Warm-starting Contextual Bandits: Robustly Combining Supervised and Bandit Feedback

Chicheng Zhang [1] **Alekh Agarwal** [1] **Hal Daumé III** [1 2] **John Langford** [1] **Sahand N Negahban** [3]

Abstract

We investigate the feasibility of learning from a mix of both fully-labeled supervised data and contextual bandit data. We specifically consider settings in which the underlying learning signal may be different between these two data sources. Theoretically, we state and prove no-regret algorithms for learning that is robust to misaligned cost distributions between the two sources. Empirically, we evaluate some of these algorithms on a large selection of datasets, showing that our approach is both feasible, and helpful in practice.

1. Introduction

In many real-world settings, a system must learn from multiple types of feedback; we consider the specific setting of learning jointly from fully labeled "supervised" examples and from online feedback "contextual bandit" (abbrev. CB) examples. For instance, in a system that chooses personalized content to display on a webpage, an expert may be able to provide an initial set of fully labeled examples to get a system started. After deployment, however, the system can only measure its performance (e.g., dwell time) on the content it displays and not other (counterfactual) options. In an automated translation system, professional translators can provide initial translations to seed a system, but the system may be able to further improve its performance based on, e.g., user satisfaction measures (Sokolov et al., 2015; Nguyen et al., 2017).

In both these settings (content display and translation), we desire an approach that is able to use the fully supervised expert data to "warm-start" a system, which later learns from CB feedback (Auer et al., 2002b; Langford & Zhang, 2007; Chu et al., 2011; Dudik et al., 2011; Agrawal & Goyal, 2013; Agarwal et al., 2014). Doing so has the added advantage of ensuring that such a system does not need to suffer too much error in an initial exploration phase, which may be necessary in user-facing systems or in error- or safety-critical settings (Tewari & Murphy, 2017). However, it is generally unreasonable to assume that the expert supervision and the CB feedback in such settings are perfectly aligned: the "best" decision according to an expert may not necessarily match a user's choice. We need algorithms that operate well even in the case of unknown degrees of misalignment; we introduce a hypothesis class-specific notion of cost similarity used in our analysis, but not our algorithms (§2). We also highlight how simple strategies for combining the two sources without robustness to misalignment can perform significantly worse than learning from the ground truth source alone (§2.1).

Furthermore, different applications can differ in terms of which source—supervised or CB—is considered "ground truth". For example, while the CB feedback from users is the better signal about their preferences in content personalization (§3), the expert translations provide the ground truth in the translation setting for which user satisfaction is an imperfect proxy (§4). We develop algorithms for *both* settings, which effectively "search" for a good balance between fitting the CB feedback and supervised labels. In both cases, we provide regret bounds showing the value of the complementary data sources, dependent on their cost discrepancy and respective sample sizes. Importantly, our theory shows that our methods perform close to an oracle that knows the similarity of the two sources beforehand and uses it to optimally weight their examples, with a small additional penalty from searching for this weighting.

Empirically, we perform experiments based on fully-labeled examples from which CB feedback is simulated. We focus on the setting when CB data is ground truth and the supervised warm-start might have differing levels of bias. In an experimental study over hundreds of datasets (§5) we demonstrate the efficacy of our algorithm. As a snapshot, Figure 1 shows the empirical cumulative distribution functions (CDF) of algorithms across a number of experimental conditions, where each (x, y) value on the curve indicates that there is a y fraction of experimental conditions where the normalized error[1] of a method is below x. The plot ag-

[1]Microsoft Research [2]University of Maryland [3]Yale University. Correspondence to: Chicheng Zhang <Chicheng.Zhang@microsoft.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] See §5 for a formal definition of normalized error.

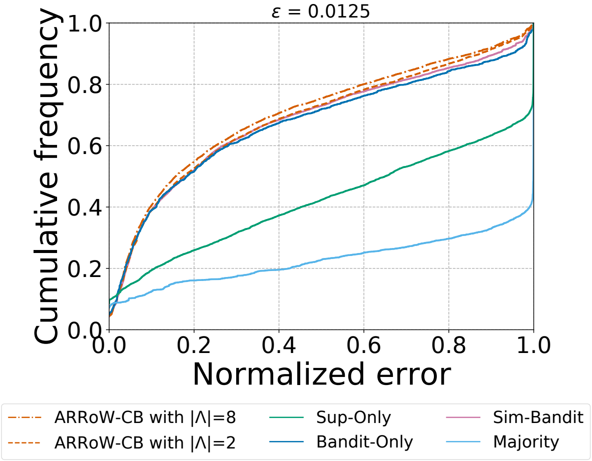

Figure 1: Empirical CDFs of the performance of different methods across a number of datasets and experimental conditions (See §5 for descriptions of all algorithms, settings, and aggregation method). Our method ARRoW-CB has a parameter $|\Lambda|$, and we evaluate it with $|\Lambda|$ set to 8 and 2; SIM-BANDIT is a baseline also leveraging the warm-start; SUP-ONLY, MAJORITY[1] and BANDIT-ONLY learn using only the supervised and CB sources respectively. All CB methods use the ϵ-greedy strategy with $\epsilon = 0.0125$.

gregates across settings where the CB and supervised signals are perfectly aligned as well as where they are not. Overall, our main algorithm, namely ARRoW-CB with $|\Lambda| = 8$, outperforms all baselines in this aggregated summary, in particular beating the two algorithms (ARRoW-CB with $|\Lambda| = 2$ and SIM-BANDIT) that leverage both CB and supervised sources. More detailed results are presented in §5.

Relation to prior work. A theoretical study of domain adaptation (Ben-David et al., 2010; Mansour et al., 2009) and learning from multiple sources (Crammer et al., 2008) are the closest prior works. In these works, all data sources provide the same supervised feedback rather than the supervised/CB modality we investigate here, with the two sources having very different information per sample. Another related line of work is on "safe" CB learning (Kazerouni et al., 2017; Sun et al., 2017) to maintain performance better than a baseline policy at all times, somewhat related to our supervised ground truth setting. However, they do not study the distributional mismatch concerns central to our work.

