\in\Delta^m$.

Assumption 2 (γ-Stability)**.** *Suppose $\mathcal{X}$ is finite. For any labeling function h^* and unlabeled training sets $\mathbf{x},\mathbf{x}'$ such that $\|\mathbf{p}_{\mathbf{x}}-\mathbf{p}_{\mathbf{x}'}\|_\infty\leq\gamma$ the following holds: For all $R\in\mathcal{R}(\mathbf{x},\mathbf{h}^*(\mathbf{x}))$ there exists $R'\in\mathcal{R}(\mathbf{x}',\mathbf{h}^*(\mathbf{x}'))$ such that*

$$R(\mathbf{h}(\mathbf{x}))<\infty \text{ if and only if } R'(\mathbf{h}(\mathbf{x}'))<\infty$$

for all labeling functions $\mathbf{h}$.

Assumption 3 (Reciprocity)**.** *For all labeled training sets $(\mathbf{x},\mathbf{y})$ and $R\in\mathcal{R}(\mathbf{x},\mathbf{y})$, if $R(\mathbf{y}')<\infty$ then $R\in\mathcal{R}(\mathbf{x},\mathbf{y}')$.*

Let A be a (possibly randomized) learning algorithm that takes a set of unlabeled training examples $\hat{\mathbf{x}}$ and a label regularization function R as input, and outputs an estimated parameter $\hat{\boldsymbol{\theta}}$. Also, if under distribution $\mathcal{D}$ each example $x\in\mathcal{X}$ is associated with exactly one label $h^*(x)\in\mathcal{Y}$, then we write $\mathcal{D}=\mathcal{D}_{\mathcal{X}}\cdot h^*$, where the *data distribution* $\mathcal{D}_{\mathcal{X}}$ is the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$.

2 Theorems 1, 2 and 3

Theorem 1. *Suppose loss function L has ϵ-uniform convergence. If $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \sim \mathcal{D}^m$ then with probability at least $1-\delta$ for all parameters $\boldsymbol{\theta} \in \Theta$ and label regularization function $R \in \mathcal{R}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$*

$$E_{\mathcal{D}}[L(\boldsymbol{\theta}, x, y)] \leq \max_{\mathbf{q} \in \Delta^m} \left(E_{\hat{\mathbf{x}}, \mathbf{q}}[L(\boldsymbol{\theta}, x, y)] - R(\mathbf{q})\right) + R(\hat{\mathbf{y}}) + \epsilon(\delta, m).$$

Proof. We have

$$\begin{aligned} E_{\mathcal{D}}[L(\boldsymbol{\theta}, x, y)] =& E_{\mathcal{D}}[L(\boldsymbol{\theta}, x, y)] - R(\hat{\mathbf{y}}) + R(\hat{\mathbf{y}}) \leq E_{\hat{\mathbf{x}}, \hat{\mathbf{y}}}[L(\boldsymbol{\theta}, x, y)] - R(\hat{\mathbf{y}}) + R(\hat{\mathbf{y}}) + \epsilon(\delta, m) \\ &\leq \max_{\mathbf{q} \in \Delta^m} \left(E_{\hat{\mathbf{x}}, \mathbf{q}}\left[L(\boldsymbol{\theta}, x, y)\right] - R(\mathbf{q})\right) + R(\hat{\mathbf{y}}) + \epsilon(\delta, m) \end{aligned}$$

where the first inequality follows from Definition 1. □

The proof of Theorem 2 is fairly complicated, but a similar result can be proved quite easily if we assume that the labeled training set $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is fixed, rather than drawn from a distribution, and that the learning algorithm A is deterministic. This allows us to argue completely deterministically, making it easy to select a labeled training set that achieves the desired lower bound. We call this simpler result Theorem 2′, and intend its proof to serve as a warm-up that conveys the intuition behind the proof of Theorem 2.

