UNDERSTANDING NEW TASKS THROUGH THE LENS OF TRAINING DATA VIA EXPONENTIAL TILTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deploying machine learning models on new tasks is a major challenge due to differences in distributions of the train (source) data and the new (target) data. However, the training data likely captures some of the properties of the new task. We consider the problem of reweighing the training samples to gain insights into the distribution of the target task. Specifically, we formulate a distribution shift model based on the exponential tilt assumption and learn train data importance weights minimizing the KL divergence between labeled train and unlabeled target datasets. The learned train data weights can then be used for downstream tasks such as target performance evaluation, fine-tuning, and model selection. We demonstrate the efficacy of our method on WATERBIRDS and BREEDS benchmarks.

1 INTRODUCTION

Machine learning models are often deployed in a target domain that differs from the domain in which they were trained and validated in. This leads to the practical challenges of adapting and evaluating the performance of models on a new domain without costly labeling of the dataset of interest. For example, in the Inclusive Images challenge (Shankar et al., 2017), the training data largely consists of images from countries in North America and Western Europe. If a model trained on this data is presented with images from countries in Africa and Asia, then (i) it is likely to perform poorly, and (ii) its performance in the training (source) domain may not mirror its performance in the target domain. However, due to the presence of a small fraction of images from Africa and Asia in the source data, it may be possible to reweigh the source samples to mimic the target domain.

In this paper, we consider the problem of learning a set of importance weights so that the reweighted source samples closely mimic the distribution of the target domain. We pose an exponential tilt model of the distribution shift between the train and the target data and an accompanying method that leverages unlabeled target data to fit the model. Although similar methods are widely used in statistics Rosenbaum & Rubin (1983) and machine learning Sugiyama et al. (2012) to train and evaluate models *under covariate shift* (where the decision function/boundary does not change), one of the main benefits of our approach is it allows *concept drift* (where the decision boundary/function are expected to differ) (Cai & Wei, 2019; Gama et al., 2014) between the source and the target domains. We summarize our contributions below:

- In Section 3 we develop a model and an accompanying method for learning source importance weights to mimic the distribution of the target domain *without* labeled target samples.
- In Section 4 we establish theoretical guarantees on the quality of the weight estimates and their utility in the downstream tasks of fine-tuning and model selection.
- We demonstrate applications of our method on WATERBIRDS (Sagawa et al., 2019) (Section 5), BREEDS (Santurkar et al., 2020) (Section 6) and synthetic (Appendix C) datasets.

2 RELATED WORK

Out-of-distribution generalization is essential for safe deployment of ML models. There are two prevalent problem settings: domain generalization and subpopulation shift (Koh et al., 2020). Domain generalization typically assumes access to several datasets during training that are related to the same task, but differ in their domain or environment (Blanchard et al., 2011; Muandet et al.,

2013). The goal is to learn a predictor that can generalize to unseen related datasets via learning invariant representations (Ganin et al., 2016; Sun & Saenko, 2016), invariant risk minimization (Arjovsky et al., 2019; Krueger et al., 2021), or meta-learning (Dou et al., 2019). Domain generalization is a very challenging problem and recent benchmark studies demonstrate that corresponding methods rarely improve over vanilla empirical risk minimization (ERM) on the source data unless given access to labeled target data for model selection (Gulrajani & Lopez-Paz, 2020; Koh et al., 2020).

Subpopulation shift setting assumes that both train and test data consist of the same groups with different group fractions. This setting is typically approached via distributionally robust optimization (DRO) to maximize worst group performance (Duchi et al., 2016; Sagawa et al., 2019), various reweighing strategies (Shimodaira, 2000; Byrd & Lipton, 2019; Sagawa et al., 2020; Idrissi et al., 2021). These methods require group annotations which could be expensive to obtain in practice. Several methods were proposed to sidestep this limitation, however they still rely on a validation set with group annotations for model selection to obtain good performance (Hashimoto et al., 2018; Liu et al., 2021; Zhai et al., 2021; Creager et al., 2021). Our method is most appropriate for the subpopulation shift setting (see Section 3), however it differs in that it does not require group annotations, but requires unlabeled target data.

Model selection on out-of-distribution (OOD) data is an important and challenging problem as noted by several authors (Gulrajani & Lopez-Paz, 2020; Koh et al., 2020; Zhai et al., 2021; Creager et al., 2021). Xu & Tibshirani (2022); Chen et al. (2021b) propose solutions specific to covariate shift based on parametric bootstrap and reweighing; Garg et al. (2022); Guillory et al. (2021); Yu et al. (2022) align model confidence and accuracy with a threshold; Jiang et al. (2021); Chen et al. (2021a) train several models and use their ensembles or disagreement. Our importance weighting approach is computationally simpler than the latter and is more flexible in comparison to the former, as it allows for concept drift and can be used in downstream tasks beyond model selection as we demonstrate both theoretically and empirically.

Domain adaptation is another closely related problem setting. Domain adaptation (DA) methods require access to labeled source and unlabeled target domains during training and aim to improve target performance via a combination of distribution matching (Ganin et al., 2016; Sun & Saenko, 2016; Shen et al., 2018), self-training (Shu et al., 2018; Kumar et al., 2020), data augmentation (Cai et al., 2021; Ruan et al., 2021), and other regularizers. DA methods are typically challenging to train and require retraining for every new target domain. On the other hand, our importance weights are easy to learn for a new domain allowing for efficient fine-tuning, similar to test-time adaptation methods (Sun et al., 2020; Wang et al., 2020; Zhang et al., 2020), which adjust the model based on the target unlabeled samples. Our importance weights can also be used to define additional regularizers to enhance existing DA methods.

Importance weighting has often been used in the domain adaptation literature on label shift (Lipton et al., 2018; Azizzadenesheli et al., 2019; Maity et al., 2020) and covariate shift (Sugiyama et al., 2007; Hashemi & Karimi, 2018) but the application has been lacking in the area of concept drift models (Cai & Wei, 2019; Maity et al., 2021), due to the reason that it is generally impossible to estimate the weights without seeing labeled data from the target. In this paper, we introduce an exponential tilt model which accommodates concept drift while allowing us to estimate the importance weights for the distribution shift.

3 THE EXPONENTIAL TILT MODEL

Notation We consider a K-class classification problem. Let $\mathcal{X} \in \mathbf{R}^d$ and $\mathcal{Y} \triangleq [K]$ be the space of inputs and set of possible labels, and P and Q be probability distributions on $\mathcal{X} \times \mathcal{Y}$ for the source and target domains correspondingly. A (probabilistic) classifier is a map $f : \mathcal{X} \to \Delta^{K-1}$. We define $p\{x, Y = k\}$ as the weighted source class conditional density, *i.e.* $p\{x, Y = k\} = p\{x \mid Y = k\} \times P\{Y = k\}$, where $p\{x \mid Y = k\}$ is the density of the source feature distribution in class k and $P\{Y = k\}$ is the class probability in source. We similarly define $q\{x, Y = k\}$ for target.