Finally, there is a substantial literature on active learning from different data sources (Donmez & Carbonell, 2008; Urner et al., 2012; Yan et al., 2011; Malago et al., 2014; Zhang & Chaudhuri, 2015; Yan et al., 2018), combining multiple labeling oracles of varying quality. The CB setting studied has important differences from active learning and the techniques do not carry over directly.

2. Notation and Problem Specification

We begin with some notation. For an event A, $I(A) = 1$ if A is true, and 0 otherwise. Denote by $[K]$ the set $\{1, 2, \ldots, K\}$. We use $\mathbb{1}_K$ to denote the all 1's vector in $\mathbb{R}^K$ and Δ^{K-1} for the K dimensional probability simplex.

In this paper, we study the problem of cost-sensitive interactive learning from multiple data sources. Specifically, we consider distributions over cost-sensitive examples (x, c), where $x \in \mathcal{X}$ is a context and c is a cost vector in $[0, 1]^K$; K being the number of actions (or "classes"). There are two distributions D^s (supervised) and D^b (CB), which have identical marginals over the context x, but different conditional distributions over cost vectors given x. We use the notation c^b (resp. c^s) to denote the cost vector c drawn from D^b (resp. D^s) to avoid writing D^b and D^s as subscripts in expectations. The interaction between the learner and the environment is described as follows:

Warm-start: The learner receives S, a dataset of n^s fully supervised examples drawn i.i.d. from D^s.

Interaction: For $t = 1, 2, \ldots, n^b$, the environment draws $(x_t, c^b_t) \sim D^b$ and reveals x_t to the learner, based on which the learner chooses a (possibly random) action $a_t \in [K]$ and observes $c^b_t(a_t)$, but not the cost of any other action.

In this paper, we focus on two learning settings: *CB ground truth setting* and *supervised ground truth setting*. In the CB ground truth setting (resp. supervised ground truth setting), the goal of the learner is to optimize the costs drawn from distribution D^b (resp. D^s).

To help make decisions, the learner is given a finite policy class Π that contains policies $\pi : \mathcal{X} \to [K]$[3]. The performance of the algorithm is measured by its *regret* to the retrospective-best policy in Π. We consider two notions of regret over the sequence $\langle x_t \rangle_{t=1}^{n^b}$, based on whether we consider the CB costs (c^b) or the supervised costs (c^s) as the ground truth:

$$\text{CB: } \mathbf{R}^b(\langle x_t, a_t \rangle_{t=1}^{n^b}) = \textstyle\sum_{t=1}^{n^b} \mathbb{E}\big[c^b(a_t) \,\big|\, x_t\big] - \min_{\pi \in \Pi} \textstyle\sum_{t=1}^{n^b} \mathbb{E}\big[c^b(\pi(x_t)) \,\big|\, x_t\big], \qquad (1)$$

$$\text{supervised: } \mathbf{R}^s(\langle x_t, a_t \rangle_{t=1}^{n^b}) = \textstyle\sum_{t=1}^{n^b} \mathbb{E}\big[c^s(a_t) \,\big|\, x_t\big] - \min_{\pi \in \Pi} \textstyle\sum_{t=1}^{n^b} \mathbb{E}\big[c^s(\pi(x_t)) \,\big|\, x_t\big]. \qquad (2)$$

In the content recommendation example (CB ground truth), x_t encodes a user profile and the system predicts which articles (a_t) to display. Here, c^b can be the negative dwell-

[1] SUP-ONLY and MAJORITY do not explore or update on CB examples and we plot the average costs of their policies over all CB examples.

[3] More generally, $\pi : \mathcal{X} \to \Delta^{K-1}$ and $\pi(x)$ is the distribution over actions given x.

time of users and c^s is the annotation of editors, which can have disagreements with c^b. The learner aims to optimize the dwell time over all displayed articles.

In the translation example (supervised ground truth), x_t encodes the text to be translated and a_t encodes its translation. Here, the learner aims to minimize errors against the expert translation (c^s) on x_t's, *despite the fact* that the system never sees these costs in its interaction phase. Note that the learner only observes the user feedback costs (c^b) in this interaction phase, which are imperfect proxies for c^s, and the only direct observations of c^s are on the warm-start examples. Nevertheless, we seek to optimize the accuracy of our translations given to the users, and hence regret is still measured over the interaction phase.

The utility of non ground truth examples are different in the two learning settings. In the CB ground truth setting, relying on the CB examples alone is sufficient to ensure vanishing regret asymptotically. The supervised warm-start primarily helps with a smaller regret in the initial phases of learning. On the other hand, in the supervised ground truth setting, the CB examples can have an asymptotically meaningful effect on the regret: for instance, if $D^s = D^b$, then utilizing CB examples can lead to a vanishing regret, whereas using supervised examples alone cannot.

We can leverage examples from a different source only when the cost structures are at least somewhat related. Therefore, we introduce a measure of similarity of two distributions over cost-sensitive examples.

Definition 1. *D_2 is said to be (α, Δ)-similar to D_1 with respect to Π, if for any policy π, $\mathbb{E}_{D_2} c(\pi(x)) - \mathbb{E}_{D_2} c(\pi^*(x)) \geq \alpha \left(\mathbb{E}_{D_1} c(\pi(x)) - \mathbb{E}_{D_1} c(\pi^*(x))\right) - \Delta$, where $\pi^* = \operatorname{argmin}_{\pi \in \Pi} \mathbb{E}_{D_1} c(\pi(x))$.*

If we have a larger α and smaller Δ, examples from D_2 are more useful for learning under D_1. Prior similarity notions, such as in Ben-David et al. (2010), roughly assume a bound on $\max_{\pi \in \Pi} |\mathbb{E}_{D_1}[c(\pi(x))] - \mathbb{E}_{D_2}[c(\pi(x))]|$. The one-sided bound in our definition (instead of absolute value bound) on regret and an additional scaling factor α yield additional flexibility. Note that Definition 1 is only used in our analysis; our algorithms do not require knowledge of α and Δ. We give a more general condition which implies (α, Δ)-similarity, along with several examples in Appendix B.