Theorem 2′ (Warm-up)**.** *Suppose Assumptions 1 and 3 hold for label regularization function family $\mathcal{R}$. For all learning algorithms A and unlabeled training sets $\hat{\mathbf{x}}$ there exist labels $\hat{\mathbf{y}}$ such that*

$$E_{\hat{\mathbf{x}}, \hat{\mathbf{y}}}[L(\hat{\boldsymbol{\theta}}, x, y)] \geq \max_{\mathbf{q} \in \Delta^m} \left(E_{\hat{\mathbf{x}}, \mathbf{q}}[L(\hat{\boldsymbol{\theta}}, x, y)] - R(\mathbf{q})\right) + \min_{\mathbf{q} \in \Delta^m} R(\mathbf{q})$$

for some $R \in \mathcal{R}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, where $\hat{\boldsymbol{\theta}}$ is the parameter output by A.

Proof. Choose any labels $\mathbf{y}'$ and any $R \in \mathcal{R}(\hat{\mathbf{x}}, \mathbf{y}')$, and let $\hat{\boldsymbol{\theta}}$ be the parameter output by algorithm A when given $\hat{\mathbf{x}}$ and R as input. If we let

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}: R(\mathbf{y}) < \infty} E_{\hat{\mathbf{x}}, \mathbf{y}}[L(\hat{\boldsymbol{\theta}}, x, y)]$$

then by Assumption 3 we have $R \in \mathcal{R}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$. So if $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is the labeled training set we can force algorithm A to output parameter $\hat{\boldsymbol{\theta}}$. Thus

$$\begin{aligned} E_{\hat{\mathbf{x}}, \hat{\mathbf{y}}}[L(\hat{\boldsymbol{\theta}}, x, y)] &= \max_{\mathbf{y}: R(\mathbf{y}) < \infty} \left(E_{\hat{\mathbf{x}}, \mathbf{y}}[L(\hat{\boldsymbol{\theta}}, x, y)]\right) - \min_{\mathbf{q} \in \Delta^m} R(\mathbf{q}) + \min_{\mathbf{q} \in \Delta^m} R(\mathbf{q}) \\ &\geq \max_{\mathbf{q} \in \Delta^m} \left(E_{\hat{\mathbf{x}}, \mathbf{q}}[L(\hat{\boldsymbol{\theta}}, x, y)] - R(\mathbf{q})\right) + \min_{\mathbf{q} \in \Delta^m} R(\mathbf{q}) \end{aligned}$$

where the inequality follows from Assumption 1. □

Theorem 2′ was proved by selecting a worst-case labeling $\hat{\mathbf{y}}$ for the fixed training set $\hat{\mathbf{x}}$. Selecting such a labeling is not so straightforward in the case when $\hat{\mathbf{x}}$ is drawn from a distribution, because different values for $\hat{\mathbf{x}}$ may require different (and inconsistent) labelings. As a consequence, Theorem 2 requires a significantly extended analysis that leverages Assumption 2; this assumption ensures that a single worst-case labeling exists with high-probability whenever the number of training examples is sufficiently large. Further, when the learning algorithm A is randomized, it can avoid suffering an arbitrarily large loss simply by guessing the label of every example; the existence of this strategy is the reason for an extra constant factor in the lower bound in Theorem 2 versus the lower bound in Theorem 2′.

Theorem 2. *Suppose Assumptions 1, 2 and 3 hold for label regularization function family $\mathcal{R}$, the loss function L is 0-1 loss, and the set of all possible examples $\mathcal{X}$ is finite. For all learning algorithms A and data distributions $\mathcal{D}_{\mathcal{X}}$ there exists a labeling function h^* such that if $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \sim \mathcal{D}^m$ (where $\mathcal{D} = \mathcal{D}_{\mathcal{X}} \cdot h^*$) and $m \geq O(\frac{1}{\gamma^2} \log \frac{|\mathcal{X}|}{\delta})$ then with probability at least $\frac{1}{4} - 2\delta$*

$$E_{\mathcal{D}}[L(\hat{\boldsymbol{\theta}}, x, y)] \geq \frac{1}{4} \max_{\mathbf{q} \in \Delta^m} \left(E_{\hat{\mathbf{x}}, \mathbf{q}}[L(\hat{\boldsymbol{\theta}}, x, y)] - R(\mathbf{q})\right) + \min_{\mathbf{q} \in \Delta^m} R(\mathbf{q}) - \epsilon(\delta, m)$$

for some $R \in \mathcal{R}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, where $\hat{\boldsymbol{\theta}}$ is the parameter output by A, and γ is the constant from Assumption 2.

