Strength from Weakness: Fast Learning Using Weak Supervision

Joshua Robinson [1] **Stefanie Jegelka** [1] **Suvrit Sra** [1]

Abstract

We study generalization properties of weakly supervised learning, that is, learning where only a few "strong" labels (the actual target for prediction) are present but many more "weak" labels are available. In particular, we show that *pretraining* using weak labels and *finetuning* using strong can accelerate the learning rate for the strong task to the fast rate of $\mathcal{O}(1/n)$, where n is the number of strongly labeled data points. This acceleration can happen even if, by itself, the strongly labeled data admits only the slower $\mathcal{O}(1/\sqrt{n})$ rate. The acceleration depends continuously on the number of weak labels available, and on the relation between the two tasks. Our theoretical results are reflected empirically across a range of tasks and illustrate how weak labels speed up learning on the strong task.

1. Introduction

While access to large amounts of labeled data has enabled the training of big models with great successes in applied machine learning, labeled data remains a key bottleneck. In numerous settings (e.g., scientific measurements, experiments, medicine), obtaining a large number of labels can be prohibitively expensive, error prone, or otherwise infeasible.

When labels are scarce, a common alternative is to use additional sources of information: "weak labels" that contain information about the "strong" target task and are more readily available, e.g., a related task, or noisy versions of strong labels from non-experts or cheaper measurements.

Such a setting is called *weakly supervised learning*, and, given its great practical relevance, it has received much attention (Zhou, 2018; Pan & Yang, 2009; Liao et al., 2005; Dai et al., 2007; Huh et al., 2016). A prominent example that enabled breakthrough results in computer vision and is now standard is *pretraining*, where one first trains a complex model on a related, large data task, and to then uses the learned features for finetuning on the small-data target task (Girshick et al., 2014; Donahue et al., 2014; Zeiler & Fergus, 2014; Sun et al., 2017). Numerous approaches to weakly supervised learning have succeeded in a variety of tasks; beyond computer vision (Oquab et al., 2015; Durand et al., 2017; Carreira & Zisserman, 2017; Fries et al., 2019). Examples include clinical text classification (Wang et al., 2019), sentiment analysis (Medlock & Briscoe, 2007), social media content tagging (Mahajan et al., 2018) and many others. Weak supervision is also closely related to unsupervised learning methods such as complementary and contrastive learning (Xu et al., 2019; Chen & Batmanghelich, 2019; Arora et al., 2019), and particularly to self-supervised learning (Doersch et al., 2015), where feature maps learned via supervised training on artificially constructed tasks have been found to even outperform ImageNet learned features on certain downstream tasks (Misra & van der Maaten, 2019).

In this paper, we make progress towards building theoretical foundations for weakly supervised learning, i.e., where we have a few strong labels, but too few to learn a good model in a conventional supervised manner. Specifically we ask,

> *Under what conditions can large amounts of weakly labeled data provably help us learn a better model than strong labels alone?*

We answer this question by analyzing a generic feature learning algorithm that learns features by pretraining on the weak task, and fine-tunes a model on those features for the strong downstream task. While generalization bounds for supervised learning typically scale as $\mathcal{O}(1/\sqrt{n})$, where n is the number of strongly labeled data points, we show that the pretrain-finetune algorithm can do better, achieving the superior rate of $\widetilde{\mathcal{O}}(n^{-\gamma})$ for $1/2 \leq \gamma \leq 1$, where γ depends on how much weak data is available, and on generalization error for the weak task. This rate smoothly interpolates between $\widetilde{\mathcal{O}}(1/n)$ in the best case, when weak data is plentiful and the weak task is not too difficult, and slower rates when less weak data is available or the weak task itself is hard.

One instantiation of our results for categorical weak labels says that, if we can train a model with $\mathcal{O}(1/\sqrt{m})$ excess risk for the weak task (where m is the amount of weak data), and $m = \Omega(n^2)$, then we obtain a "fast rate" $\widetilde{\mathcal{O}}(1/n)$ on the

[1]Massachusetts Institute of Technology, Cambridge, MA 02139. Correspondence to: Joshua Robinson <joshrob@mit.edu>.

Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

excess risk of the *strong task*. This speedup is significant compared to the commonly observed $\mathcal{O}(1/\sqrt{n})$ "slow rates".

To obtain any such results, it is necessary to capture the task relatedness between weak and strong tasks. We formulate a general sufficient condition: that there exists a shared mutual embedding for which predicting the strong label is "easy", and predicting the weak label is possible. "Easy" prediction is formalized via the *central condition* (van Erven et al., 2012; 2015), a property that ensures that learning improves quickly as more data is added. We merely assume existence of such an embedding; *a priori* we do not know what this shared embedding is. Our theoretical analysis shows that learning an estimate of the embedding by pretraining on the weak task still allows fast learning on the strong task.

In short, we make the following contributions:

- We introduce a theoretical framework for analyzing weakly supervised learning problems.
- We propose the shared embedding plus central condition as a viable way to quantify relatedness between weak and strong tasks. The condition merely posits the existence of such an embedding; this makes obtaining generalization bounds non-trivial.
- We obtain generalization bounds for the strong task. These bounds depend continuously on two key quantities: 1) the growth rate of the number m of weak labels in terms of the number n of strong labels, and 2) generalization performance on the weak task.
- We show that in the best case, when m is sufficiently larger than n, weak supervision delivers *fast rates*.

We validate our theoretical findings, and observe that our fast and intermediate rates are indeed observed in practice.

1.1. Examples of Weak Supervision

Coarse Labels. It is often easier to collect labels that capture only part of the information about the true label of interest (Zhao et al., 2011; Guo et al., 2018; Yan et al., 2015; Taherkhani et al., 2019). A particularly pertinent example is semantic labels obtained from hashtags attached to images (Mahajan et al., 2018; Li et al., 2017). Such tags are generally easy to gather in large quantities, but tend to only capture certain aspects of the image that the person tagging them focused on. For example, an image with the tag `#dog` could easily also contain children, or other label categories that have not been explicitly tagged.

Crowd Sourced Labels. A primary way for obtaining large labeled data is via crowd-sourcing using platforms such as Amazon Mechanical Turk (Khetan et al., 2018; Kleindessner & Awasthi, 2018). Even for the simplest of labeling tasks, crowd-sourced labels can often be noisy (Zhang & Sabuncu, 2018; Branson et al., 2017; Zhang et al., 2014), which becomes worse for labels requiring expert knowledge. Typically, more knowledgeable labelers are more expensive (e.g., professional doctors versus medical students for a medical imaging task), which introduces a tradeoff between label quality and cost that the user must carefully manage.