We consider the problem of learning importance weights on samples from a source domain so that the weighted source samples mimic the target distribution. We assume that the learner has access to labeled samples $\{(X_{P,i}, Y_{P,i})\}_{i=1}^{n_P}$ from the source domain and and unlabeled samples $\{X_{Q,i}\}_{i=1}^{n_Q}$ from the target domain. The learner's goal is to estimate a weight function $\omega : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$ such

that

$$\mathbf{E}\left[\omega(X_P, Y_P) g(X_P, Y_P)\right] \approx \mathbf{E}\big[g(X_Q, Y_Q)\big] \text{ for all (reasonable) } g : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}. \tag{3.1}$$

Ideally, $\omega = \frac{dQ}{dP}$ is the likelihood ratio between the source and target domains (this leads to equality in (3.1)), but learning this weight function is generally impossible without labeled samples from the target domain (David et al., 2010). Thus we must impose additional restrictions on the source and target domains.

The exponential tilt model We assume that there is a vector of sufficient statistics $T : \mathcal{X} \to \mathbf{R}^p$ and the parameters $\{\theta_k \in \mathbf{R}^p, \alpha_k \in \mathbf{R}\}_{k=1}^K$ such that

$$\log \frac{q\{x, Y = k\}}{p\{x, Y = k\}} = \theta_k^\top T(x) + \alpha_k \text{ for all } k \in [K]; \tag{3.2}$$

i.e. $q\{x, Y = k\}$ is a member of the exponential family with base measure $p\{x, Y = k\}$ and sufficient statistics T. We call (3.2) the **exponential tilt** model. It implies the importance weights between the source and target samples are

$$\omega(x, y) = \exp(\theta_y^\top T(x) + \alpha_y).$$

The exponential tilt model is motivated by the rich theory of exponential families in statistics. In machine learning, it was used for learning with noisy labels and for improving worst-group performance when group annotations are available (Li et al., 2020; 2021). It is also closely related to several common models in transfer learning and domain adaptation. In particular, it implies there is a linear concept drift between the source and target domains. It also extends the widely used **covariate shift** (Sugiyama & Kawanabe, 2012) and **label shift** models (Alexandari et al., 2020; Lipton et al., 2018; Azizzadenesheli et al., 2019; Maity et al., 2020; Garg et al., 2020) of distribution shifts. It extends the covariate shift model because the exponential tilt model permits (linear in $T(X)$) **concept drifts** between the source and target domains; it extends the label shift model because it allows the class conditionals to differ between the source and target domains. It does, however, come with a limitation: implicit in the model is the assumption that there is some amount of overlap between the source and target domains. In the subpopulation shift setting, this assumption is always satisfied, while in domain generalization it may be violated if the new domain drastically differs from the source data (see Appendix C for a synthetic data example).

Choosing T We comment on the choice of the sufficient statistics T under the subpopulation shift setting. The goal of T is to identify the common subpopulations in the source and target domains, *i.e.*

$$(X_P, Y_P) \mid \{T(X_P) = t, Y_P = k\} \overset{d}{\approx} (X_Q, Y_Q) \mid \{T(X_Q) = t, Y_P = k\}.$$

If T segments the source domain into its subpopulations (*i.e.* the subpopulations are $\{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid T(x) = t, y = k\}$ for different values of t's and k's), then it is possible to achieve perfect reweighing of the source domain with the exponential tilt model: the weight of the $\{T(X) = t, Y = k\}$ subpopulation is $\exp(\theta_k^\top t + \alpha_k)$. However, in practice, such a T that perfectly segments the subpopulations may not exist (*e.g.* the subpopulations may overlap) or is very hard to learn (*e.g.* we don't have prior knowledge of the subpopulations to guide T).

If no prior knowledge of the domains is available, we can use a neural network to parameterize T and learn its weights along with the tilt parameters, or simply use a pre-trained feature extractor as T, which we demonstrate to be sufficiently effective in our empirical studies. We also study the effects of misspecification of T using a synthetic dataset example in Appendix C.

Fitting the exponential tilt model We fit the exponential tilt model via distribution matching. This step is based on the observation that under the exponential tilt model (3.2)

$$q_X\{x\} = \textstyle\sum_{k=1}^K p\{x, Y = k\} \exp(\theta_k^\top T(x) + \alpha_k), \tag{3.3}$$

where q_X is the (marginal) density of the inputs in the target domain. It is possible to obtain an estimate $\widehat{q}_X$ of q_X from the unlabeled samples $\{X_{i,Q}\}_{i=1}^n$ and estimates $\widehat{p}\{x, Y = k\}$ of the $p\{x, Y = k\}$'s from the labeled samples $\{(X_{i,P}, Y_{i,P})\}_{i=1}^m$. This suggests we find θ_k's and α_k's such that

$$\textstyle\sum_{k=1}^K \widehat{p}\{x, Y = k\} \exp(\theta_k^\top T(x) + \alpha_k) \approx \widehat{q}_X\{x\}.$$

Note that the θ_k's and α_k's are dependent because $\widehat{q}_X$ must integrate to one. We enforce this restriction as a constraint in the distribution matching problem:

$$\{(\hat{\theta}_k, \hat{\alpha}_k)\}_{k=1}^K \in \begin{cases} \arg\min_{\{(\theta_k,\alpha_k)\}_{k=1}^K} D\left(\widehat{q}_X\{x\} \| \sum_{k=1}^K \widehat{p}\{x, Y=k\} \exp(\theta_k^\top T(x) + \alpha_k)\right) \\ \text{subject to } \int_{\mathcal{X}} \sum_{k=1}^K \widehat{p}\{x, Y=k\} \exp(\theta_k^\top T(x) + \alpha_k) dx = 1\,, \end{cases} \tag{3.4}$$

where D is a discrepancy between probability distributions on $\mathcal{X}$. Although there are many possible choices of D, we pick the Kullback-Leibler (KL) divergence in the rest of this paper because it leads to some computational benefits. We reformulate the above optimization for KL-divergence to relax the constraint which we state in the following lemma.

Lemma 3.1. *If D is the Kullback-Leibler (KL) divergence then optima in* (3.3) *is achieved at $\{(\hat{\theta}_k, \hat{\alpha}_k)\}_{k=1}^K$ where*

$$\{(\hat{\theta}_k, \hat{\alpha}'_k)\}_{k=1}^K \in \arg\max_{\{(\theta_k,\alpha'_k)\}_{k=1}^K} \mathbf{E}_{\widehat{Q}_X}\left[\log\left\{\sum_{k=1}^K \widehat{\eta}_{P,k}(X) exp(\theta_k^\top T(X) + \alpha'_k\right\}\right] - \log\left\{\mathbf{E}_{\widehat{P}}\left[exp(\theta_Y^\top T(X) + \alpha'_Y)\right]\right\}$$

$\widehat{\eta}_P = \{\widehat{\eta}_{P,k}\}_{k=1}^K$ is a probabilistic classifier for P and $\hat{\alpha}_k = \hat{\alpha}'_k - \log\left\{\mathbf{E}_{\widehat{P}}\left[exp(\hat{\theta}_Y^\top T(X) + \hat{\alpha}'_Y)\right]\right\}$.