Proof. Noting that $\mathcal{X} = \{\tilde{x}_1, \ldots, \tilde{x}_N\}$, let $\overline{\mathbf{p}}$ be an N-length vector whose ith component is the probability assigned by data distribution $\mathcal{D}_{\mathcal{X}}$ to example $\tilde{x}_i$. A straightforward calculation using the Chernoff bound shows that if $\hat{\mathbf{x}} \sim \mathcal{D}_{\mathcal{X}}^m$ and $m \geq O(\frac{1}{\gamma^2} \log \frac{|\mathcal{X}|}{\delta})$ then $\|\mathbf{p}_{\hat{\mathbf{x}}} - \overline{\mathbf{p}}\|_\infty \leq \frac{\gamma}{2}$ with probability $1 - \delta$.

Since $1 - \delta > 0$, there must exist $\overline{\mathbf{x}} \in \mathcal{X}^m$ such that $\|\mathbf{p}_{\overline{\mathbf{x}}} - \overline{\mathbf{p}}\|_\infty \leq \frac{\gamma}{2}$. Now choose any labels $\mathbf{y}'$ and $R_{\overline{\mathbf{x}}} \in \mathcal{R}(\overline{\mathbf{x}}, \mathbf{y}')$, and let

$$\overline{\mathbf{y}} = \arg \max_{\mathbf{y}: R_{\overline{\mathbf{x}}}(\mathbf{y}) < \infty} E_{\overline{\mathbf{x}}, \mathbf{y}}[L(\theta, x, y)]$$

where we let $\theta = A(\overline{\mathbf{x}}, R_{\overline{\mathbf{x}}})$. By Assumption 1 and linearity, $\overline{\mathbf{y}}$ assigns identical labels to identical examples in $\overline{\mathbf{x}}$. Thus we can select a labeling function h^* such that $\mathbf{h}^*(\overline{\mathbf{x}}) = \overline{\mathbf{y}}$. This is the labeling function asked for by the statement of the theorem, with the caveat that we will need to modify h^* later in the proof.

We are now ready to define the behavior of the labeler, i.e. the choice of $R \in \mathcal{R}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ for each labeled training set $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ (where $\hat{\mathbf{y}} = \mathbf{h}^*(\hat{\mathbf{x}})$). Say that $\mathbf{x}$ is γ-*close* if $\|\mathbf{p}_{\mathbf{x}} - \mathbf{p}_{\overline{\mathbf{x}}}\|_\infty \leq \gamma$. For every γ-close $\mathbf{x}$, Assumption 2 permits us to fix an $R_{\mathbf{x}} \in \mathcal{R}(\mathbf{x}, \mathbf{h}^*(\mathbf{x}))$ such that $R_{\mathbf{x}}(\mathbf{h}(\mathbf{x})) < \infty$ if and only if $R_{\overline{\mathbf{x}}}(\mathbf{h}(\overline{\mathbf{x}})) < \infty$ for all labeling functions $\mathbf{h}$. So if the training set $\hat{\mathbf{x}}$ is γ-close, we demand that the labeler return $R_{\hat{\mathbf{x}}}$ to the learning algorithm A. The labeler's behavior when the training set $\hat{\mathbf{x}}$ is not γ-close can be arbitrary.