Object Detection. A common computer vision task is to draw bounding boxes around objects in an image (Oquab et al., 2015). A popular alternative to expensive bounding box annotations is a set of words describing the objects present, without localization information (**?**Bilen & Vedaldi, 2016; Branson et al., 2017; Wan et al., 2018). This setting too is an instance of coarse labeling.

Model Personalization. In examples like recommender systems (Ricci et al., 2011), online advertising (Naumov et al., 2019), and personalized medicine (Schork, 2015), one needs to make predictions for individuals, while information shared by a larger population acts as supportive, weak supervision (Desrosiers & Karypis, 2011).

2. Weakly Supervised Learning

We begin with some notation. The spaces $\mathcal{X}$ and $\mathcal{Y}$ denote as usual the space of features and strong labels. In *weakly supervised learning*, we have in addition $\mathcal{W}$, the space of weak labels. We receive the tuple (X, W, Y) drawn from the product space $\mathcal{X} \times \mathcal{W} \times \mathcal{Y}$. The goal is to then predict the strong label Y using the features X, and possibly benefiting from the related information captured by W.

More specifically, we work with two datasets: (1) a weakly labeled dataset $\mathcal{D}_m^{\text{weak}}$ of m examples drawn independently from the marginal distribution $P_{X,W}$; and (2) a dataset $\mathcal{D}_n^{\text{strong}}$ of n strong labeled examples drawn from the marginal $P_{X,Y}$. Typically, $n \ll m$.

We then use the weak labels to learn an embedding in a latent space $\mathcal{Z} \subseteq \mathbb{R}^s$. In particular, we assume that there exists an unknown "good" embedding $Z = g_0(X) \in \mathcal{Z}$, using which a linear predictor A^* can determine W, i.e., $A^*Z = A^*g_0(X) = W$ in the regression case, and $\sigma(A^*g_0(X)) = W$ in the classification setting, where σ is the sigmoid function. The g_0 assumption holds, for example, whenever W is a deterministic function of X. This assumption is made to simplify the exposition; if it does not hold exactly then one can still obtain a generalization guarantee by introducing an additive term to the final generalization bound equal to the smallest error attainable by any measurable hypothesis, reflecting the inherent noise in the problem of predicting W from X.

Using the latent space $\mathcal{Z}$, we define two function classes: *strong predictors* $\mathcal{F} \subset \{f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}\}$, and *weak feature maps* $\mathcal{G} \subset \{g : \mathcal{X} \to \mathcal{Z}\}$. Later we will assume that class $\mathcal{F}$ is parameterized, and identify functions f in $\mathcal{F}$

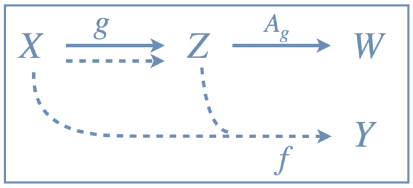

Figure 1. Schema for weakly supervised learning using Algorithm 1. The dotted lines denote the flow of strong data, and the solid lines the flow of weak data.

with parameter vectors. We then learn a predictor $f \in \mathcal{F}$ by replacing the latent vector Z with an embedding $\hat{g}(X) \in \mathcal{Z}$ that we learn from weakly labeled data. Corresponding to each of these function classes we introduce two loss functions.

First, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ measures loss of the strong predictor; we assume this loss to be continuously differentiable in its first argument. We will equivalently write $\ell_f(x,z,y) := \ell(f(x,z),y)$ for predicting from a latent vector $z \in \mathcal{Z}$; similarly, for predicting from an estimate $\hat{z} = g(x)$, we write the loss as $\ell_{f(\cdot,g)}(x,y) := \ell(f(x,g(x)),y)$.

Second, $\ell^{\text{weak}} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}_+$ measures loss for the weak task. This loss also applies to measuring loss of feature maps $g : \mathcal{X} \to \mathcal{Z}$, by using the best possible downstream linear classifier, i.e., $\ell_g^{\text{weak}}(x,w) = \ell^{\text{weak}}(A_g^\top g(x), w)$ where $A_g \in \arg\min_{A \in \mathbb{R}^{|\mathcal{Y}| \times s}} \mathbb{E}\ell^{\text{weak}}(Ag(X), W)$. In the classification case, for notational simplicity, we fold the softmax function into ℓ^{weak} to convert numeric predictions to a probability distribution.

Our primary goal is to learn a model $\hat{h} = \hat{f}(\cdot, \hat{g}) : \mathcal{X} \to \mathcal{Y}$ that achieves low risk $\mathbb{E}[\ell_{\hat{h}}(X,Y)]$.

To that end, we seek to bound the *excess risk*:

$$\mathbb{E}_P[\ell_{\hat{h}}(X,Y) - \ell_{h^*}(X,Y)], \tag{1}$$

for $h^* = f^*(\cdot, g^*)$ where g^* and f^* are given by

$$g^* \in \operatorname{argmin}_{g \in \mathcal{G}} \mathbb{E}[\ell_g^{\text{weak}}(X,W)],$$
$$f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}[\ell_{f(\cdot,g^*)}(X,Y)].$$

The comparison of $\hat{h}$ to h^* based on the best embedding g^* for the weak task is the most natural one for the pretrain-finetune algorithm that we analyze (see Algorithm 1).

We are interested in studying the *rate* at which the excess risk (1) goes to zero. Specifically, we are interested in studying the *learning rate* parameter γ for which the excess risk is $\mathcal{O}(n^{-\gamma})$. We refer to $\gamma \leq 1/2$ as a *slow rate*, and $\gamma \geq 1$ as a *fast rate* (possibly ignoring logarithmic factors, i.e., $\widetilde{\mathcal{O}}(1/n)$). When $1/2 < \gamma < 1$ we have *intermediate rates*.