One benefit of minimizing the KL divergence is that the learner does not need to estimate the $p\{x, Y=k\}$'s; they merely need to train a discriminative model to estimate $\widehat{\eta}_P$ from the (labeled) samples from the source domain. In other words, the learner does not need to train a generative model.

We plug the fitted $\hat{\theta}_k$'s and $\hat{\alpha}_k$'s into (3.5) to obtain Exponential Tilt Reweighting Alignment (ExTRA) importance weights:

$$\widehat{\omega}(x, y) = \exp(\hat{\theta}_y^\top T(x) + \hat{\alpha}_y). \tag{3.5}$$

We summarize the ExTRA procedure in Algorithm 1 in Appendix B.2.

Next we describe two downstream tasks where ExTRA weights can be used:

1. **ExTRA model evaluation in the target domain.** Practitioners may estimate the performance of a model in the target domain by reweighing the empirical risk in the source domain:

$$\mathbf{E}\left[\ell(f(X_Q), Y_Q)\right] \approx \frac{1}{n_P} \sum_{i=1}^{n_P} \ell(f(X_{P,i}), Y_{P,i}) \widehat{\omega}(X_{P,i}, Y_{P,i}), \tag{3.6}$$

   where ℓ is a loss function. This allows to evaluate model accuracy in the target domain without target labeled samples *even in the presence of concept drift between the training and target domain*.
2. **ExTRA fine-tuning for target domain performance.** Since the reweighted empirical risk (in the source domain) is a good estimate of the risk in the target domain, practitioners may fine-tune models for the target domain by minimizing the reweighted empirical risk:

$$\widehat{f}_Q \in \arg\min_{f \in \mathcal{F}} \mathbf{E}_{\widehat{P}}\left[\ell(f(X), Y) \widehat{\omega}(X, Y)\right]. \tag{3.7}$$

We note that the correctness of (3.4) depends on the identifiability of the θ_k's and α_k's from (3.3); *i.e.* the uniqueness of the parameters that satisfy (3.3). As long as the tilt parameters are identifiable, then (3.4) provides consistent estimates of them. However, without additional assumptions on the $p\{x, Y=k\}$'s and T, the tilt parameters are generally unidentifiable from (3.3). Next we elaborate on the identifiability of the exponential tilt model.

4 Theoretical properties of exponential tilting

4.1 Identifiability of the exponential tilt model

To show that the θ_k's and α_k's are identifiable from (3.3), we must show that there is a unique solution to (3.3). Unfortunately, this is not always the case. For example, consider a linear discriminant analysis (LDA) problem in which the class conditionals drift between the source and target domains:

$$p\{x, Y=k\} = \pi_k \phi(x - \mu_{P,k}), \quad q\{x, Y=k\} \quad = \pi_k \phi(x - \mu_{Q,k})\,,$$

where ϕ is the standard multivariate normal density, $\pi_k \in (0,1)$ are the class proportions in both source and target domains, and $\mu_{P,k}$'s (resp. the $\mu_{Q,k}$'s) are the class conditional means in the source (resp. target) domains. We see that this problem satisfies the exponential tilt model with $T(x) = x$:

$$\log \tfrac{q\{x,Y=k\}}{p\{x,Y=k\}} = (\mu_{Q,k} - \mu_{P,k})^\top x - \tfrac{1}{2}\|\mu_{Q,k}\|_2^2 + \tfrac{1}{2}\|\mu_{P,k}\|_2^2 \,.$$

This instance of the exponential tilt model is not identifiable. Any permutation of the class labels $\sigma : [K] \to [K]$ also leads to the same (marginal) distribution of inputs:

$$\begin{aligned}&\textstyle\sum_{k=1}^K p\{x, Y=k\}\exp\left((\mu_{Q,k} - \mu_{P,k})^\top x + \tfrac{1}{2}\|\mu_{P,k}\|_2^2 - \tfrac{1}{2}\|\mu_{Q,k}\|_2^2\right)\\ &\quad = \textstyle\sum_{k=1}^K p\{x, Y=k\}\exp\left((\mu_{Q,\sigma(k)} - \mu_{P,k})^\top x + \tfrac{1}{2}\|\mu_{P,k}\|_2^2 - \tfrac{1}{2}\|\mu_{Q,\sigma(k)}\|_2^2\right).\end{aligned}$$

From this example, we see that the non-identifiability of the exponential tilt model is closely related to the label switching problem in clustering. Intuitively, the exponential tilt model in the preceding example is too flexible because it can tilt any $p\{x, Y = k\}$ to $q\{x, Y = l\}$. Thus there is ambiguity in which $p\{x, Y = k\}$ tilts to which $q\{x, Y = l\}$. In the rest of this subsection, we present an identification restriction that guarantees the identifiability of the exponential tilt model.

A standard identification restriction in related work on domain adaptation is a clustering assumption. For example, Tachet et al. (2020) assume there is a partition of $\mathcal{X}$ into disjoint sets $\mathcal{X}_k$ such that $\text{supp}(P\{\cdot \mid Y = k\})$, $\text{supp}(Q\{\cdot \mid Y = k\}) \subset \mathcal{X}_k$ for all $k \in [K]$. This assumption is strong: it implies there is a perfect classifier in the source and target domains. Here we consider a weaker version of the clustering assumption: there are sets $\mathcal{S}_k$ such that

$$P\{Y = k \mid X \in \mathcal{S}_k\} = Q\{Y = k \mid X \in \mathcal{S}_k\} = 1.$$

We note that the $\mathcal{S}_k$'s can be much smaller than the $\mathcal{X}_k$'s; this permits the supports of $P\{\cdot \mid Y = k\}$ and $P\{\cdot \mid Y = l\}$ to overlap.

Definition 4.1 (anchor set)**.** *A set $\mathcal{S}_k \subset \mathcal{X}$ is an **anchor set** for class k if $p\{x, Y = k\} > 0$ and $p\{x, Y = l\} = 0$, $l \neq k$ for all $x \in \mathcal{S}_k$.*

Proposition 4.2 (identifiability from anchor sets)**.** *If there are anchor sets $\mathcal{S}_k$ for all K classes (in the source domain) and $T(\mathcal{S}_k)$ is p-dimensional, then there is at most one set of θ_k's and α_k's that satisfies* (3.3)*.*

This identification restriction is also closely related to the linear independence assumption in Gong et al. (2016). Inspecting the proof of proposition 4.2 (see Appendix A.3), we see that the anchor set assumption implies linear independence of $\{p_k(x)\exp(\theta_k^\top T(x) + \alpha_k)\}_{k=1}^K$ for any set of θ_k's and α_k's. We study the anchor set assumption empirically in a synthetic experiment in Appendix C. Our experiments show that the assumption is mild and is violated only under extreme data scenarios.