Let $X^\gamma \subseteq \mathcal{X}$ be the set of all examples $\tilde{x} \in \mathcal{X}$ such that $\tilde{x}$ appears in at least one γ-close $\mathbf{x}$. For any $R_{\mathbf{x}}$ such that $\mathbf{x}$ is γ-close (where was $R_{\mathbf{x}}$ defined above), consider the collection of label sets $\{Y_{\tilde{x}} : \tilde{x} \in \mathcal{X}\}$ satisfying the guarantee in Assumption 1. Note that, by Assumption 2, a *single* collection of label sets satisfies the guarantee for *all* γ-close $\mathbf{x}$. Let $\{Y_{\tilde{x}}^\gamma : \tilde{x} \in X^\gamma\}$ be one such collection. Now consider *any* labeling function h satisfying $h(\tilde{x}) \in Y_{\tilde{x}}^\gamma$ for all $\tilde{x} \in X^\gamma$. If $\mathbf{x}$ is γ-close, then by Assumption 3 we have $R_{\mathbf{x}} \in \mathcal{R}(\mathbf{x}, \mathbf{h}(\mathbf{x}))$. We will use this fact below when modifying h^*.

We are now ready to modify h^* in a way that forces the learning algorithm A to suffer large loss. Let $\boldsymbol{\theta}(A, \mathbf{x}, R)$ denote the parameter returned by learning algorithm A on training set $\mathbf{x}$ and label regularization function R, which is a random variable due to the possible randomization of algorithm A. Now partition X^γ into two disjoint sets: $X^{\gamma,1} = \{\tilde{x} \in X^\gamma : |Y_{\tilde{x}}^\gamma| = 1\}$ and $X^{\gamma,2+} = \{\tilde{x} \in X^\gamma : |Y_{\tilde{x}}^\gamma| > 1\}$. For each $\tilde{x} \in X^\gamma$ we modify $h^*(\tilde{x})$ as follows: If $\tilde{x} \in X^{\gamma,1}$ then set $h^*(\tilde{x}) = \tilde{y}$, where $Y_{\tilde{x}}^\gamma = \{\tilde{y}\}$. Otherwise, if $\tilde{x} \in X^{\gamma,2+}$ then set $h^*(\tilde{x}) = \tilde{y}$, where $\tilde{y}$ satisfies

$$\Pr_{A, \mathbf{x} \sim \mathcal{D}_{\mathcal{X}}^m} \left[h_{\boldsymbol{\theta}(A, \mathbf{x}, R_{\mathbf{x}})}(\tilde{x}) \neq \tilde{y} \;\middle|\; \mathbf{x} \text{ is } \gamma\text{-close} \right] \geq \frac{1}{2} \tag{1}$$

where $\Pr_{A, \mathbf{x} \sim \mathcal{D}_{\mathcal{X}}^m}[\cdot]$ denotes probability with respect to the randomization of learning algorithm A and the choice of $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}^m$. Note that such a label $\tilde{y}$ must exist, because $|Y_{\tilde{x}}^\gamma| > 1$. Importantly, by the fact given above, this modification of h^* does not affect the previously defined behavior of the labeler, because we still have $R_{\mathbf{x}} \in \mathcal{R}(\mathbf{x}, \mathbf{h}^*(\mathbf{x}))$ for all γ-close $\mathbf{x}$.

Let $I \subseteq [m]$. Define the random variable $Z_I = \frac{1}{|I|} \sum_{i \in I} \mathbf{1}\{h_{\boldsymbol{\theta}(A, \mathbf{x}, R_{\mathbf{x}})}(x_i) \neq h^*(x_i)\}$. We have

$$\begin{aligned}
&E_{A, \mathbf{x} \sim \mathcal{D}_{\mathcal{X}}^m} \left[Z_I \;\middle|\; \mathbf{x} \text{ is } \gamma\text{-close and } x_i \in X_\gamma^{2+} \text{ for all } i \in I \right] \\
&= \frac{1}{|I|} \sum_{i \in I} \Pr_{A, \mathbf{x} \sim \mathcal{D}_{\mathcal{X}}^m} \left[h_{\boldsymbol{\theta}(A, \mathbf{x}, R_{\mathbf{x}})}(x_i) \neq h^*(x_i) \;\middle|\; \mathbf{x} \text{ is } \gamma\text{-close and } x_i \in X_\gamma^{2+} \text{ for all } i \in I \right] \\
&\geq \frac{1}{2}
\end{aligned} \tag{2}$$

which follows from Eq. (1).