Algorithm 1 Pretrain-finetune meta-algorithm

input $\mathcal{D}_m^{\text{weak}}$, $\mathcal{D}_n^{\text{strong}}$, $\mathcal{F}$, $\mathcal{G}$
Obtain weak embedding $\hat{g} \leftarrow \text{Alg}_m(\mathcal{G}, P_{X,W})$
Form dataset $\mathcal{D}_n^{\text{aug}} = \{(x_i, z_i, y_i)\}_{i=1}^n$ where $z_i := \hat{g}(x_i)$ for $(x_i, y_i) \in \mathcal{D}_n^{\text{strong}}$
Define distribution $\hat{P}(X,Z,Y) = P(X,Y)\mathbb{1}\{Z = \hat{g}(X)\}$
Obtain strong predictor $\hat{f} \leftarrow \text{Alg}_n(\mathcal{F}, \hat{P})$
return $\hat{h}(\cdot) := \hat{f}(\cdot, \hat{g}(\cdot))$

2.1. Pretrain-finetune meta-algorithm

The algorithm we analyze solves two supervised learning problems in sequence. The first step runs an algorithm,

$$\hat{g} \leftarrow \text{Alg}_m(\mathcal{G}, P_{X,W})$$

on m i.i.d. observations from $P_{X,W}$, and outputs a feature map $\hat{g} \in \mathcal{G}$. Using the resulting $\hat{g}$ we form an augmented dataset $\mathcal{D}_n^{\text{aug}} = \{(x_i, z_i, y_i)\}_{i=1}^n$, where $z_i := \hat{g}(x_i)$ for $(x_i, y_i) \in \mathcal{D}_n^{\text{strong}}$. Therewith, we have n i.i.d. samples from the distribution $\hat{P}(X,Z,Y) := P(X,Y)\mathbb{1}\{Z = \hat{g}(X)\}$. The second step then runs an algorithm,

$$\hat{f} \leftarrow \text{Alg}_n(\mathcal{F}, \hat{P})$$

on n i.i.d samples from $\hat{P}$, and outputs a strong predictor $\hat{f} \in \mathcal{F}$. The final output is then simply the composition $\hat{h} = \hat{f}(\cdot, \hat{g})$. This procedure is summarized in Algorithm 1 and the high level schema in Figure 1.

Algorithm 1 is generic because in general the two supervised learning steps can use any learning algorithm. Our analysis treats the case where $\text{Alg}_n(\mathcal{F}, \hat{P})$ is empirical risk minimization (ERM) but is *agnostic* to the choice of learning algorithm $\text{Alg}_m(\mathcal{G}, P_{X,W})$. Our results use high level properties of these two steps, in particular their generalization error, which we introduce next.

We break the generalization analysis into two terms depending on the bounds for each of the two supervised learning steps. We introduce here the notation $\text{Rate}(\cdot)$ to enable a more convenient discussion of these rates. We describe our notation in the format of definitions to expedite the statement of the theoretical results in Section 3.

Definition 1 (Weak learning)**.** Let $\text{Rate}_m(\mathcal{G}, P_{X,W}; \delta)$ be such that a (possibly randomized) algorithm $\text{Alg}_m(\mathcal{G}, P_{X,W})$ that takes as input a function class $\mathcal{G}$ and m i.i.d. observations from $P_{X,W}$, returns a weak embedding $\hat{g} \in \mathcal{G}$ for which,

$$\mathbb{E}_P \ell_{\hat{g}}^{\text{weak}}(X,W) \leq \text{Rate}_m(\mathcal{G}, P_{X,W}; \delta),$$

with probability at least $1 - \delta$.

We are interested in two particular cases of loss function ℓ^{weak}: (i) $\ell^{\text{weak}}(w,w') = \mathbb{1}\{w \neq w'\}$ when $\mathcal{W}$ is a categorical space; and (ii) $\ell^{\text{weak}}(w,w') = \|w - w'\|$ (for some norm $\|\cdot\|$ on $\mathcal{W}$) when $\mathcal{W}$ is a continuous space.

Definition 2 (Strong learning)**.** Let $\text{Rate}_n(\mathcal{F}, Q; \delta)$ be such that a (possibly randomized) algorithm $\text{Alg}_n(\mathcal{F}, Q)$ that takes as input a function space $\mathcal{F}$, and n i.i.d. observations from a distribution $Q(\mathcal{X} \times \mathcal{Z} \times \mathcal{Y})$, returns a strong predictor $\hat{f} \in \mathcal{F}$ for which,

$$\mathbb{E}_{U\sim Q}\Big[\ell_{\hat{f}}(U) - \ell_{f^*}(U)\Big] \leq \text{Rate}_n(\mathcal{F}, Q; \delta)$$

with probability at least $1-\delta$.

Henceforth, we drop δ from the rate symbols, for example writing $\text{Rate}_m(\mathcal{G}, P_{X,W})$ instead of $\text{Rate}_m(\mathcal{G}, P_{X,W}; \delta)$.

It is important to note that the algorithms $\text{Alg}_m(\mathcal{G}, P_{X,W})$ and $\text{Alg}_n(\mathcal{F}, Q)$ can use any loss functions during training. In particular, even though we assume ℓ^{weak} to be a metric, it is possible to use non-metric losses such as cross entropy during training. This is because the only requirement we place on these algorithms is that they imply generalization bounds in terms of the losses ℓ^{weak} and ℓ respectively. For concreteness, our analysis focuses the case where $\text{Alg}_n(\mathcal{F}, Q)$ is ERM using loss ℓ.

3. Excess Risk Analysis

In this section we analyze Algorithm 1 with the objective of obtaining high probability excess risk bounds (see (1)) for the strong predictor $\hat{h} = \hat{f}(\cdot, \hat{g})$. Informally, the main theorem we prove is the following.

Theorem 3 (Informal)**.** *Suppose that $\text{Rate}_m(\mathcal{G}, P_{X,W}) = \mathcal{O}(m^{-\alpha})$ and that $\text{Alg}_n(\mathcal{F}, \hat{P})$ is ERM. Under suitable assumptions on $(\ell, P, \mathcal{F})$, Algorithm 1 obtains excess risk,*

$$\mathcal{O}\Big(\frac{\alpha\beta \log n + \log(1/\delta)}{n} + \frac{1}{n^{\alpha\beta}}\Big)$$

with probability $1-\delta$, when $m = \Omega(n^{\beta})$ for $\mathcal{W}$ discrete, or $m = \Omega(n^{2\beta})$ for $\mathcal{W}$ continuous.

For the prototypical scenario where $\text{Alg}_m(\mathcal{G}, P_{X,W}) = \mathcal{O}(1/\sqrt{m})$, one obtains fast rates when $m = \Omega(n^2)$, and $m = \Omega(n^4)$, in the discrete and continuous cases, respectively. More generally, if $\alpha\beta < 1$ then $\mathcal{O}(n^{-\alpha\beta})$ is the dominant term and we observe intermediate or slow rates.