4.2 Consistency in estimation of the tilt parameters and the importance weights

Here, we establish a convergence rate for the estimated tilt parameters (Lemma (3.1)) and the Ex-TRA importance weights (Equation (3.1)). To simplify the notation, we define $S(x) = (1, T(x)^\top)^\top$ as the extended sufficient statistics for the exponential tilt and denote the corresponding tilt parameters as $\xi_k = (\alpha_k, \theta_k^\top)^\top$. We let $\xi_k^\star = (\alpha_k^\star, \theta_k^{\star\top})^\top$'s be the true values of the tilt parameters ξ_k's and let $\xi = (\xi_1^\top, \ldots, \xi_K^\top)^\top \in \mathbf{R}^{K(p+1)}$ be the long vector containing all the tilt parameters. We recall that estimating the parameters from the optimization stated in Lemma 3.1 requires a classifier $\widehat{\eta}_P$ on the source data. So, we define our objective for estimating ξ through a generic classifier $\eta : \mathcal{X} \to \Delta^K$. Denoting $\eta_k(x)$ as the k-th co-ordinate of $\eta(x)$ we define the expected log-likelihood objective as:

$$\mathfrak{L}(\eta, \xi) = \mathbf{E}_{Q_X}[\log\{\textstyle\sum_{k=1}^K \eta_k(X)\exp(\xi_k^\top S(X))\}] - \log[\mathbf{E}_P\{\exp(\xi_Y^\top S(X))\}]\,,$$

and its empirical version as

$$\hat{\mathfrak{L}}(\eta, \xi) = \mathbf{E}_{\widehat{Q}_X}[\log\{\textstyle\sum_{k=1}^K \eta_k(X)\exp(\xi_k^\top S(X))\}] - \log[\mathbf{E}_{\widehat{P}}\{\exp(\xi_Y^\top S(X))\}]\,.$$

To establish the consistency of MLE we first make an assumption that the loss $\xi \mapsto -\mathfrak{L}(\eta_P^\star, \xi)$ is strongly convex at the true parameter value.

Assumption 4.3. *The loss $\xi \mapsto -\mathfrak{L}(\eta_P^\star, \xi)$ is strongly convex at $\xi^\star$, i.e., there exists a constant $\mu > 0$ such that for any ξ it holds:*

$$-\mathfrak{L}(\eta_P^\star, \xi) \geq -\mathfrak{L}(\eta_P^\star, \xi^\star) - \partial_\xi \mathfrak{L}(\eta_P^\star, \xi^\star)^\top (\xi - \xi^\star) + \tfrac{\mu}{2}\|\xi - \xi^\star\|_2^2 \,.$$

We note that the assumption is a restriction on the distribution Q rather than the objective itself. For technical convenience we next assume that the feature space is bounded.

Assumption 4.4. *$\mathcal{X}$ is bounded, i.e., there exists an $M > 0$ such that $\mathcal{X} \subset B(0, M)$.*

Recall, from Lemma 3.1, that we need a fitted source classifier $\widehat{\eta}_P$ to estimate the tilt parameter: $\xi^\star$ is estimated by maximizing $\hat{\mathfrak{L}}(\widehat{\eta}_P, \xi)$ rather than the unknown $\hat{\mathfrak{L}}(\eta_P^\star, \xi)$. While analyzing the convergence of $\hat{\xi}$ we are required to control the difference $\hat{\mathfrak{L}}(\hat{\eta}_P, \xi) - \hat{\mathfrak{L}}(\eta_P^\star, \xi)$. To ensure the difference is small, assume the pilot estimate of the source regression function $\widehat{\eta}_P$ is consistent at some rate r_{n_P}.

Assumption 4.5. *Let $f_{P,k}^\star(x) = \log\{\eta_{P,k}^\star(x)\} - \frac{1}{K}\sum_{j=1}^K \log\{\eta_{P,j}^\star(x)\}$. We assume that there exist an estimators $\{\hat{f}_{P,k}(x)\}_{k=1}^K$ for $\{f_{P,k}^\star(x)\}_{k=1}^K$ such that the following holds: there exists a constant $c > 0$ and a sequence $r_{n_P} \to 0$ such that for almost surely $[\mathbb{P}_X]$ it holds*

$$\mathbf{P}(\|\hat{f}_P(x) - f_P^\star(x)\|_2 > t) \leq \mathit{exp}(-ct^2/r_{n_P}^2),\ t > 0 \,.$$

We use the estimated logits $\{\hat{f}_{P,k}(x)\}_{k=1}^K$ to construct the regression functions as $\widehat{\eta}_{P,k}(x) = \exp(\hat{f}_{P,k}(x))/\{\sum_{j=1}^K \exp(\hat{f}_{P,j}(x))\}$, which we use in the objective stated in Lemma 3.1 to analyze the convergence of the tilt parameter estimates and the ExTRA weights. With the above assumptions we're now ready to state concentration bounds for $\hat{\xi} - \xi^\star$ and $\hat{\omega} - \omega^\star$, where the true importance weight $\omega^\star$ is defined as $\omega^\star(x, y) = \exp({\xi_y^\star}^\top S(x))$.

Theorem 4.6. *Let the assumptions 4.3, 4.4 and 4.5 hold. For the sample sizes n_P, n_Q define $\alpha_{n_P,n_Q} = r_{n_P}\sqrt{\log(n_Q)} + \{(p+1)K/n_P\}^{1/2} + \{(p+1)K/n_Q\}^{1/2}$. There exists constants $k_1, k_2 > 0$ such that for any $\delta > 0$ with probability at least $1 - (2K+1)\delta$ the following hold:*

$$\|\hat{\xi} - \xi^\star\|_2 \leq k_1 \alpha_{n_P,n_Q}\sqrt{\log(1/\delta)},\ \textit{and}\ \|\hat{\omega} - \omega^\star\|_{1,P} \leq k_2 \alpha_{n_P,n_Q}\sqrt{\log(1/\delta)}.$$

In Theorem 4.6 we notice that as long as $r_{n_P}\log(n_Q) \to 0$ for $n_P, n_Q \to \infty$ we have $\alpha_{n_P,n_Q} \to 0$. This implies both *the estimated tilt parameters and the ExTRA weights converge to their true values as the sample sizes $n_P, n_Q \to \infty$.*

We next provide theoretical guarantees for the downstream tasks (1) fine-tuning and (2) target performance evaluation that we described in Section 3.

Fine-tuning We establish a generalization bound for the fitted model (3.7) using weighted-ERM on source domain. We denote $\mathcal{F}$ as the classifier hypothesis class. For $f \in \mathcal{F}$ and a weight function $\omega : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}_{\geq 0}$ define the weighted loss function and its empirical version on source data as:

$$\mathcal{L}_P(f, w) = \mathbf{E}_P[\omega(X, Y)\ell(f(X), Y)],\ \hat{\mathcal{L}}_P(f, w) = \mathbf{E}_{\widehat{P}}[\omega(X, Y)\ell(f(X), Y)] \,.$$

We also define the loss function on the target data as: $\mathcal{L}_Q(f) = \mathbf{E}_Q\left[\ell(f(X), Y)\right]$. If $\{(\theta_k^\star, \alpha_k^\star)\}_{k=1}^K$ is the true value of the tilt parameters in (3.2), *i.e.*, the following holds:

$$q\{x, Y = k\} = p\{x, Y = k\}\exp\{\alpha_k^\star + (\theta_k^\star)^\top T(x)\};\ k \in [K] \,,$$

then defining $\omega^\star(x, k) = \exp\{\alpha_k^\star + (\theta_k^\star)^\top T(x)\}$ as the true weight we notice that $\mathcal{L}_P(f, \omega^\star) = \mathcal{L}_Q(f)$, which is easily observed by setting $g(x, y) = \ell(f(x), y)$ in the display (3.1).