For any random variable $Z \in [0, 1]$ we know that $E[Z] \leq \Pr[Z \geq a] + a$ for all $a > 0$. Combining this inequality for $a = \frac{1}{4}$ with the bound in Eq. (2) yields the following: If $\mathbf{x}$ is γ-close and $x_i \in X_\gamma^{2+}$ for all $i \in I$ then

$$\frac{1}{|I|} \sum_{i \in I} \mathbf{1}\{h_{\boldsymbol{\theta}(A, \mathbf{x}, R_{\mathbf{x}})}(x_i) \neq h^*(x_i)\} \geq \frac{1}{4} \tag{3}$$

with probability at least $\frac{1}{4}$.

Recall that the training set $\hat{\mathbf{x}}$ is drawn from $\mathcal{D}_{\mathcal{X}}^m$ and that $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(A, \hat{\mathbf{x}}, R_{\hat{\mathbf{x}}})$. Assume that $\hat{\mathbf{x}}$ is γ-close; this occurs with probability $1 - \delta$. Let $I = \{i \in [m] : \hat{x}_i \in X_\gamma^{2+}\}$ be the indices of examples in $\hat{\mathbf{x}}$ that are in X_γ^{2+}. We have

$$
\begin{aligned}
E_{\hat{\mathbf{x}},\hat{\mathbf{y}}}[L(\hat{\boldsymbol{\theta}}, x, y)] &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{h_{\hat{\boldsymbol{\theta}}}(\hat{x}_i) \neq h^*(\hat{x}_i)\} \\
&= \frac{|I|}{m}\frac{1}{|I|}\sum_{i\in I} \mathbf{1}\{h_{\hat{\boldsymbol{\theta}}}(\hat{x}_i) \neq h^*(\hat{x}_i)\} + \frac{1}{m}\sum_{i\in[m]\setminus I} \mathbf{1}\{h_{\hat{\boldsymbol{\theta}}}(\hat{x}_i) \neq h^*(\hat{x}_i)\} \\
&\geq \frac{|I|}{m}\frac{1}{4} + \frac{1}{m}\sum_{i\in[m]\setminus I} \mathbf{1}\{h_{\hat{\boldsymbol{\theta}}}(\hat{x}_i) \neq h^*(\hat{x}_i)\} && (4) \\
&= \frac{1}{m}\frac{1}{4}\sum_{i\in I} \max_{y\in Y_{\hat{x}_i}^\gamma} [L(\hat{\boldsymbol{\theta}}, \hat{x}_i, y)] + \frac{1}{m}\sum_{i\in[m]\setminus I} \mathbf{1}\{h_{\hat{\boldsymbol{\theta}}}(\hat{x}_i) \neq h^*(\hat{x}_i)\} && (5) \\
&= \frac{1}{m}\frac{1}{4}\sum_{i\in I} \max_{y\in Y_{\hat{x}_i}^\gamma} [L(\hat{\boldsymbol{\theta}}, \hat{x}_i, y)] + \frac{1}{m}\sum_{i\in[m]\setminus I} \max_{y\in Y_{\hat{x}_i}^\gamma} [L(\hat{\boldsymbol{\theta}}, \hat{x}_i, y)] && (6) \\
&\geq \frac{1}{4}\frac{1}{m}\sum_{i=1}^{m} \max_{y\in Y_{\hat{x}_i}^\gamma} [L(\hat{\boldsymbol{\theta}}, \hat{x}_i, y)] \\
&= \frac{1}{4}\max_{\mathbf{y}: R_{\hat{\mathbf{x}}}(\mathbf{y})<\infty} \left(E_{\hat{\mathbf{x}},\mathbf{y}}[L(\hat{\boldsymbol{\theta}}, x, y)]\right). && (7)
\end{aligned}
$$