Finally, we define some additional notation. In the t-th interaction round, our algorithms compute $\hat{c}_t$, an estimate of the unobserved vector c_t^b. We use $\mathbb{E}_S$ to denote sample averages on S and abbreviate $\mathbb{E}_{S_t}$ by $\mathbb{E}_t$ where $S_t = \{(x_\tau, \hat{c}_\tau)\}_{\tau=1}^t$ is the log of the CB examples up to time t.

2.1. Failure of Simple Strategies

The settings we have described so far might appear deceptively simple. It should be easy to include some additional supervised examples, which contain more feedback, into a CB algorithm. We now illustrate the difficulty of this task when the two distributions D^s and D^b are misaligned.

Consider the special case of 2-armed bandits (CB with a dummy context), where the CB source is the ground truth. D^s and D^b are deterministic with costs $(0.5, 0.5 + \frac{\Delta}{2})$ and $(0.5, 0.5 - \frac{\Delta}{2})$ for the two arms respectively, so that they are $(1, \Delta)$-similar. Suppose we see $n^s = \Omega(1/\Delta^3)$ examples in warmstart, and use them to initialize the means and confidence intervals on each arm to run the UCB algorithm (Auer et al., 2002a). Proposition 1 in Appendix C show that the optimal arm according to D^b, which is arm 2, is never played for the first $O(\exp(1/\Delta))$ rounds, incurring regret $\Omega(\Delta \exp(1/\Delta))$. So for any $\Delta < 0.5$, the regret is strictly larger than that of a UCB algorithm which ignores the warm-start and incurs at most $\tilde{O}(1/\Delta)$ regret. On the other hand, if $D^b = D^s$, then the UCB strategy described above incurs no regret.

What we observe here is a failure in competing simultaneously with two baselines: naively warmstarting by weighting examples from the two sources equally, or just ignoring the supervised source entirely. We will next describe an algorithm to compete not just with these two, but many possible weightings of the two sources. This extreme failure case shows that an arbitrary low-regret CB algorithm cannot handle biased warm-start data without extra care. Using additional randomization can help, but is not adequate by itself as we will see in our theory and experiments.

3. Contextual Bandit Ground Truth Setting

In this section, we study the setting where D^b, the distribution over CB examples, is considered the ground truth, as in the content recommendation example. Recall that in this setting, one could ignore the supervised warm-start examples and still achieve vanishing regret; the main goal here is to show that using the warm start data can help further reduce the regret, especially in early stages of learning.

3.1. Algorithm

Intuition of our approach. The key challenge in designing an algorithm for the CB ground truth setting is understanding how to effectively combine two data sources which might have unknown differences in their distributions. For the simpler supervised learning setting, Proposition 4 in Appendix I shows that it suffices to always use one of the two sources depending on the bias and relative number of examples. This has two caveats though: the bias is not known in practice, and completely ignoring one data source is ob-

Algorithm 1 Adaptive Reweighting for Robustly Warm-starting Contextual Bandits (ARRoW-CB)

```
Require: Supervised dataset $S$ from $D^s$ of size $n^s$, number of interaction rounds $n^b$, exploration probability $\epsilon$, weighted combination parameters $\Lambda$, policy class $\Pi$.
1: for $t = 1, 2, \ldots, n^b$ do
2:     Observe instance $x_t$ from $D^b$.
3:     Define $p_t := \frac{1-\epsilon}{t-1} \sum_{\tau=1}^{t-1} \pi_\tau^{\lambda_t}(x_t) + \frac{\epsilon}{K} \mathbb{1}_K$ for $t \geq 2$ and $p_t := \frac{1}{K} \mathbb{1}_K$ for $t = 1$.
4:     Predict $a_t \sim p_t$, and receive feedback $c_t^b(a_t)$.
5:     Define the inverse propensity score (IPS) cost vector $\hat{c}_t(a) := \frac{c_t^b(a_t)}{p_{t,a_t}} I(a = a_t)$, for $a \in [K]$.
6:     For every $\lambda \in \Lambda$, train $\pi_t^\lambda$ by minimizing over $\pi \in \Pi$:
           $$\lambda \sum_{\tau=1}^{t-1} \hat{c}_\tau(\pi(x_\tau)) + (1-\lambda) \sum_{(x,c^s) \in S} c^s(\pi(x)). \quad (3)$$
7:     Set $\lambda_{t+1} \leftarrow \operatorname{argmin}_{\lambda \in \Lambda} \sum_{\tau=1}^{t} \hat{c}_\tau(\pi_\tau^\lambda(x_\tau))$.
8: end for
```

viously wasteful when the two sources are identical. We choose to instead consider cost minimization on a dataset where the two sources are combined with different weights, and seek to learn these weights adaptively.

With these insights, we return to the actual problem setting of warm-starting a CB learner with supervised examples. Our algorithm for this setting is presented in Algorithm 1. The main idea is to minimize the empirical risk on a weighted dataset containing examples from the two sources. Our algorithm picks the mixture weighting by online model selection over a set of weighting parameters Λ, where we use the ground truth CB data at each time step to evaluate which $\lambda \in \Lambda$ has the best performance so far. For each $\lambda \in \Lambda$, we estimate a $\pi^\lambda \in \Pi$ as the empirical risk minimizer (ERM) for the λ-mixture between CB and supervised examples. We focus on the simplest ϵ-greedy algorithm for CBs, leaving similar modifications in more advanced CB algorithms for future work.

So long as $\{0, 1\} \subseteq \Lambda$, Algorithm 1 allows for relying on one source alone, while using a larger set of Λ significantly improves its empirical performance (see §5).[4]

We need some additional notation to present our regret bound. We define $V_t(\lambda)$ that governs the deviation of λ-weighted empirical costs for all policies in Π as $V_t(\lambda) := 2\sqrt{(\frac{\lambda^2 K t}{\epsilon} + (1-\lambda)^2 n^s) \ln \frac{8n^b|\Pi|}{\delta}} + (\frac{\lambda K}{\epsilon} + (1-\lambda)) \ln \frac{8n^b|\Pi|}{\delta}$, and G_t that bounds the excess cost of the ERM solution using weighted combination parameter λ as

$$G_t(\lambda, \alpha, \Delta) := \frac{(1-\lambda)n^s\Delta + 2V_t(\lambda)}{\lambda t + (1-\lambda)n^s\alpha}.$$

We prove the following theorem in Appendix E.