To establish our generalization bound we require Rademacher complexity (Bartlett & Mendelson, 2002) (denoted as $\mathcal{R}_{n_P}(\mathcal{G})$; see Appendix A.1 for details) and the following assumption on the loss function.

Assumption 4.7. *The loss function ℓ is bounded, i.e., for some $B > 0$, $|\ell\{f(x), y\}| \leq B$ for any $f \in \mathcal{F}$, $x \in \mathcal{X}$ and $y \in [K]$.*

With the above definitions and the assumption we establish our generalization bound.

Lemma 4.8. *For a weight function ω and the source samples $\{(X_{P,i}, Y_{P,i})\}_{i=1}^{n_P}$ of size n_P let $\hat{f}_\omega = \arg\min_{f\in\mathcal{F}} \hat{\mathcal{L}}_P(f, \omega)$. There exists a constant $c > 0$ such that the following generalization bound holds with probability at least $1 - \delta$*

$$\mathcal{L}_Q(\hat{f}_\omega) - \min_{f\in\mathcal{F}} \mathcal{L}_Q(f) \leq 2\mathcal{R}_{n_P}(\mathcal{G}) + B\|\omega - \omega^\star\|_{1,P} + c\sqrt{\frac{\log(1/\delta)}{n_P}}\,, \tag{4.1}$$

where $\mathcal{R}_{n_P}(\mathcal{G})$ is the Rademacher complexity of $\mathcal{G} = \{\omega^\star(x, y)\ell(f(x), y) : f \in \mathcal{F}\}$ defined in Appendix A.1.

In Theorem 4.6 we established an upper bound for the estimated weights $\hat{\omega}$, which concludes that $\hat{f}_{\hat{\omega}}$ has the following generalization bound: for any $\delta > 0$, with probability at least $1 - (2K + 2)\delta$

$$\mathcal{L}_Q(\hat{f}_{\hat{\omega}}) - \min_{f\in\mathcal{F}} \mathcal{L}_Q(f) \leq 2\mathcal{R}_{n_P}(\mathcal{G}) + k_2\alpha_{n_P,n_Q}\sqrt{\log(1/\delta)} + c\sqrt{\log(1/\delta)/n_P}\,,$$

where k_2 is the constant in Theorem 4.6 and c is the constant in Lemma 4.8. The generalization bound implies that for large sample sizes ($n_P, n_Q \to \infty$) the target accuracy of weighted ERM on source data well approximates the accuracy of ERM on target data.

Target performance evaluation We provide a theoretical guarantee for the target performance evaluation (3.6) using our importance weights. Here we only consider the functions $g : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$ which are bounded by some $B > 0$, *i.e.* $|g(x, y)| \leq B$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The simplest and the most frequently used example is the model accuracy which uses 0-1-loss as the loss function: for a model f the loss $g(x, y) = \mathbb{I}\{f(x) = y\}$ is bounded with $B = 1$. For such functions we notice that $\mathbf{E}_Q[g(X, Y)] = \mathbf{E}_P[g(X, Y)\omega^\star(X, Y)]$, as observed in display (3.1). This implies the following bound on the target performance evaluation error

$$\begin{aligned}\left|\mathbf{E}_Q[g(X, Y)] - \mathbf{E}_P[g(X, Y)\hat{\omega}(X, Y)]\right| &= \left|\mathbf{E}_P[g(X, Y)\omega^\star(X, Y)] - \mathbf{E}_P[g(X, Y)\hat{\omega}(X, Y)]\right| \\ &\leq B\mathbf{E}_P[|\hat{\omega}^\star(X, Y) - \omega^\star(X, Y)|] \leq B\|\hat{\omega} - \omega^\star\|_{1,P}\,.\end{aligned}$$

We recall the concentration bound for $\|\hat{\omega} - \omega^\star\|_{1,P}$ from Theorem 4.6 and conclude that the *estimated target performance in* (3.6) *converges to the true target performance* at rate α_{n_P,n_Q}.

5 WATERBIRDS CASE STUDY

To demonstrate the efficacy of the ExTRA algorithm for reweighing the source data we (i) verify the ability of ExTRA to upweigh samples most relevant to the target task; (ii) evaluate the utility of weights in downstream tasks such as fine-tuning and (iii) model selection.

WATERBIRDS dataset combines bird photographs from the Caltech-UCSD Birds-200-2011 (CUB) dataset (Wah et al., 2011) and the image backgrounds from the Places dataset (Zhou et al., 2017). The birds are labeled as one of $\mathcal{Y}$ = {waterbird, landbird} and placed on one of $\mathcal{A}$ = {water background, land background}. The images are divided into four groups: landbirds on land (0); landbirds on water (1); waterbirds on land (2); waterbirds on water (3). The source dataset is highly imbalanced, i.e. the smallest group (2) has 56 samples. We embed all images with a pre-trained ResNet18 (He et al., 2016). See Appendix B.1 for details.

We consider five subpopulation shift target domains: all pairs of domains with different bird types and the original test set (Sagawa et al., 2019) where all 4 groups are present with proportions vastly different from the source. For all domains, we fit ExTRA weights (with ResNet18 features as $T(x)$) from 10 different initializations and report means and standard deviations for the metrics. See Appendix B.2 for the implementation details.

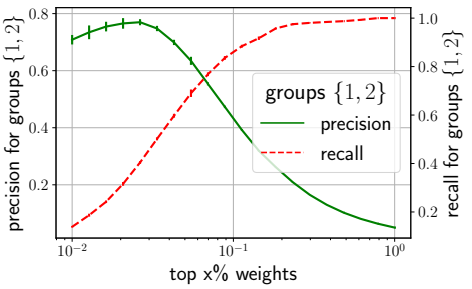

Figure 1: ExTRA precision and recall for samples with top $x\%$ weights.

ExTRA weights quality For a given target domain it is most valuable to upweigh the samples in the source data corresponding to the groups comprising

Table 1: Model selection results on WATERBIRDS

	target accuracy			rank correlation		
target groups	ExTRA	SrcVal	ATC-NE	ExTRA	SrcVal	ATC-NE
{0, 2}	0.819±0.012	0.854	**0.871**	0.419±0.01	**0.807**	0.760
{1, 2}	**0.741**±0.047	0.616	0.646	**0.747**±0.106	-0.519	-0.590
{0, 3}	**0.978**±0.001	**0.978**	0.976	**0.962**±0.004	0.956	0.906
{1, 3}	**0.757**±0.011	0.737	0.747	**0.361**±0.168	-0.318	-0.411
{0, 1, 2, 3}	**0.856**±0.034	0.803	0.818	**0.658**±0.295	0.263	0.178
average	**0.83**	0.798	0.812	**0.753**	0.166	0.110

that domain. The most challenging is the target $\{1, 2\}$ consisting only of birds appearing on their atypical backgrounds. Groups $\{1, 2\}$ correspond to 5% of the source data making them most difficult to "find". To quantify the ability of ExTRA to upweigh these samples we report precision (proportion of samples from groups $\{1, 2\}$ within the top $x\%$ of the weights) and recall (proportion of $\{1, 2\}$ samples within the top $x\%$ of the weights) in Figure 1. We notice that samples corresponding to 10% largest ExTRA weights contain slightly over 80% of the groups $\{1, 2\}$ in the source data (recall). This demonstrates the ability of ExTRA to upweigh relevant samples. We present examples of upweighted images and results for other target domains in Appendix B.3.