By Eq. (3), Eq. (4) holds with probability $\frac{1}{4}$. Eq. (5) holds because, when L is 0-1 loss, $\max_{y\in Y}[L(\boldsymbol{\theta}, x, y)] = 1$ whenever $|Y| > 1$. Eq. (6) holds because, for all $\tilde{x} \in X_\gamma^1$, we set $h^*(\tilde{x}) = \tilde{y}$, where $Y_{\tilde{x}}^\gamma = \{\tilde{y}\}$. Eq. (7) follows from Assumption 1.

Continuing, we have

$$
\begin{aligned}
E_{\mathcal{D}}[L(\hat{\boldsymbol{\theta}}, x, y)] &\geq E_{\hat{\mathbf{x}},\hat{\mathbf{y}}}[L(\hat{\boldsymbol{\theta}}, x, y)] - \epsilon(\delta, m) && (8) \\
&\geq \frac{1}{4}\max_{\mathbf{y}: R_{\hat{\mathbf{x}}}(\mathbf{y})<\infty} \left(E_{\hat{\mathbf{x}},\mathbf{y}}[L(\hat{\boldsymbol{\theta}}, x, y)]\right) - \min_{\mathbf{q}\in\Delta^m} R_{\hat{\mathbf{x}}}(\mathbf{q}) + \min_{\mathbf{q}\in\Delta^m} R_{\hat{\mathbf{x}}}(\mathbf{q}) - \epsilon(\delta, m) && (9) \\
&\geq \frac{1}{4}\max_{\mathbf{q}\in\Delta^m} \left(E_{\hat{\mathbf{x}},\mathbf{q}}[L(\hat{\boldsymbol{\theta}}, x, y)] - R_{\hat{\mathbf{x}}}(\mathbf{q})\right) + \min_{\mathbf{q}\in\Delta^m} R_{\hat{\mathbf{x}}}(\mathbf{q}) - \epsilon(\delta, m) && (10)
\end{aligned}
$$

where Eq. (8) holds with probability $1 - \delta$ by Defintion 1, Eq. (9) follows from Eq. (7) and Eq. (10) follows from Assumption 1.

Taking the union bound over all events that were conditioned on in the preceding argument, we find that Eq. (10) holds with probability $\frac{1}{4} - 2\delta$, and this proves the theorem. □

Theorem 3. *Suppose the loss function L is 0-1 loss. There exists a label regularization function family $\mathcal{R}$ that satisfies Assumptions 1 and 2, but not Assumption 3, and a learning algorithm A such that for all distributions $\mathcal{D}$ if $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \sim \mathcal{D}^m$ then with probability at least $1 - \delta$*

$$
E_{\mathcal{D}}[L(\hat{\boldsymbol{\theta}}, x, y)] \leq \max_{\mathbf{q}\in\Delta^m} \left(E_{\hat{\mathbf{x}},\mathbf{q}}[L(\hat{\boldsymbol{\theta}}, x, y)] - R(\mathbf{q})\right) + \min_{\mathbf{q}\in\Delta^m} R(\mathbf{q}) + \epsilon(\delta, m) - 1
$$

for some $R \in \mathcal{R}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, where $\hat{\boldsymbol{\theta}}$ is the parameter output by A.

Proof. Consider a one-to-one correspondence $f : \mathcal{Y}^m \to \mathcal{Y}^m$ such that if $\mathbf{y}' = f(\mathbf{y})$ then $y'(i) \neq y(i)$ for all $i \in [m]$. In other words, f maps each labeling $\mathbf{y}$ to a labeling $\mathbf{y}'$ that assigns a different label to every example. Clearly, as long as $|\mathcal{Y}| > 1$ such an f can always be chosen.