Theorem 1. *Suppose D^s is (α, Δ)-similar to D^b. Then for any $\delta < 1/e$, with probability $1 - \delta$, the average CB regret of Algorithm 1 can be bounded as:*

$$\frac{1}{n^b} \mathbf{R}^b(\langle x_t, a_t \rangle_{t=1}^{n^b}) \leq \epsilon + 3\sqrt{\frac{\ln \frac{8n^b|\Pi|}{\delta}}{n^b}} + 32\sqrt{\frac{K \ln \frac{8n^b|\Lambda|}{\delta}}{n^b \epsilon}} + \min_{\lambda \in \Lambda} \frac{\ln(en^b)}{n^b} \sum_{t=1}^{n^b} G_t(\lambda, \alpha, \Delta) \quad (4)$$

The bound (4) consists of many intuitive terms. The first ϵ term comes from uniform exploration; the second term is from the deviation of costs under D^b. The next term is the average regret incurred in performing model selection for λ; in our experiments $|\Lambda| = 8$ so that it can be thought of as $\widetilde{O}(\sqrt{K/(n^b\epsilon)})$. The final term involving a minimum over λ's is effectively finding the weighted combination which minimizes a bias-variance tradeoff in combining the two sources. Here the bias is controlled by Δ and in place of variance we use $V_t(\lambda)$ for high-probability results. Contrasting with learning with CB examples alone, we replace a $\sqrt{\frac{K \ln(|\Pi|/\delta)}{n^b\epsilon}}$ term with the middle term independent of $\ln|\Pi|$ and the average of G_t's which can be much smaller in favorable cases as we discuss below.

Identical distributions: A very friendly setting has $D^s = D^b$, corresponding to $(1, 0)$-similarity. Since the theorem holds with a minimum over all λ's in the set Λ, we can pick specific values λ_0 of our choice. One choice of λ_0 motivated from prior work (Ben-David et al., 2010) is to pick it such that $\lambda_0/(1-\lambda_0) = \epsilon/K$ to equalize the variance of the two sources, meaning each supervised example is worth K/ϵ CB examples. This setting of $\lambda_0 = \frac{\epsilon}{K+\epsilon}$ yields $G_t(\lambda_0, 1, 0) = O\left(\sqrt{\frac{K \ln \frac{n^b|\Pi|}{\delta}}{\epsilon t + K n^s}} + \frac{K \ln \frac{n^b|\Pi|}{\delta}}{\epsilon t + K n^s}\right)$. That is, after t CB samples, the effective sample size is $n^s + Kt/\epsilon$.

Comparison with no warmstart: Whenever $1 \in \Lambda$, the minimum over $\lambda \in \Lambda$ in Theorem 1 can be bounded by its value at $\lambda = 1$, which corresponds to ignoring the warmstart examples and using bandit examples alone. For this special case, we have the following corollary.

Corollary 1. *Under conditions of Theorem 1, suppose that $1 \in \Lambda$. Then for any $\delta < 1/e$, with probability $1 - \delta$,*

$$\frac{1}{n^b} \mathbf{R}^b(\langle x_t, a_t \rangle_{t=1}^{n^b}) \leq \epsilon + O\left(\sqrt{\frac{\ln \frac{n^b|\Pi|\,|\Lambda|}{\delta}}{n^b\epsilon}}\right).$$

[4] If we approximate the computation of the best policy in Step 6 using an online oracle as in prior works (Agarwal et al., 2014; Langford & Zhang, 2007), then the entire algorithm can be implemented in a streaming fashion since 7 for selecting the best λ also uses an online estimate a la Blum et al. (1999) for each λ as opposed to a holdout estimate for the current policy π_t^λ.

The corollary follows from using the value of $G_t(1, \alpha, \Delta)$ along with some algebra, and shows that the regret of ARRoW-CB is never worse than using the bandit source alone, up to a term scaling as $\ln |\Lambda|$. In particular, the usual choice of $\epsilon = O((n^b)^{-2/3})$ implies a $O((n^b)^{2/3})$ regret bound. Since a small value of $|\Lambda|$ suffices in our experiments, this is a negligible cost for robustness to arbitrary bias in the warmstart examples. Similarly comparing to $\lambda = 0$ lets us obtain a comparison against using the warmstart alone up to a model selection penalty, when $0 \in \Lambda$. The minimization over a richer set of λ leaves room for further improvements as shown in the case of $D^s = D^b$ above (which used a different setting of λ_0). Further improvements are also possible in the algorithm by using different λ values after reach round, which is not captured in the theory here.

4. Supervised Ground Truth Setting

In §3, we developed an algorithm and proved regret bounds for combining supervised and CB feedback, in the case where the CB cost is considered the ground truth. In this section, we consider the reverse setting where the supervised source constitutes the ground truth, recalling the motivating example in an automated translation setting from the introduction. Here, we wish to leverage the CB examples for learning the best policy relative to the distribution D^s.

Note that this setting is qualitatively different, since we only have a fixed number n^s of ground-truth examples while the number of CB examples grows over time. If we assign relative weights to individual supervised and CB examples as in Algorithm 1, the CB examples will eventually outweigh the supervised ones for any $\lambda > 0$, which is not desirable when the supervised source is the ground truth. In Algorithm 2, we address this problem by first computing the average costs of every policy on the supervised and CB examples separately, and then choosing a policy that minimizes a weighted combination of these averages. As a consequence, the relative weight of each CB example diminishes as their number grows, with the overall bias incurred from the CB source staying bounded.

Another difference between Algorithm 2 and Algorithm 1 is that, as opposed to using the CB examples collected online, we use subsets of warm start examples to guide the selection of weighted combination parameter λ. To this end, we introduce an epoch structure in the algorithm. In particular, at each epoch e, λ_e and π_e^λ's are updated exactly once, where a separate validation set is used to pick λ. In addition, we play with uniform randomization around the most recent policy as opposed to a running average of all policies trained so far, an outcome of using a separate validation set (line 12 of Algorithm 2) instead of progressive validation (7 of Algorithm 1). Since the exploration policy at the next epoch depends on the previous validation set, we must use a "fresh" validation set at each epoch. Avoiding this splitting is an interesting question for future work.