Model fine-tuning We demonstrate the utility of ExTRA weights in the fine-tuning (3.7). The basic goal of such importance weighing is to improve the performance in the target in comparison to training on uniform source weights S -> T, *i.e.* ERM. Another baseline is the DRO model (Hashimoto et al., 2018) that aims to maximize worst-group performance without access to the group labels. We consider two additional baselines that utilize group annotations to improve worst-group performance: re-weighing the source to equalize group proportions (RW_{gr}) and group DRO (gDRO) (Sagawa et al., 2019). The aforementioned baselines do not try to adjust to the target domain. Finally, we compare to πT -> T that fine-tunes the model only using the subset of the source samples corresponding to the target domain groups. In all cases we use logistic regression as model class.

We compare target accuracy across domains in Figure 2. Analogous comparison with area under the receiver operator curve can be found in Figure 6 in Appendix B.3.1. Model trained with ExTRA weights outperforms all "fair" baselines and matches the performance of the three baselines that had access to additional information. In all target domains ExTRA fine-tuning is comparable with the πT -> T supporting its ability to upweigh relevant samples. Notably, on {1,2} domain of both minority groups *and* on {0,3} domain of both majority groups, ExTRA outperforms RW_{gr} and gDRO that utilize group annotations. This emphasizes the advantage of adapting to the target domain instead of pursuing a more conservative goal of worst-group performance maximization. Finally, we note that ExTRA fine-tuning did not perform as well on the domain {1,3}, however in this case the πT -> T also did not do well.

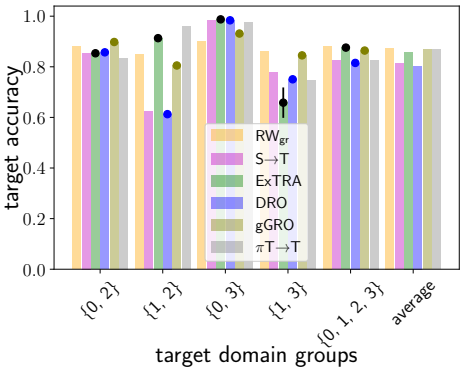

Figure 2: Performance on WATERBIRDS.

Model selection out-of-distribution is an important task, that is difficult to perform without target data labels and group annotations (Gulrajani & Lopez-Paz, 2020; Zhai et al., 2021). We evaluate the ability of choosing a model for the target domain based on accuracy on the ExTRA reweighted source validation data. We compare to the standard source validation model selection (SrcVal) and to the recently proposed ATC-NE (Garg et al., 2022) that uses negative entropy of the predicted probabilities on the target domain to score models. We fit a total of 120 logistic regression models with different weighting (uniform, label balancing, and group balancing) and varying regularizers. See Appendix B.2 for details.

In Table 1 we compare the target performance of models selected using each of the model evaluation scores and rank correlation between the corresponding model scores and true target accuracies.

Model selection with ExTRA results in the best target performance and rank correlation on 4 out of 5 domains and on average. Importantly, the rank correlation between the true performance and ExTRA model scores is always positive, unlike the baselines, suggesting its reliability in providing meaningful information about the target domain performance.

6 BREEDS CASE STUDY

BREEDS (Santurkar et al., 2020) is a subpopulation shift benchmark derived from ImageNet (Deng et al., 2009). It uses the class hierarchy to define groups within classes. For example, in the Entity-30 task considered in this experiment, class fruit is represented by strawberry, pineapple, jackfruit, Granny Smith in the source and buckeye, corn, ear, acorn in the target. This is an extreme case of subpopulation shift where source and target groups have zero overlap. We modify the dataset by adding a small fraction π of random samples from the target to the source for two reasons: (i) our exponential tilt model requires some amount of overlap between source and target; (ii) arguably, in practice, it is more likely that the source dataset has at least a small representation of all groups.

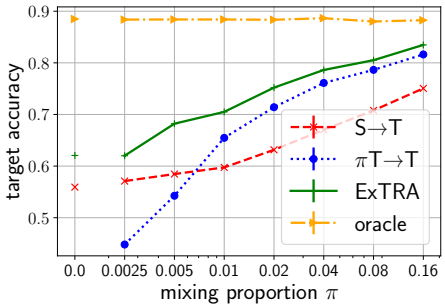

Figure 3: Performance on BREEDS.

Our goal is to show that ExTRA can identify the target samples mixed into the source for efficient fine-tuning. We obtain feature representations from a pre-trained self-supervised SwAV (Caron et al., 2020). To obtain the ExTRA weights we use SwAV features as sufficient statistic. We then train logistic regression models on (i) the source dataset re-weighted with ExTRA, (ii) uniformly weighted source (S -> T), (iii) target samples mixed into the source (πT -> T), (iv) all target samples (oracle). See Appendix B.1, B.2 for details. We report performance for varying mixing proportion π in Figure 3. First, we note that even when $\pi = 0$, i.e. source and target have completely disjoint groups (similar to domain generalization), ExTRA improves over the vanilla S -> T. Next, we see that S -> T improves very slowly in comparison to ExTRA as we increase the mixing proportion; πT -> T improves faster as we increase the number of target samples it has access to, but never suppresses ExTRA and matches its improvement slope for the larger π values. We conclude that ExTRA can effectively identify target samples mixed into source that are crucial for the success of fine-tuning *and* find source samples most relevant to the target task allowing it to outperform πT -> T. We report precision and recall analogous to the corresponding WATERBIRDS experiment in Appendix B.3.

7 CONCLUSION

In this paper, we developed an importance weighing method for approximating expectations of interest on new domains leveraging unlabeled samples (in addition to a labeled dataset from the source domain). We demonstrated the applicability of our method on downstream tasks such as model evaluation/selection and fine-tuning both theoretically and empirically. Unlike other importance weighing methods that only allow covariate shift between the source and target domains, we permit concept drift between the source and target. Though we demonstrate the efficacy of our method in synthetic setup of concept drift (Appendix C), in a future research it would be interesting to investigate the performance in more realistic setups (*e.g.* CIFAR10.2 to CIFAR10.2 (Lu et al., 2020), Imagenet to Imagenetv2 (Recht et al., 2019)).