In order to obtain any such result, it is necessary to quantify how the weak and strong tasks relate to one another – if they are completely unrelated, then there is no reason to expect the representation $\hat{g}(X)$ to benefit the strong task. The next subsection addresses this question.

3.1. Relating weak and strong tasks

Next, we formally quantify the relaship between the weak and strong task, via two concepts. First, we assume that the two tasks share a mutual embedding $Z = g_0(X)$. Alone, this is not enough, since otherwise one could simply take the trivial embedding $X = g_0(X)$, which will not inform the strong task. The *central condition* that we introduce in this section quantifies how the embedding makes the strong task "easy". Second, we assume a shared stability via a "relative Lipschitz" property: small perturbations to the feature map g that do not hurt the weak task, do not affect the strong prediction loss much either.

Definition 4. We say that f is L-Lipschitz relative to $\mathcal{G}$ if for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $g, g' \in \mathcal{G}$,

$$|\ell_{f(\cdot,g)}(x,y) - \ell_{f(\cdot,g')}(x,y)| \leq L\ell^{\text{weak}}(\beta_g^\top g(x), \beta_{g'}^\top g'(x))).$$

We say the function class $\mathcal{F}$ is L-Lipschitz relative to $\mathcal{G}$, if every $f \in \mathcal{F}$ is L-Lipschitz relative to $\mathcal{G}$.

The Lipschitz terminology is justified since the domain uses the pushforward pseudometric $(z, z') \mapsto \ell^{\text{weak}}(A_g^\top z, A_{g'}^\top z')$, and the range is a subset of $\mathbb{R}_+$. In the special case where $\mathcal{Z} = \mathcal{W}$, $g(X)$ is actually an estimate of the weak label W and relative Lipschitzness reduces to conventional Lipschitzness of $\ell(f(x,w),y)$ in w.

The central condition is well-known to yield fast rates for supervised learning (van Erven et al., 2015); it directly implies that we could learn a map $(X, Z) \mapsto Y$ with $\widetilde{\mathcal{O}}(1/n)$ excess risk. The difficulty with this naive view is that at test time we would need access to the latent value $Z = g_0(X)$, an implausible requirement. To circumnavigate this hurdle, we replace g_0 with $\hat{g}$ by solving the supervised problem $(\ell, \hat{P}, \mathcal{F})$, for which we will have access to data.

But it is not clear whether this surrogate problem would continue to satisfy the central condition. One of our main theoretical contributions is to show that $(\ell, \hat{P}, \mathcal{F})$ indeed satisfies a weak central condition (Theorems 7 and 8), and to show that this weak central condition still enables strong excess risk guarantees (Theorem 9).

We are now ready to define the central condition. In essence, this condition requires that (X, Z) is highly predictive of Y, which, combined with the fact that $g_0(X) = Z$ has zero risk on W, links the weak and strong tasks together.

Definition 5 (The Central Condition)**.** A learning problem $(\ell, P, \mathcal{F})$ on $\mathcal{U} := \mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$ is said to satisfy the ε-*weak η-central condition* if there exists an $f^* \in \mathcal{F}$ such that

$$\mathbb{E}_{U\sim P(\mathcal{U})}[e^{-\eta(\ell_f(U) - \ell_{f^*}(U))}] \leq e^{\eta\varepsilon},$$

for all $f \in \mathcal{F}$. The 0-weak central condition is known as the strong central condition.

We assume that the strong central condition holds for our weakly supervised problem $(\ell, P, \mathcal{F})$ with $P = P_{\mathcal{U}} = P_{X,Z,Y}$ where $Z = g_0(X)$. A concrete but general example of a class of weakly supervised problems satisfying the shared embedding assumption and central condition are those where weak and strong labels share a common

latent embedding Z, and Y is a logistic model on Z. In detail let $Z = g_0(X)$ be an arbitrary embedding of input X, and let W be a deterministic function of Z. Suppose also that $Y = \sigma(A^*Z) = f^*(Z)$ for some matrix A^*, where σ denotes the Softmax function. Then, as observed by Foster et al. (2018), the learning problem $(\ell, P, \mathcal{F})$ is Vovk mixable, and hence the central condition holds (van Erven et al., 2015), where ℓ is the logistic loss, and $\mathcal{F} = \{Z \mapsto \sigma(AZ) : A \in \mathbb{R}^{|\mathcal{Y}| \times s}\}$.

It is important to note that if one knows that a problem $(\ell, P, \mathcal{F})$ on $\mathcal{U} := \mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$ satisfies the central condition, then it is not a priori clear if one can construct a hypothesis set $\widetilde{\mathcal{F}} \subseteq \{\mathcal{X} \to \mathcal{Y}\}$ such that $(\ell, P_{X,Y}, \widetilde{\mathcal{F}})$ satisfies the central condition. In the standard supervised setting with samples only from $P_{X,Y}$ this is likely impossible in general. This is because in the later case the feature Z is no longer an input to the model and so potentially valuable predictive features are lost. However, one perspective on the analysis in this section is that we show that with the added support of samples from $P_{X,W}$, the hypothesis class $\widetilde{\mathcal{F}} = \{f(\cdot, \hat{g}(\cdot)) : f \in \mathcal{F}\}$, where $\hat{g}(X)$ is learned using the weak labeled samples, does indeed satisfy the central condition with a slightly larger ε.

The central condition and related literature. The central condition unifies many well-studied conditions known to imply fast rates (van Erven et al., 2015), including Vapnik and Chervonenkis' original condition, that there is an $f^* \in \mathcal{F}$ with zero risk (Vapnik & Chervonenkis, 1971; 1974). The popular strong-convexity condition (Kakade & Tewari, 2009; Lecué et al., 2014) is also a special case, as is (stochastic) exponential concavity, which is satisfied by density estimation: where $\mathcal{F}$ are probability densities, and $\ell_f(u) = -\log f(u)$ is the logarithmic loss (Audibert et al., 2009; Juditsky et al., 2008; Dalalyan et al., 2012). Another example is Vovk mixability (Vovk, 1990; 1998), which holds for online logistic regression (Foster et al., 2018), and also holds for uniformly bounded functions with the square loss. A modified version of the central condition also generalizes the Bernstein condition and Tsybakov's margin condition (Bartlett & Mendelson, 2006; Tsybakov et al., 2004).