Now suppose each $\mathcal{R}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ contains a single label regularization function R such that $R(\mathbf{q}) = 0$ if $\mathbf{q} = f(\hat{\mathbf{y}})$ and $R(\mathbf{q}) = \infty$ otherwise. Note that this violates Assumption 3.

Now consider a learning algorithm A that does the following: Given $(\hat{\mathbf{x}}, R)$, algorithm A finds the (unique) labeling $\mathbf{y}'$ that minimizes R, then recovers the correct labeling $\hat{\mathbf{y}}$ by setting $\hat{\mathbf{y}} =$

$f^{-1}(\mathbf{y}')$, and then finds $\hat{\boldsymbol{\theta}}$ that minimizes $E_{\hat{\mathbf{x}},\hat{\mathbf{y}}}[L(\boldsymbol{\theta}, x, y)]$. Note that since the function f was chosen arbitrarily, finding $\mathbf{y}'$ will be computationally infeasible in general. Now we have

$$\begin{aligned} E_{\mathcal{D}}[L(\hat{\boldsymbol{\theta}}, x, y)] \leq & E_{\hat{\mathbf{x}},\hat{\mathbf{y}}}[L(\hat{\boldsymbol{\theta}}, x, y)] + \epsilon(\delta, m) \leq E_{\hat{\mathbf{x}}, f(\hat{\mathbf{y}})}[L(\hat{\boldsymbol{\theta}}, x, y)] + \epsilon(\delta, m) - 1 \\ = & \max_{\mathbf{q} \in \Delta^m} \left(E_{\hat{\mathbf{x}},\mathbf{q}}[L(\hat{\boldsymbol{\theta}}, x, y)] - R(\mathbf{q}) \right) + \min_{\mathbf{q} \in \Delta^m} R(\mathbf{q}) + \epsilon(\delta, m) - 1 \end{aligned}$$

where the first inequality follows from Definition 1, the second inequality follows from the choice of f, and the last equality follows from the choice of R. □

3 Analysis of Algorithm 1

Algorithm 1 GAME: Game for Adversarially Missing Evidence

1: **Given:** Constants $\epsilon_1, \epsilon_2 > 0$.
2: Find $\tilde{\mathbf{q}}$ such that $\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}, \tilde{\mathbf{q}}) \geq \max_{\mathbf{q} \in \Delta^m} \min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}, \mathbf{q}) - \epsilon_1$
3: Find $\tilde{\boldsymbol{\theta}}$ such that $F(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{q}}) \leq \min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}, \tilde{\mathbf{q}}) + \epsilon_2$
4: **Return:** Parameter estimate $\tilde{\boldsymbol{\theta}}$.

Recall that the goal of Algorithm 1 is to find a parameter $\boldsymbol{\theta}^*$ that realizes the minimum

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{q} \in \Delta^m} \left(E_{\hat{\mathbf{x}},\mathbf{q}}[L(\boldsymbol{\theta}, x, y)] - R(\mathbf{q}) \right) + \alpha \left\| \boldsymbol{\theta} \right\|^2. \tag{11}$$

Before we can analyze Algorithm 1, we need a definition.

Definition 2. *A function $f : S \to \mathbb{R}$ is* κ-strongly convex *if for all $x, y \in S$ and $\lambda \in [0, 1]$*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{1}{2} \kappa \lambda (1-\lambda) \left\| x - y \right\|^2.$$

So a κ-strongly convex function is one that is "curved" everywhere, where the amount of curvature is given by κ. It is easy to show that $F(\boldsymbol{\theta}, \mathbf{q})$ is an α-strongly convex function of $\boldsymbol{\theta}$, for any fixed $\mathbf{q}$. This is because the loss function L is convex in $\boldsymbol{\theta}$, and the addition of the term $\alpha \left\| \boldsymbol{\theta} \right\|^2$ makes it α-strongly convex. The next lemma proves that all approximate minimizers of a strongly convex function must be near each other.