Algorithm 2 Combining contextual bandit and supervised data when supervised source is the ground truth

```
Require: Supervised dataset $S$ from $D^s$ of size $n^s$, number of interaction rounds $n^b$, exploration probability $\epsilon$, weighted combination parameters $\Lambda$, policy set $\Pi$.
 1: Let $E = \lceil \log n^b \rceil$ be the number of epochs.
 2: Define $t_e = \min(2^e, n^b)$ for $e \geq 1$, and $t_0 = 0$.
 3: Partition $S$ to $E+1$ equally sized sets $S^{\text{tr}}, S_1^{\text{val}}, \ldots, S_E^{\text{val}}$.
 4: for $e = 1, 2, \ldots, E$ do
 5:   for $t = t_{e-1} + 1, t_{e-1} + 2, \ldots, t_e$ do
 6:     Observe instance $x_t$ from $D^b$.
 7:     Define $p_t := (1-\epsilon)\pi_{e-1}^{\lambda_{e-1}}(x_t) + \frac{\epsilon}{K}\mathbb{1}_K$ for $e \geq 2$, and $p_t := \frac{1}{K}\mathbb{1}_K$ for $e = 1$.
 8:     Predict $a_t \sim p_t$ and receive feedback $c_t^b(a_t)$.
 9:     Define the IPS cost vector $\hat{c}_t(a) := \frac{c_t^b(a_t)}{p_{t,a_t}} I(a = a_t)$, for $a \in [K]$.
10:   end for
11:   For each $\lambda \in \Lambda$, train $\pi_e^\lambda$ as:
        $\arg\min_{\pi \in \Pi} \lambda \mathbb{E}_{t_e} \hat{c}(\pi(x)) + (1-\lambda)\mathbb{E}_{S^{\text{tr}}} c^s(\pi(x))$.
12:   Set $\lambda_e \leftarrow \arg\min_{\lambda \in \Lambda} \mathbb{E}_{S_e^{\text{val}}} c^s(\pi_e^\lambda(x))$.
13: end for
```

For the main result, we need the following notation for the deviation of λ-weighted empirical costs, where $E = \lceil \log n^b \rceil$ is the total number of epochs:

$$W_t(\lambda) := 2\sqrt{\left(\frac{\lambda^2 K}{t\epsilon} + \frac{(1-\lambda)^2(E+1)}{n^s}\right)\ln\frac{8E|\Pi|}{\delta}} + \left(\frac{\lambda K}{t\epsilon} + \frac{(1-\lambda)(E+1)}{n^s}\right)\ln\frac{8E|\Pi|}{\delta}.$$

Theorem 2. *Suppose that D^b is (α, Δ)-similar to D^s. Then for any $\delta < 1/e$, with probability $1-\delta$, the average supervised regret of Algorithm 2 can be bounded as:*

$$\frac{1}{n^b}\mathbf{R}^s(\langle x_t, a_t\rangle_{t=1}^{n^b}) \leq \epsilon + 3\sqrt{\frac{\ln\frac{8|\Pi|}{\delta}}{n^b}} + \sqrt{\frac{2(E+1)\ln\frac{8E|\Lambda|}{\delta}}{n^s}} + \min_{\lambda \in \Lambda}\frac{2}{n^b}\sum_{t=1}^{n^b}\frac{\lambda\Delta + 2W_t(\lambda)}{(1-\lambda)+\lambda\alpha}. \qquad (5)$$

The first term is the cost of exploration, while the second is the gap between the conditional and unconditional expectations over costs in defining the regret. The third term captures the complexity of model selection while the final is the performance upper bound for the best λ in our weighted combination set Λ. As before, this significantly improves upon the $O(\sqrt{\frac{\ln|\Pi|/\delta}{n^s}})$ bound from using supervised examples alone whenever the two sources have sufficient similarity. The proof can be found in Appendix F.

Identical distributions: When $D^s = D^b$, which implies that D^s is $(1, 0)$-similar to D^b, a single choice of $\lambda =$

$\frac{n^b\epsilon}{n^sK+n^b\epsilon}$ will ensure that the last term in Equation (5) is at most $\tilde{O}\left(\sqrt{\frac{K\ln\frac{8E|\Pi|}{\delta}}{Kn^s+n^b\epsilon}}+\frac{K\ln\frac{8E|\Pi|}{\delta}}{Kn^s+n^b\epsilon}\right)$ (See Proposition 2 in Appendix G). That is, after n^b CB samples, the effective sample size is at most $n^s + n^b\epsilon/K$.

5. Experiments

Experimentally, we focus on the question of learning with the CB costs as the ground truth (§3). Our experiments seek to address the following questions: **a)** How much benefit does a small amount of supervised warm-start provide? **b)** How much benefit does the bandit feedback provide? **c)** How robust is our algorithm under a realistic mismatch in cost structures? **d)** How robust is our algorithm under adversarial cost structures (the "safety" question)?

We consider the following set of approaches:
Bandit-Only: a baseline that only uses CB examples.
Majority: always predicts $a \in \text{argmin}_{a\in[K]}\mathbb{E}_{(x,c)\sim D^b}[c(a)]$ independent of the context, without exploration.
Sup-Only: a baseline that uses the best policy on supervised examples, without exploration.
Sim-Bandit: a baseline that runs the CB algorithm on warm-start examples as well, providing cost for the chosen action only (from the supervised set) and then continues on the remaining CB examples.
ARRoW-CB with $\Lambda = \{0, \frac{1}{8}\zeta, \frac{1}{4}\zeta, \frac{1}{2}\zeta, \zeta, \frac{1}{2}+\frac{1}{2}\zeta, \frac{3}{4}+\frac{1}{4}\zeta, 1\}$ (abbrev. **ARRoW-CB with** $|\Lambda| = 8$), where $\zeta = \epsilon/(K+\epsilon)$; this is chosen because ζ is an approximate minimizer of $G_t(\lambda, 1, 0)$, and the $|\Lambda|$ used aims to ensure that $\min_{\lambda\in\Lambda} G_t(\lambda,\alpha,\Delta)$ is close to $\min_{\lambda\in[0,1]} G_t(\lambda,\alpha,\Delta)$ (see Prop. 3). For computational considerations, we use the last policy $\pi_t^{\lambda_t}$ rather than the averaged policy $\frac{1}{t-1}\sum_{\tau=1}^{t-1}\pi_\tau^{\lambda_t}$ in line 3 of Algorithm 1.
ARRoW-CB with $\Lambda = \{0,1\}$(abbrev. **ARRoW-CB with** $|\Lambda| = 2$): as argued in Proposition 3, choosing λ in $\{0,1\}$ also approximately minimizes $G_t(\lambda,\alpha,\Delta)$.