As noted earlier, Z is not observable at train or test time, so we cannot simply treat the problem as a single supervised learning problem. Therefore, obtaining fast or intermediate rates is a nontrivial challenge. We approach this challenge by splitting the learning procedure into two supervised tasks (Algorithm 1). In its second step, Algorithm 1 replaces $(\ell, P, \mathcal{F})$ with $(\ell, \hat{P}, \mathcal{F})$. Our strategy to obtain generalization bounds is first to guarantee that $(\ell, \hat{P}, \mathcal{F})$ satisfies the weak central condition, and then to show that the weak central condition implies the desired generalization guarantees.

The rest of this section develops the theoretical machinery needed for obtaining our bounds. We summarize the key steps of our argument below.

1. Decompose the excess risk into two components: the excess risk of the weak predictor and the excess risk on the learning problem $(\ell, \hat{P}, \mathcal{F})$ (Proposition 6).
2. Show that the learning problem $(\ell, \hat{P}, \mathcal{F})$ satisfies a relaxed version of the central condition - the "weak central condition" (Propositions 7 and 8).
3. Show that the ε-weak central condition yields excess risk bounds that improve as ε decreases (Prop. 9).
4. Combine all previous results to obtain generalization bounds for Algorithm 1 (Theorem 10).

3.2. Generalization Bounds for Weakly Supervised Learning

The first item on the agenda is Proposition 6 which obtains a generic bound on the excess risk in terms of $\text{Rate}_m(\mathcal{G}, P_{X,W})$ and $\text{Rate}_n(\mathcal{F}, \hat{P})$.

Proposition 6 (Excess risk decomposition)**.** *Suppose that f^* is L-Lipschitz relative to $\mathcal{G}$. Then the excess risk $\mathbb{E}[\ell_{\hat{h}}(X,Y) - \ell_{h^*}(X,Y)]$ is bounded by,*

$$2L\mathit{Rate}_m(\mathcal{G}, P_{X,W}) + \mathit{Rate}_n(\mathcal{F}, \hat{P}).$$

The first term corresponds to excess risk on the weak task, which we expect to be small since that environment is data-rich. Hence, the problem of obtaining excess risk bounds reduces to bounding the second term, $\text{Rate}_n(\mathcal{F}, \hat{P})$. This second term is much more opaque; we spend the rest of the section primarily analyzing it.

We now prove that if $(\ell, P, \mathcal{F})$ satisfies the ε-weak central condition, then the artificial learning problem $(\ell, \hat{P}, \mathcal{F})$ obtained by replacing the true population distribution P with the estimate $\hat{P}$ satisfies a slightly weaker central condition. We consider the categorical and continuous $\mathcal{W}$-space cases separately, obtaining an improved rate in the categorical case. In both cases, the proximity of this weaker central condition to the ε-weak central condition is governed by $\text{Rate}_m(\mathcal{G}, P_{X,W})$, but the dependencies are different.

Proposition 7 (Categorical weak label)**.** *Suppose that $\ell^{weak}(w, w') = \mathbb{1}\{w \neq w'\}$ and that ℓ is bounded by $B > 0$, $\mathcal{F}$ is Lipschitz relative to $\mathcal{G}$, and that $(\ell, P, \mathcal{F})$ satisfies the ε-weak central condition. Then $(\ell, \hat{P}, \mathcal{F})$ satisfies the $\varepsilon + \mathcal{O}\left(e^B \mathit{Rate}_m(\mathcal{G}, P_{X,W})\right)$-weak central condition with probability at least $1 - \delta$.*

Next, we consider the norm induced loss. In this case it is also possible to obtain obtain the weak central condition for the artificially augmented problem $(\ell, \hat{P}, \mathcal{F})$.

Proposition 8 (Continuous weak label)**.** *Suppose that* $\ell^{weak}(w,w') = \|w - w'\|$ *and that* ℓ *is bounded by* $B > 0$*,* $\mathcal{F}$ *is L-Lipschitz relative to* $\mathcal{G}$*, and that* $(\ell, P, \mathcal{F})$ *satisfies the* ε*-weak central condition. Then* $(\ell, \hat{P}, \mathcal{F})$ *satisfies the* $\varepsilon + \mathcal{O}\Big(\sqrt{Le^B Rate_m(\mathcal{G}, P_{X,W})}\Big)$*-weak central condition with probability at least* $1-\delta$*.*

For both propositions, a slight modification of the proofs easily eliminates the e^B term when $\text{Rate}_m(\mathcal{G}, P_{X,W}) \leq \mathcal{O}(e^{-B})$. Since we typically consider the regime where $\text{Rate}_m(\mathcal{G}, P_{X,W})$ is close to zero, Propositions 7 and 8 essentially say that replacing P by $\hat{P}$ only increases the weak central condition parameter slightly.

The next, and final, step in our argument is to obtain a generalization bound for ERM under the ε-weak central condition. Once we have this bound, one can obtain good generalization bounds for the learning problem $(\ell, \hat{P}, \mathcal{F})$ since the previous two propositions guarantee that it satisfies the weak central condition from some small ε. Combining this observation with the results from the previous section finally allows us to obtain generalization bounds on Algorithm 1 when $\text{Rate}_n(\mathcal{F}, \hat{P})$ is ERM.

For this final step, we assume that our strong predictor class $\mathcal{F}$ is parameterized by a vector in $\mathbb{R}^d$, and identify each f with this parameter vector. We also assume that the parameters live in an L_2 ball of radius R. By Lagrangian duality this is equivalent to our learning algorithm being ERM with L_2-regularization for some regularization parameter.

Proposition 9. *Suppose* $(\ell, Q, \mathcal{F})$ *satisfies the* ε*-weak central condition,* ℓ *is bounded by* $B > 0$*, each* $\mathcal{F}$ *is* L'*-Lipschitz in its parameters in the* ℓ_2 *norm,* $\mathcal{F}$ *is contained in the Euclidean ball of radius R, and* $\mathcal{Y}$ *is compact. Then when* $Alg_n(\mathcal{F}, Q)$ *is ERM, the excess risk* $\mathbb{E}_Q[\ell_{\hat{f}}(U) - \ell_{f^*}(U)]$ *is bounded by,*

$$\mathcal{O}\Big(V\frac{d\log(RL'/\varepsilon) + \log(1/\delta)}{n} + V\varepsilon\Big),$$

with probability at least $1-\delta$*, where* $V = B + \varepsilon$*.*

Any parameterized class of functions that is continuously differentiable in its parameters satisfies the L'-Lipschitz requirement since we assume the parameters live in a closed ball of radius R. The $\mathcal{Y}$ compactness assumption can be dropped in the case where $y \mapsto \ell(y, \cdot)$ is Lipschitz.