Lemma 1. *If f is κ-strongly convex and $x^* = \arg\min_x f(x)$ and $f(\tilde{x}) \leq f(x^*) + \epsilon$ then*

$$\left\| x^* - \tilde{x} \right\| \leq \sqrt{\frac{2}{\kappa} \epsilon}.$$

Proof. Choose any $\lambda \in [0, 1)$. We have

$$\begin{aligned} f(x^*) \leq & f(\lambda x^* + (1-\lambda)\tilde{x}) \\ \leq & \lambda f(x^*) + (1-\lambda) f(\tilde{x}) - \frac{1}{2} \kappa \lambda (1-\lambda) \left\| x^* - \tilde{x} \right\|^2 \\ \leq & f(x^*) + (1-\lambda)\epsilon - \frac{1}{2} \kappa \lambda (1-\lambda) \left\| x^* - \tilde{x} \right\|^2 \end{aligned}$$

where we used the definitions of x^*, κ-strongly convex functions, and $\tilde{x}$, in that order.

Some algebra yields $\left\| x^* - \tilde{x} \right\| \leq \sqrt{\frac{2}{\kappa \lambda} \epsilon}$ where we were able to cancel a $(1-\lambda)$ factor from both sides because $\lambda < 1$. Now taking the limit of this upper bound as $\lambda \to 1$ proves the lemma. □

We now prove that Algorithm 1 produces a good estimate of $\boldsymbol{\theta}^*$, the minimum of the objective (11).

Theorem 4. *The parameter $\tilde{\boldsymbol{\theta}}$ output by Algorithm 1 satisfies*

$$\left\| \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\| \leq \sqrt{\frac{8}{\alpha} (\epsilon_1 + \epsilon_2)}.$$

Proof. $F(\boldsymbol{\theta}, \mathbf{q})$ is convex in $\boldsymbol{\theta}$ and concave in $\mathbf{q}$, and Δ^m is convex and compact. Therefore, by Sion's minimax theorem [1] we have

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{q} \in \Delta^m} F(\boldsymbol{\theta}, \mathbf{q}) = \max_{\mathbf{q} \in \Delta^m} \min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}, \mathbf{q}) \triangleq v^* \tag{12}$$

where we defined v^* to be the common value of both sides of the equation. Also note that $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \max_{\mathbf{q} \in \Delta^m} F(\boldsymbol{\theta}, \mathbf{q})$, by definition.

We will show that both $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ are approximate minimizers of the function $F(\boldsymbol{\theta}, \tilde{\mathbf{q}})$. We have

$$F(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{q}}) \geq \min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}, \tilde{\mathbf{q}}) \geq v^* - \epsilon_1 \tag{13}$$

where we used, in order: minimization over $\boldsymbol{\theta}$; the definition of Algorithm 1 and Eq. (12). We also have

$$F(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{q}}) \leq \min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}, \tilde{\mathbf{q}}) + \epsilon_2 \leq v^* + \epsilon_2 \tag{14}$$

where we used, in order: Algorithm 1; maximization over $\mathbf{q}$ and Eq. (12). Putting these together, we obtain

$$v^* - \epsilon_1 - \epsilon_2 \leq F(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{q}}) - \epsilon_2 \leq \min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}, \tilde{\mathbf{q}}) \leq F(\boldsymbol{\theta}^*, \tilde{\mathbf{q}}) \leq v^*$$

where we used, in order: Eq. (13); Eq. (14); minimization over $\boldsymbol{\theta}$; maximization over $\mathbf{q}$ and the definition of $\boldsymbol{\theta}^*$ and Eq. (12).

The last line implies that both $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ are $(\epsilon_1 + \epsilon_2)$-approximate minimizers of $F(\boldsymbol{\theta}, \tilde{\mathbf{q}})$. And since $F(\boldsymbol{\theta}, \tilde{\mathbf{q}})$ is α-strongly convex in $\boldsymbol{\theta}$, Lemma 1 and the triangle inequality together imply the theorem. □

References

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