In subsequent discussions, if not explicitly mentioned, ARRoW-CB refers to ARRoW-CB with $|\Lambda| = 8$.

All the algorithms (other than Sup-Only and Majority, which do not explore) use ϵ-greedy exploration, with most of the results presented using $\epsilon = 0.0125$. We additionally present the results for $\epsilon = 0.1$ and $\epsilon = 0.0625$ in Appendix J. In general, the increased uniform exploration for larger ϵ leads to some performance penalty in the CB algorithms relative to Sup-Only, when the bias is small. However, the added exploration gives robustness to large bias as it is readily detected in more adversarial noise settings.

Datasets. We compare these approaches on 524 binary and multiclass classification datasets from Bietti et al. (2018), which in turn are from `openml.org`. For each dataset, we use the multiclass label in the dataset to generate cost vectors c^b and c^s respectively. That is, given an example $(x,y) \in \mathcal{X}\times[K]$, $c^b(a) = I(a \neq y)$. We vary the number of warm-start examples and CB examples as follows: for a dataset of size n, we vary the number of warm-start examples n^s in $\{0.005n, 0.01n, 0.02n, 0.04n\}$, and the number of CB examples n^b in $\{0.92n, 0.46n, 0.23n, 0.115n\}$. Define the *warm-start ratio* as the ratio n^b/n^s. We group different settings of (dataset, n^s, n^b) by n^b/n^s, so that a separate plot is generated for each ratio in $R = \{2.875, 5.75, 11.5, 23, 46, 92, 184\}$. We filter out the settings where n^s is below 100.

Evaluation Criteria. For each (dataset, n^s, n^b) combination c, we can compute $e_{c,a}$ to be the average cost of algorithm a on the CB examples. Because the range of $e_{c,a}$ can vary significantly over different settings c, we normalize these to yield the *normalized error* of an algorithm on a dataset: $\texttt{err}_{c,a} := \frac{e_{c,a}-e_c^*}{\max_b e_{c,b}-e_c^*}$, where e_c^* is the error achieved by a fully supervised one-versus-all learning algorithm trained on all the examples with original labels in this dataset. Lower normalized error indicates better performance. We plot the cumulative distribution function (CDF) of the normalized errors for each algorithm. That is, for an algorithm a, at each point x, the y value is the fraction of c's such that $\texttt{err}_{c,a} \le x$. In general, a high CDF value at a small x indicates that the algorithm is performing well over a large number of (dataset, n^s, n^b) combinations.

In some of the plots when investigating the effect of a particular type or level of noise, we find it useful to aggregate the plots further over all warm-start ratios in creating the CDF and this aggregation is done by a pointwise averaging of the individual CDFs.

Comparison with baselines using both sources. We present the CDFs of all algorithms under various noise models in Figure 2, with detailed results for individual noise levels, warm-start ratios and different ϵ values in Appendix J. In Figure 2, we aggregate over warm-start ratios as described earlier. We can see from the figures that ARRoW-CB's CDFs (approximately) dominate those of Sim-Bandit and ARRoW-CB with $|\Lambda| = 2$, which use weightings of 0.5, and the best of $\{0,1\}$ respectively. These gains highlight the importance of being more careful about selecting a good weighting, despite the earlier intuition from Proposition 4. We see that there is a potentially added benefit of using different λ's in different phases of learning which might even outperform the best setting in hindsight.

Results for aligned cost structures. In Fig. 2a, we consider the setting $c^s = c^b$. Here, ARRoW-CB's CDF dominates all other algorithms other than Sup-Only. For Sup-Only, the warm-start policy is used greedily with no explo-