Observe that the bound in Proposition 9 depends linearly on d, the number of parameters of $\mathcal{F}$. Since we consider the regime where n is small, the user might use only a small model (e.g., a shallow network) to parameterize $\mathcal{F}$, so d may not be too large. On the other hand, the bound is independent of the complexity of $\mathcal{G}$. This is important since the user may want to use a powerful model class for g to profit from the bountiful amounts of weak labels.

Proposition 9 gives a generalization bound for any learning problem $(\ell, Q, \mathcal{F})$ satisfying the weak central condition, and may therefore be of interest in the theory of fast rates more broadly. For our purposes, however, we shall apply it only to the particular learning problem $(\ell, \hat{P}, \mathcal{F})$. In this case, the ε shall depend on $\text{Rate}_m(\mathcal{G}, P_{X,W})$, yielding strong generalization bounds when $\hat{g}$ has low excess risk.

Combining Proposition 9 with both of the two previous propositions yields fast rates guarantees (Theorem 10) for the double estimation algorithm (Algorithm 1) for ERM. The final bound depends on the rate of learning for the weak task, and on the quantity of weak data available m.

Theorem 10 (Main result)**.** *Suppose the assumptions of Proposition 9 hold,* $(\ell, P, \mathcal{F})$ *satisfies the central condition, and that* $Rate_m(\mathcal{G}, P_{X,W}) = \mathcal{O}(m^{-\alpha})$*. Then, when* $Alg_n(\mathcal{F}, \hat{P})$ *is ERM we obtain excess risk* $\mathbb{E}_P[\ell_{\hat{h}}(X,Y) - \ell_{h^*}(X,Y)]$ *that is bounded by,*

$$\mathcal{O}\Big(\frac{d\alpha\beta\log RL'n + \log\frac{1}{\delta}}{n} + \frac{L}{n^{\alpha\beta}}\Big),$$

with probability at least $1-\delta$*, if either of the following conditions hold,*

1. $m = \Omega(n^\beta)$ *and* $\ell^{weak}(w,w') = \mathbb{1}\{w \neq w'\}$ *(discrete* $\mathcal{W}$*-space).*
2. $m = \Omega(n^{2\beta})$ *and* $\ell^{weak}(w,w') = \|w - w'\|$ *(continuous* $\mathcal{W}$*-space).*

To reduce clutter we absorb the dependence on B into the big-$\mathcal{O}$. The key quantities governing the ultimate learning rate are α, the learning rate on the weak task, and β, which determines the amount of weak labels relative to strong.

4. Experiments

We experimentally study two types of weak labels: noisy, and coarse. We study two cases: when the amount of weak data grows linearly with the amount of strong data, and when the amount of weak data grows quadratically with the amount of strong data (plus a baseline). Note that in a log-log plot the *negative of the gradient* is the learning rate γ such that excess risk is $\mathcal{O}(n^{-\gamma})$. All image-based experiments use either a ResNet-18 or ResNet-34 for the weak feature map g (see Appendix C for full details).

4.1. Noisy Labels

We simulate a noisy labeler who makes labelling mistakes in a data dependent way (as opposed to independent random noise) by training an auxiliary deep network on a held out dataset to classify at a certain accuracy - for our CIFAR-10 experiments we train to 90% accuracy. This is intended to mimic human annotators working on a crowd sourcing platform. The predictions of the auxiliary network are used

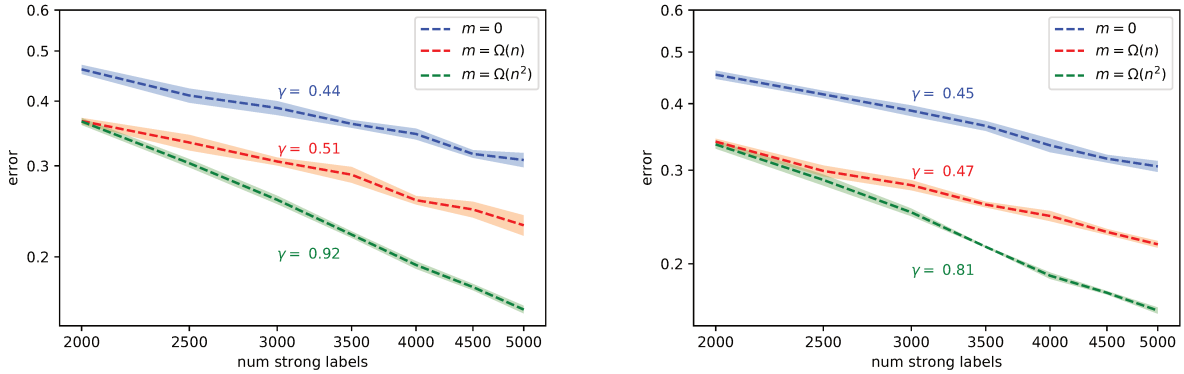

Figure 2. Generalization error on CIFAR-10 using noisy weak labels for different growth rates of m. Left hand diagram is for simulated "noisy labeler", the right hand picture is for random noise.

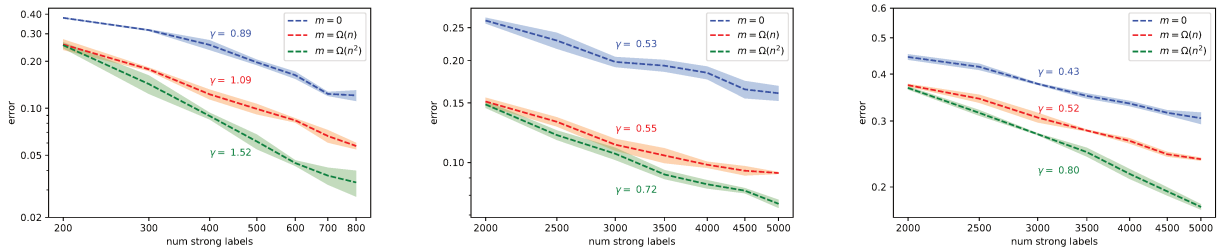

Figure 3. Coarse labels. Generalization error on various datasets using coarse weak label grouping for different growth rates of m. Datasets left to right: MNIST, SVHN, and CIFAR-10.

as weak labels. We also run experiments using independent random noise, flipping each label randomly with 10% chance. See Figure 2 for results.