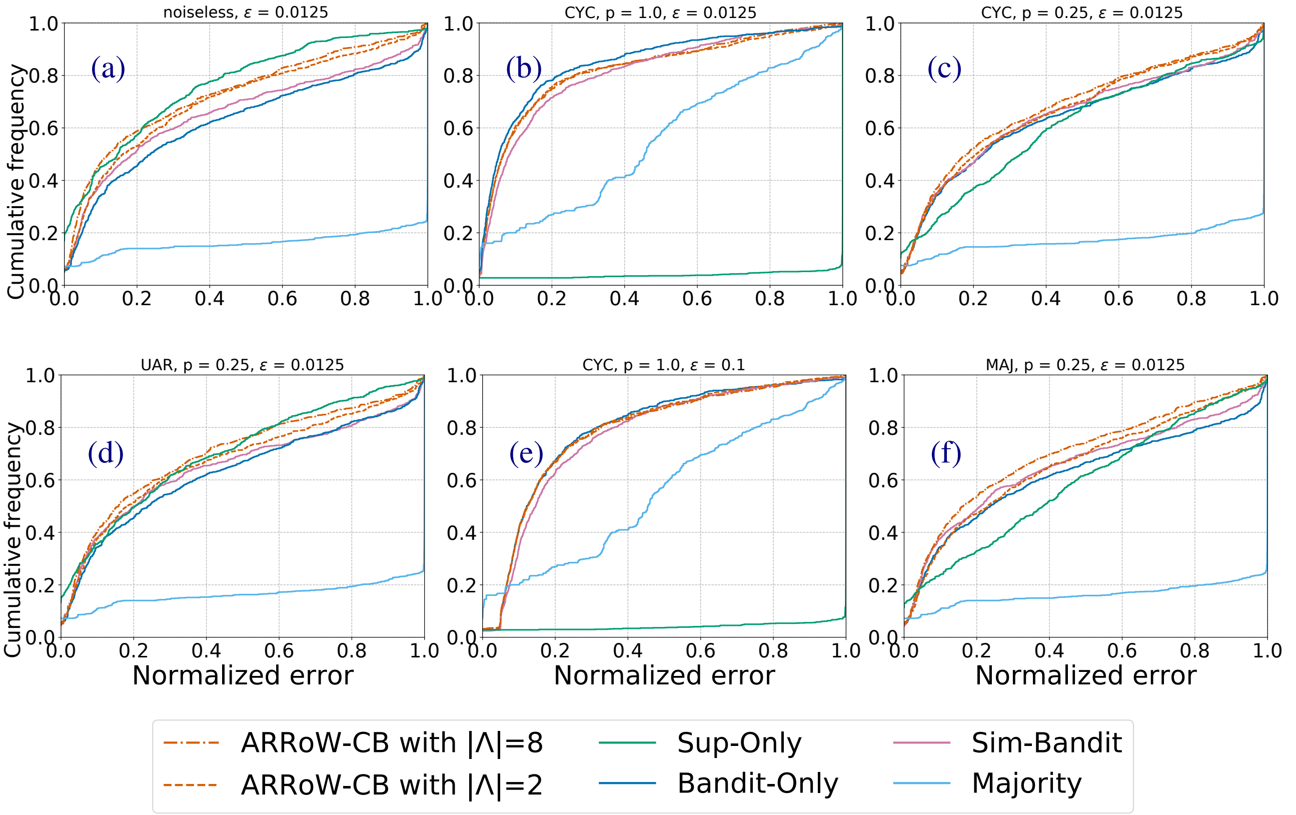

Figure 2: Comparison of all algorithms in the CB ground truth setting using the empirical CDF of the normalized performance scores. Left: unbiased warm-start examples with noiseless (top) and UAR with probability 0.5 (down) costs on warm-start examples. Middle: extreme noise rate using CYC noise type with probability 1.0. All CB algorithms use $\epsilon = 0.0125$ for exploration (top) and $\epsilon = 0.1$ (bottom). Right: moderate and potentially helpful noise rates. The corruption added to the warm-start examples are of types CYC (top) and MAJ (down) respectively, with probability 0.25.

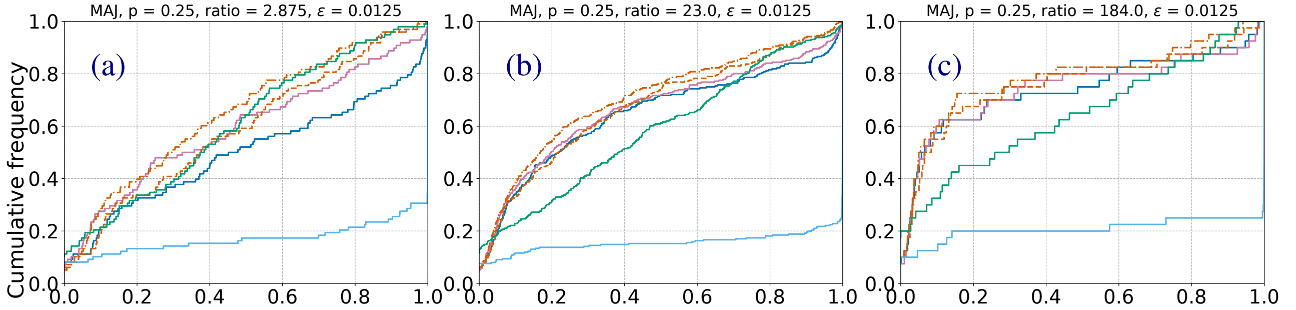

Figure 3: Effect of varying warm-start ratios for MAJ noise with $p = 0.25$. The warm-start ratios vary from 2.875 (left), 23 (middle) to 184 (right). Each CDF aggregates over all conditions of this noise type with the same warm-start ratio.

ration, making it a very strong baseline when there is no bias. We observe that ARRoW-CB uses the warm-start much more effectively than both the Sim-Bandit and ARRoW-CB with $|\Lambda| = 2$ baselines. Our next experiments consider a uniform at random (UAR) noise setting, where the supervised data is unbiased (with respect to c^b) but has higher variance. In particular, for every example (x, c^b), with probability $1 - p$ we set $c^s = c^b$ and with probability p we set c^s as the classification error against a uniform random label. From Claim 2 in the appendix, D^s is $(1 - p, 0)$-similar to D^b. We plot the CDFs of the algorithms in the case where $p = 0.25$ in Fig. 2d. The ordering of the CDFs stays the same, with SUP-ONLY less dominant (unsurprisingly), and with the gaps between methods reduced with the reduced utility of the warm-start data.

Results with adversarial noise. We next conduct an experiment where c^s and c^b are highly misaligned in order to understand how robust ARRoW-CB is to adversarial conditions. We consider the cycling noise model (CYC), where we set the supervised costs to be "off-by-one" from the CB costs. Specifically, if c^b declares action a to be the zero-cost action, then, with probability p, c^s corrupts the costs so that action $(a+1) \bmod K$ becomes the zero-cost action. From Claim 3 in the appendix, D^s is $(1, 2p)$-similar to D^b. The CDF results for this experiment are in Figs. 2b, 2e ($p = 1$) for $\epsilon = 0.0125$ and $\epsilon = 0.1$, and Fig. 2c ($p = 0.25$) for $\epsilon = 0.0125$ respectively. Again, ARRoW-CB is dominant amongst methods which use both the sources. In the case of $p = 1.0$, BANDIT-ONLY performs the best as the warm start examples are misleading. ARRoW-CB performs slightly worse (Fig. 2b) due to the model selection overhead, as discussed following Theorem 1. This gap is reduced when we increase the ϵ value in ϵ-greedy to 0.1 (Fig. 2e). In the case of $p = 0.25$ (Fig. 2c), ARRoW-CB outperforms all the methods, showing that it can utilize warm start examples even if they are moderately biased.