In each case, both the generalization error when using additional weak data is lower, *and* the learning rate itself is higher. Indeed, the learning rate improvement is significant. For simulated noisy labels, $\gamma = 0.44$ when $m = 0$, and $\gamma = 0.92$ for $m = \Omega(n^2)$. Random noisy labels has a similar result with $\gamma = 0.45$ and $\gamma = 0.81$ for $m = 0$, and $m = \Omega(n^2)$ respectively. The experimental results are in line with our theoretical result that the learning rate should double when β doubles.

4.2. Coarse labels

We also consider fine tuning on a small subset CIFAR-100 with one of 100 categories. For weak pretraining we use a larger dataset of examples labeled with 20 semantically meaningful "super-categories". Each super category contains exactly 5 of the 100 fine grained categories. For example, the categories"maple", "oak", "palm", "pine", and "willow" are all part of the super-category"trees". The results are presented in Figure 4. In the natural language domain we consider the TREC fast-based question categorization dataset. Similarly to CIFAR-100, examples are naturally divided up into six coarse groups of questions concerning numerics, humans, locations etc. They are further divided up into 50 fine grained classes, used as strong labels. Since we observed the natural language experiments to be much noisier than the vision tasks, we ran 20 repeats to get a reliable average to observe the central tendency.

Again, generalization error is consistently lower, and learning rate constantly high for larger m growth rate. The differences are generally very significant, e.g. for CIFAR-100 where top-5 accuracy learning rate is $\gamma = 0.40$ for $m = 0$, and $\gamma = 0.82$ for $m = \Omega(n^2)$, and for MNIST $\gamma = 0.89$ and $\gamma = 1.52$ for $m = 0$ and $m = \Omega(n^2)$ respectively. The contrast is slightly less pronounced on the TREC experiments, but the same broad trend is observed.

5. Related Work

Weakly supervised learning. There exists previous work on the case where one *only* has weak labels. Khetan et al. (2018) consider crowd sourced labels and use an EM-style

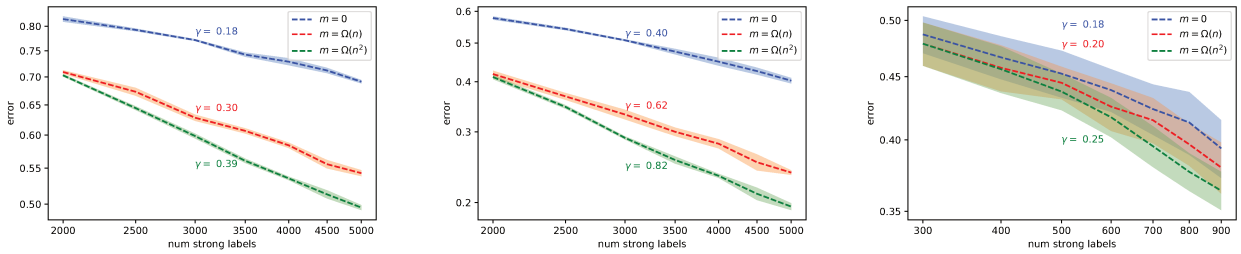

Figure 4. Left and middle: Generalization error on CIFAR-100 using coarse weak labels for different growth rates of m. Left diagram is top-1 accuracy, and middle diagram is top-5 accuracy. Right: top-1 error for the TREC question sentiment analysis task.

algorithm to model the quality of individual workers. Another approach proposed by Ratner et al. (2016; 2019) uses correlation between multiple different weak label sources to estimate the ground truth label. A different approach is to use pairwise semantic (dis)similarity as a form of weak signal about unlabeled data (Arora et al., 2019; Bao et al., 2018) or to use complementary labels, which give you a label telling you a class that the input is *not* in (Xu et al., 2019; Ishida et al., 2017).

Fast rates. There is a large body of work studying a variety of favorable situations under which it is possible to obtain rates better than slow-rates. From a generalization and optimization perspective, strongly convex losses enable fast rates for generalization and for fast convergence of stochastic gradient (Kakade & Tewari, 2009; Hazan et al., 2007; Foster & Syrgkanis, 2019). These works are special cases of exponentially concave learning, which is itself a special case of the central condition. There are completely different lines of work on fast rates, such as developing data dependent local Rademacher averages (Bartlett et al., 2005); and herding, which has been used to obtain fast rates for integral approximation (Welling, 2009).

Learning with a nuisance component. The two-step estimation algorithm we study in this paper is closely related to statistical learning under a nuisance component (Chernozhukov et al., 2018; Foster & Syrgkanis, 2019). In that setting one wishes to obtain excess risk bounds for the model $\hat{f}(\cdot, g_0(\cdot))$ where $W = g_0(X)$ is the true weak predictor. The analysis of learning in such settings rests crucially on the Neyman orthogonality assumption (Neyman & Scott, 1965). Our setting has the important difference of seeking excess risk bounds for the compositional model $\hat{f}(\cdot, \hat{g}(\cdot))$.

Self-supervised learning. In self-supervised learning the user artificially constructs pretext learning problems based on attributes of unlabeled data (Doersch et al., 2015; Gidaris et al., 2018). In other words, it is often possible to construct a weakly supervised learning problem where the choice of weak labels are a design choice of the user. In line with our analysis, the success of self-supervised representations relies on picking pretext labels that capture useful information about the strong label such as invariances and spacial understanding (Noroozi & Favaro, 2016; Misra & van der Maaten, 2019). Conversely, weakly supervised learning can be viewed as a special case of self-supervision where the pretext task is selected from some naturally occurring label source (Jing & Tian, 2019).

6. Discussion

Our work focuses on analyzing weakly supervised learning. We believe, however, that the same framework could be used to analyze other popular learning paradigms. One immediate extension of our analysis would be to multiple inconsistent sources of weak labels as by Khetan et al. (2018).

Other important extensions would be to include self-supervised pretraining. Cases where the marginal $P(X)$ does not shift fall within the scope of our analysis. However, a key technical difference between our setting and approaches such as image super-resolution (Ledig et al., 2017), solving jigsaw puzzles (Noroozi & Favaro, 2016), and image inpaining (Pathak et al., 2016) is that in the latter, the marginal distribution of features $P(X)$ is potentially different on the pretext task as compared to the downstream tasks of interest. Because of this difference our analysis doesn't immediately transfer over to these settings, leaving an interesting avenue for future work.

Another option is to use our representation transfer analysis to study multi-task or meta-learning settings where one wishes to reuse an embedding across multiple tasks with shared characteristics with the aim of obtaining certified performance across all tasks.

A completely different direction, based on the observation that our analysis is predicated on the idea of "cheap" weak labels and "costly" strong labels, is to ask how best to allocate a finite budget for label collection when faced with varying quality label sources.