Results with majority noise. Finally, we consider the case of a noise model that replaces the ground truth label with the majority label, roughly modeling a "lazy annotator" who occasionally defaults to the most frequent class. For the majority noise model (MAJ), with probability $1-p$, we set $c^s = c^b$ and with probability p we set c^s to a cost vector that has a zero for the most frequent label in this dataset and one elsewhere. From Claim 3 in the appendix, D^s is $(1, 2p)$-similar to D^b. The CDFs for this setting are shown in Figure 2f, where we again see ARRoW-CB dominating all the baselines (similar to Figure 2c).

In sum, we observe that ARRoW-CB is the *only* method which is the best or close across all the noise regimes; no other approach is consistently strong. In practical scenarios, where the extent of bias in the warm-start is difficult or costly to ascertain, this robust performance of ARRoW-CB is extremely desirable. If we have some prior information about the noise level, it is prudent to prefer smaller ϵ when we expect a low noise (to compete well with SUP-ONLY), while a larger ϵ is preferred in high noise situations (to quickly detect the extent of bias).

While we present aggregates over warm-start ratios here, plots for each combination of noise type, level and warm-start ratio for three values of ϵ are shown in Appendix J.

Effect of warm-start ratio. In Fig. 3, we pick a moderate noise setting and study the ordering of the different methods as the number of warm-start examples n^s increases relative to n^b. We see ARRoW-CB outperforming all methods. SUP-ONLY is strong on the left for a small ratio (2.875), while BANDIT-ONLY does well on the other extreme (184), and ARRoW-CB consistently outperform both the baselines combining the two sources.

Overall. Overall, we see that effectively using warm-start examples can certainly improve the performance of CB approaches. ARRoW-CB provides a way to do this in a robust manner, consistently outperforming most baselines. This is best evidenced in Figure 1, which further aggregates performance across the following 10 noise conditions on the warm start examples: noiseless and $\{$UAR, CYC, MAJ$\}$ corruptions with probability p in $\{0.25, 0.5, 1.0\}$.

6. Discussion and Future Work

In this paper, we study the question of incorporating multiple data sources in CB settings. We see that even in simple cases, obvious techniques do not work robustly, and some care is required to handle biases from the non-ground-truth source.

Building on our results, there are several natural avenues for future work. Doing a similar modification to more advanced exploration algorithms (e.g. (Agrawal & Goyal, 2013; Agarwal et al., 2014)) is significantly more challenging. This falls into the general category of selecting the best from an ensemble of CB algorithms (where the ensemble corresponds to different weightings of the supervised and the CB examples). In ϵ-greedy, the policy training corresponds to training the CB algorithm on reweighted data, while the model selection over λ induces the action distribution. While the first step is typically straightforward even for other CB algorithms, finding an action distribution which looks good at this round, while allows the CB algorithms for different λ values to make subsequent updates is significantly harder (for instance, when using a UCB style strategy, each λ value might suggest a completely different action and expect reward feedback about it). A possible approach is to employ ideas from the CORRAL algorithm (Agarwal et al., 2017), but the cost of model selection is linear instead of logarithmic in $|\Lambda|$, and the approach is somewhat data inefficient due to restarts. More ambitiously, it is desirable for the schedule of supervised and CB examples to not be fixed in a warm-start fashion but based on active querying, such as by sending uncertain examples to a labeler for full supervision. Studying this and considering broader sources of feedback are both interesting future research.

Acknowledgments

We thank Alberto Bietti for kindly sharing the scripts for experiments performed by Bietti et al. (2018), and helping getting the experiments running. We also thank the anonymous reviewers for their helpful feedback.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646, 2014.

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. *COLT*, 2017.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 127–135, 2013. URL http://jmlr.org/proceedings/papers/v28/agrawal13.html.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Beygelzimer, A., Langford, J., and Zadrozny, B. Weighted one-against-all. In *AAAI*, 2005.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26, 2011.

Bietti, A., Agarwal, A., and Langford, J. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

Blum, A., Kalai, A., and Langford, J. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *COLT*, 1999.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pp. 208–214, 2011. URL http://www.jmlr.org/proceedings/papers/v15/chu11a/chu11a.pdf.

Crammer, K., Kearns, M., and Wortman, J. Learning from multiple sources. *Journal of Machine Learning Research*, 9:1757–1774, 2008.

Donmez, P. and Carbonell, J. G. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 619–628. ACM, 2008.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*. Citeseer, 2011.

Karampatziakis, N. and Langford, J. Online importance weight aware updates. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 392–399, Arlington, Virginia, United States, 2011. AUAI Press. ISBN 978-0-9749039-7-2. URL http://dl.acm.org/citation.cfm?id=3020548.3020594.

Kazerouni, A., Ghavamzadeh, M., Abbasi, Y., and Roy, B. V. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 3913–3922, 2017.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 817–824. Curran Associates Inc., 2007.

Malago, L., Cesa-Bianchi, N., and Renders, J. Online active learning with strong and weak annotators. In *NIPS Workshop on Learning from the Wisdom of Crowds*, 2014.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *COLT*, 2009.

Nguyen, K., Daumé III, H., and Boyd-Graber, J. Reinforcement learning for bandit neural machine translation with simulated human feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. URL http://hal3.name/docs/#daume17simhuman.

Ross, S., Mineiro, P., and Langford, J. Normalized online learning. *UAI*, 2013.

Sokolov, A., Riezler, S., and Urvoy, T. Bandit structured prediction for learning from partial feedback in statistical machine translation. In *MT Summit*, 2015.

Sun, W., Dey, D., and Kapoor, A. Safety-aware algorithms for adversarial contextual bandit. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 3280–3288, 2017. URL http://proceedings.mlr.press/v70/sun17a.html.

Tewari, A. and Murphy, S. A. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pp. 495–517. Springer, 2017.

Urner, R., David, S. B., and Shamir, O. Learning from weak teachers. In *Artificial Intelligence and Statistics*, pp. 1252–1260, 2012.

Yan, S., Chaudhuri, K., and Javidi, T. Active learning with logged data. *ICML*, 2018.

Yan, Y., Rosales, R., Fung, G., and Dy, J. G. Active learning from crowds. In *ICML*, 2011.

Yu, B. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.

Zhang, C. and Chaudhuri, K. Active learning from weak and strong labelers. In *Advances in Neural Information Processing Systems*, pp. 703–711, 2015.