Acknowledgements This work was supported by NSF TRIPODS+X grant (DMS-1839258), NSF-BIGDATA (IIS-1741341), and the MIT-MSR TRAC collaboration. The authors thank Ching-Yao Chuang for discussions.

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. In *Int. Conference on Machine Learning (ICML)*, 2019.

Audibert, J.-Y. et al. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4), 2009.

Bao, H., Niu, G., and Sugiyama, M. Classification from pairwise similarity and unlabeled data. 2018.

Bartlett, P. L. and Mendelson, S. Empirical minimization. *Probability theory and related fields*, 135(3), 2006.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. Local Rademacher complexities. *The Annals of Statistics*, 33(4), 2005.

Bilen, H. and Vedaldi, A. Weakly supervised deep detection networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Branson, S., Van Horn, G., and Perona, P. Lean crowdsourcing: Combining humans and machines in an online system. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Carl, B. and Stephani, I. *Entropy, compactness and the approximation of operators*. Number 98. Cambridge University Press, 1990.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Chen, J. and Batmanghelich, K. Weakly supervised disentanglement by pairwise similarities. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. Boosting for transfer learning. In *Int. Conference on Machine Learning (ICML)*, 2007.

Dalalyan, A. S., Tsybakov, A. B., et al. Mirror averaging with sparsity priors. *Bernoulli*, 18(3), 2012.

Desrosiers, C. and Karypis, G. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender systems handbook*. Springer, 2011.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Int. Conference on Computer Vision (ICCV)*, 2015.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *Int. Conference on Machine Learning (ICML)*, 2014.

Durand, T., Mordan, T., Thome, N., and Cord, M. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. In *Conference on Learning Theory (COLT)*, 2019.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Logistic regression: The importance of being improper. In *Conference on Learning Theory (COLT)*, 2018.

Fries, J. A., Varma, P., Chen, V. S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M., et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. *Nature communications*, 10(1), 2019.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *preprint arXiv:1803.07728*, 2018.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Guo, Y., Liu, Y., Bakker, E. M., Guo, Y., and Lew, M. S. CNN-RNN: A large-scale hierarchical image classification framework. *Multimedia Tools and Applications*, 77(8), 2018.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3), 2007.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Huh, M., Agrawal, P., and Efros, A. A. What makes ImageNet good for transfer learning? *preprint arXiv:1608.08614*, 2016.

Ishida, T., Niu, G., Hu, W., and Sugiyama, M. Learning from complementary labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jing, L. and Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *preprint arXiv:1902.06162*, 2019.

Juditsky, A., Rigollet, P., Tsybakov, A. B., et al. Learning by mirror averaging. *The Annals of Statistics*, 36(5), 2008.

Kakade, S. M. and Tewari, A. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009.

Khetan, A., Lipton, Z. C., and Anandkumar, A. Learning from noisy singly-labeled data. *Int. Conf. on Learning Representations (ICLR)*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Int. Conf. on Learning Representations (ICLR)*, 2015.

Kleindessner, M. and Awasthi, P. Crowdsourcing with arbitrary adversaries. In *Int. Conference on Machine Learning (ICML)*, 2018.

Lecué, G., Rigollet, P., et al. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1), 2014.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *cvpr*, 2017.

Li, W., Wang, L., Li, W., Agustsson, E., and van Gool, L. Webvision database: Visual learning and understanding from web data. In *preprint arXiv:1708.02862*, 2017.

Liao, X., Xue, Y., and Carin, L. Logistic regression with an auxiliary data source. In *Int. Conference on Machine Learning (ICML)*, 2005.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M.,

Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Europ. Conference on Computer Vision (ECCV)*, 2018.

Medlock, B. and Briscoe, T. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007.

Mehta, N. A. Fast rates with high probability in exp-concave statistical learning. In *Proc. Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

Misra, I. and van der Maaten, L. Self-supervised learning of pretext-invariant representations. *preprint arXiv:1912.01991*, 2019.

Naumov, M., Mudigere, D., Shi, H.-J. M., Huang, J., Sundaraman, N., Park, J., Wang, X., Gupta, U., Wu, C.-J., Azzolini, A. G., et al. Deep learning recommendation model for personalization and recommendation systems. *preprint arXiv:1906.00091*, 2019.

Neyman, J. and Scott, E. L. Asymptotically optimal tests of composite hypotheses for randomized experiments with noncontrolled predictor variables. *Journal of the American Statistical Association*, 60(311), 1965.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Europ. Conference on Computer Vision (ECCV)*. Springer, 2016.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 2009.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *cvpr*, 2016.

Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., and Ré, C. Training complex models with multi-task weak supervision. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 33, 2019.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Ricci, F., Rokach, L., and Shapira, B. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 2011.

Schork, N. J. Personalized medicine: time for one-person trials. *Nature*, 520(7549), 2015.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Int. Conference on Computer Vision (ICCV)*, 2017.

Taherkhani, F., Kazemi, H., Dabouei, A., Dawson, J., and Nasrabadi, N. M. A weakly supervised fine label classifier enhanced by coarse supervision. In *Int. Conference on Computer Vision (ICCV)*, 2019.

Tsybakov, A. B. et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1), 2004.

van Erven, T., Grünwald, P., Reid, M., and Williamson, R. Mixability in statistical learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

van Erven, T., Grunwald, P., Mehta, N. A., Reid, M., Williamson, R., et al. Fast rates in statistical and online learning. In *Journal of Machine Learning Research*, 2015.

Vapnik, V. and Chervonenkis, A. Theory of pattern recognition, 1974.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2), 1971.

Vovk, V. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2), 1998.

Vovk, V. G. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

Wan, F., Wei, P., Jiao, J., Han, Z., and Ye, Q. Min-entropy latent model for weakly supervised object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., and Liu, H. A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making*, 19(1), 2019.

Welling, M. Herding dynamical weights to learn. In *Int. Conference on Machine Learning (ICML)*, 2009.

Xu, Y., Gong, M., Chen, J., Liu, T., Zhang, K., and Batmanghelich, K. Generative-discriminative complementary learning. In *preprint arXiv:1904.01612*, 2019.

Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W., and Yu, Y. HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition. In *Int. Conference on Computer Vision (ICCV)*, 2015.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional neural networks. In *Europ. Conference on Computer Vision (ECCV)*, 2014.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Zhao, B., Li, F., and Xing, E. P. Large-scale category structure aware image categorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 2018.