Despite its benefits, the exponential tilt model does suffer from a few limitations. Implicit in the exponential tilt assumption is that the supports of the target class conditionals have some overlap with the corresponding source class conditionals. Although this assumption is likely satisfied in many instances of domain generalization problems (and is always satisfied in the subpopulation shift setting), an interesting avenue for future studies is to accommodate support alignment in the distribution shift model, *i.e.* to align the supports for class conditioned feature distributions in source and target domains. One way to approach this is to utilize distribution matching techniques from domain adaptation literature (Ganin et al., 2016; Sun & Saenko, 2016; Shen et al., 2018), similarly to Cai et al. (2021). We hope aligning supports via distribution matching will allow our method to succeed on domain generalization problems where the support overlap assumption is violated.

Ethical statement: We recommend considering the representation of the minority groups when applying ExTRA in the context of fairness-sensitive applications. The goal of ExTRA is to approximate the distribution of the target domain, thus, in order to use ExTRA weights for fine-tuning or model selection to obtain a fair model, the target domain should be well representative of both privileged and unprivileged groups. If the target domain has miss/under-represented groups, a model obtained using ExTRA weights may be biased.

REFERENCES

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. EM with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation. *arXiv:1901.06852 [cs, stat]*, January 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, September 2019.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized Learning for Domain Adaptation under Label Shifts. *arXiv:1903.09734 [cs, stat]*, March 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 2178–2186, Red Hook, NY, USA, December 2011. Curran Associates Inc. ISBN 978-1-61839-599-3.

Jonathon Byrd and Zachary Lipton. What is the Effect of Importance Weighting in Deep Learning? In *International Conference on Machine Learning*, pp. 872–881. PMLR, May 2019.

T. Tony Cai and Hongji Wei. Transfer Learning for Nonparametric Classification: Minimax Rate and Adaptive Classifier. *arXiv:1906.02903 [cs, math, stat]*, June 2019.

Tianle Cai, Ruiqi Gao, Jason D. Lee, and Qi Lei. A Theory of Label Propagation for Subpopulation Shift. *arXiv:2102.11203 [cs, stat]*, February 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. *Advances in Neural Information Processing Systems*, 34, 2021a.

Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. Mandoline: Model evaluation under distribution shift. In *International Conference on Machine Learning*, pp. 1617–1629. PMLR, 2021b.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment Inference for Invariant Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2189–2200. PMLR, July 2021.

Shai Ben David, Tyler Lu, Teresa Luu, and David Pal. Impossibility Theorems for Domain Adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136. JMLR Workshop and Conference Proceedings, March 2010.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Qi Dou, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain Generalization via Model-Agnostic Learning of Semantic Features. *arXiv:1910.13580 [cs]*, October 2019.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *arXiv:1610.03425 [stat]*, October 2016.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, March 2014. ISSN 0360-0300. doi: 10.1145/2523813.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, January 2016. ISSN 1532-4435.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C. Lipton. A Unified View of Label Shift Estimation. *arXiv:2003.07554 [cs, stat]*, March 2020.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. *arXiv preprint arXiv:2201.04234*, 2022.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Scholkopf. Domain Adaptation with Conditional Transferable Components. In *Proceedings of Machine Learning Research*, volume 48, pp. 17, New York, New York, USA, June 2016.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1134–1144, 2021.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*, September 2020.

Mahdi Hashemi and Hassan Karimi. Weighted machine learning. *Statistics, Optimization and Information Computing*, 6(4):497–525, 2018.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness Without Demographics in Repeated Loss Minimization. *arXiv:1806.08010 [cs, stat]*, June 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. *arXiv preprint arXiv:2110.14503*, 2021.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv:2012.07421 [cs]*, December 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding Self-Training for Gradual Domain Adaptation. *arXiv:2002.11361 [cs, stat]*, February 2020.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. February 2022. doi: 10.48550/arXiv.2202.10054.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted Empirical Risk Minimization. In *International Conference on Learning Representations*, September 2020.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *arXiv preprint arXiv:2109.06141*, 2021.

Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and Correcting for Label Shift with Black Box Predictors. *arXiv:1802.03916 [cs, stat]*, July 2018.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6781–6792. PMLR, July 2021.

Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

Subha Maity, Yuekai Sun, and Moulinath Banerjee. Minimax optimal approaches to the label shift problem. *arXiv:2003.10443 [math, stat]*, April 2020.

Subha Maity, Diptavo Dutta, Jonathan Terhorst, Yuekai Sun, and Moulinath Banerjee. A linear adjustment based approach to posterior drift in transfer learning. *arXiv:2111.10841 [stat]*, December 2021.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain Generalization via Invariant Feature Representation. January 2013.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022.

Yangjun Ruan, Yann Dubois, and Chris J Maddison. Optimal representations for covariate shift. *arXiv preprint arXiv:2201.00057*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *arXiv:1911.08731 [cs, stat]*, November 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. *arXiv:2005.04345 [cs, stat]*, August 2020.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for Subpopulation Shift. August 2020. doi: 10.48550/arXiv.2008.04859.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv:1711.08536 [stat]*, November 2017.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000. ISSN 0378-3758. doi: 10.1016/S0378-3758(00)00115-4.

Avanti Shrikumar and Anshul Kundaje. Calibration with bias-corrected temperature scaling improves domain adaptation under label shift in modern neural networks. *Preprint at https://arxiv.org/abs/1901.06852 v1*, 2019.

Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-T Approach to Unsupervised Domain Adaptation. In *International Conference on Learning Representations*, February 2018.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, USA, March 2012. ISBN 978-0-262-01709-1.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate Shift Adaptation by Importance Weighted Cross Validation. *The Journal of Machine Learning Research*, 8:985–1005, December 2007. ISSN 1532-4435.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, 2012. ISBN 978-0-521-19017-6. doi: 10.1017/CBO9781139035613.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision – ECCV 2016 Workshops*, Lecture Notes in Computer Science, pp. 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8. doi: 10.1007/978-3-319-49409-8_35.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. *arXiv:1909.13231 [cs, stat]*, July 2020.

Remi Tachet, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain Adaptation with Conditional Distribution Matching and Generalized Label Shift. March 2020.

A. W. van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 2000. ISBN 978-0-387-94640-5 978-1-4757-2547-6.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*, September 2020.

Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Hui Xu and Robert Tibshirani. Estimation of prediction error with known covariate shift. *arXiv preprint arXiv:2205.01849*, 2022.

Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. *arXiv preprint arXiv:2202.05834*, 2022.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. DORO: Distributional and Outlier Robust Optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12345–12355. PMLR, July 2021.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift. *arXiv:2007.02931 [cs, stat]*, October 2020.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A PROOFS

A.1 RADEMACHER COMPLEXITY

We next define the Rademacher complexity (Bartlett & Mendelson, 2002) that has been frequently used in machine learning literature to establish a generalization bound. Instead of considering the Rademacher complexity on $\mathcal{F}$ we define the class of weighted losses $\mathcal{G}(\ell, \mathcal{F}) = \{g_f(x, y) = w^\star(x, y)\ell(f(x), y) : \ f \in \mathcal{F}\}$ and for $n \in \mathbb{N}$ we define its Rademacher complexity measure as

$$\mathcal{R}_n(\mathcal{G}) \triangleq \mathbf{E}_{\{(u_i,v_i)\}_{i=1}^n \overset{\text{IID}}{\sim} P}\left[\mathbf{E}_{\{\xi_i\}_{i=1}^n}[\sup_{f\in\mathcal{F}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \xi_i \omega^\star(u_i, v_i)\ell\{f(u_i), v_i\}]\right]$$

where $\{\xi_i\}_{i=1}^n$ are IID

A.2 PROOF OF LEMMA 3.1

If D is the Kulback-Leibler (KL) divergence, then we can rewrite the objective in (3.4) as:

$$\begin{aligned}
&\mathbf{E}_{\widehat{Q}_X}\left[\log \widehat{q}_X\{X\} - \log\left\{\textstyle\sum_{k=1}^K \widehat{p}\{X, Y = k\}\exp(\theta_k^\top T(X) + \alpha_k)\right\}\right] \\
&\quad = \mathbf{E}_{\widehat{Q}_X}\left[\log\left\{\textstyle\sum_{k=1}^K \widehat{\eta}_{P,k}(X)\exp(\theta_k^\top T(X) + \alpha_k\right\}\right] - \mathbf{E}_{\widehat{Q}_X}\left[\log\left\{\tfrac{q_X\{x\}}{p_X\{x\}}\right\}\right],
\end{aligned}$$

where the term $-\mathbf{E}_{\widehat{Q}_X}\left[\log\left\{\frac{q_X\{x\}}{p_X\{x\}}\right\}\right]$ in our objective does not involve any tilt parameters and we drop it from our objective. To simplify the notations we define

$$\begin{aligned}
O(\theta, \alpha) &\triangleq \mathbf{E}_{\widehat{Q}_X}\left[\log\left\{\textstyle\sum_{k=1}^K \widehat{\eta}_{P,k}(X)\exp(\theta_k^\top T(X) + \alpha_k\right\}\right] \\
N(\theta, \alpha) &\triangleq \mathbf{E}_{\widehat{P}}\left[\exp(\theta_Y^\top T(X) + \alpha_Y)\right].
\end{aligned}$$

where $(\theta, \alpha) \in \mathbf{R}^q$ for $q = K(p + 1)$. In terms of O and N, (3.3) is

$$(\hat{\theta}, \hat{\alpha}) = \arg\max_{(\theta,\alpha)}\{O(\theta, \alpha) \mid N(\theta, \alpha) = 1\}. \tag{A.1}$$

Let $F_1 \triangleq \{(\theta, \alpha) \mid N(\theta, \alpha) = 1\}$ be the feasible set. We introduce a change of variables:

$$c(\theta, \alpha') = (\theta, \alpha(\theta, \alpha')) \ \text{ where } \ \alpha(\theta, \alpha') = \alpha' - 1_K \log(N(\theta, \alpha')).$$

Note that for any $(\theta, \alpha') \in \mathbf{R}^q,\ c(\theta, \alpha') \in F_1$ because it holds:

$$\begin{aligned}
N(c(\theta, \alpha')) &= N(\theta, \alpha' - \log(N(\theta, \alpha')) \times 1_K) \\
&= \mathbf{E}_{\widehat{P}}\left[\exp(\theta_Y^\top T(X) + \alpha'_Y - \log(N(\theta, \alpha')))\right] \\
&= \frac{\mathbf{E}_{\widehat{P}}\left[\exp(\theta_Y^\top T(X) + \alpha'_Y)\right]}{N(\theta, \alpha')} = 1.
\end{aligned}$$

and the objective value changes to

$$\begin{aligned}
O(c(\theta, \alpha')) &= \mathbf{E}_{\widehat{Q}_X}\left[\log\left\{\textstyle\sum_{k=1}^K \widehat{\eta}_{P,k}(X)\exp[\theta_k^\top T(X) + \alpha'_k - \log(N(\theta, \alpha')]\right\}\right] \\
&= \mathbf{E}_{\widehat{Q}_X}\left[\log\left\{\textstyle\sum_{k=1}^K \widehat{\eta}_{P,k}(X)\exp[\theta_k^\top T(X) + \alpha'_k]\right\}\right] - \log(N(\theta, \alpha')) \\
&= O(\theta, \alpha') - \log(N(\theta, \alpha')).
\end{aligned}$$

Defining $F_2 = \{c(\theta, \alpha') \mid (\theta, \alpha') \in \mathbf{R}^q\}$ we see that $F_1 = F_2$, whose argument follows. We first notice that $F_2 \subset F_1$ since for any $(\theta, \alpha') \in \mathbf{R}^q$ it holds $c(\theta, \alpha') \in F_1$. We also notice that $F_1 \subset F_2$. This is argued by noticing the following: if $(\theta, \alpha') \in F_1$ then $N(\theta, \alpha') = 1$ which implies $c(\theta, \alpha') = (\theta, \alpha')$.

Here, we summarize the crux of the proof. Though there are multiple (θ, α') that produces the same value of $c(\theta, \alpha')$, each of these (θ, α')'s produce the same value for the objective

$$O(\theta, \alpha') - \log(N(\theta, \alpha')) = O(c(\theta, \alpha')),$$

and $c(\theta, \alpha')$ always satisfy the constraint. So, the optimal point (θ, α) of (A.1) corresponds to multiple (θ, α')'s and each of them maximizes $O(\theta, \alpha') - \log(N(\theta, \alpha'))$. Furthermore, we obtain the

optimal (θ, α) from any of (θ, α') (which optimizes $O(\theta, \alpha') - \log(N(\theta, \alpha'))$) using the transformation $c(\theta, \alpha')$.

The mathematical description of the change of variable follows. With the change of variable we rewrite (A.1) as

$$(\hat{\theta}, \hat{\alpha}) = c(\hat{\theta}, \hat{\alpha}'),\ (\hat{\theta}, \hat{\alpha}') = \arg\max_{(\theta,\alpha')} \{O(c(\theta, \alpha')) \mid N(c(\theta, \alpha')) = 1, (\theta, \alpha') \in \mathbf{R}^q\}$$
$$\text{or } (\hat{\theta}, \hat{\alpha}) = c(\hat{\theta}, \hat{\alpha}'),\ (\hat{\theta}, \hat{\alpha}') = \arg\max_{(\theta,\alpha')} \{O(\theta, \alpha') - \log(N(\theta, \alpha')) \mid (\theta, \alpha') \in \mathbf{R}^q\}$$

where the constraint disappear because $N(c(\theta, \alpha')) = 1$ for any $(\theta, \alpha') \in \mathbf{R}^q$. This completes the proof.

A.3 Proof of proposition 4.2

Suppose there are two sets of tilt parameters (θ_k, α_k)'s and (η_k, β_k)'s that satisfy (3.3):

$$q_X\{x\} = \sum_{k=1}^{K} p\{x, Y = k\}\exp(\theta_k^\top T(x) + \alpha_k) = \sum_{k=1}^{K} p\{x, Y = k\}\exp(\eta_k^\top T(x) + \beta_k).$$

For any $x \in \mathcal{S}_k$, the terms that include $p\{x, Y = l\}$, $l \neq k$ vanish:

$$p\{x, Y = k\}\exp(\theta_k^\top T(x) + \alpha_k) = p\{x, Y = k\}\exp(\eta_k^\top T(x) + \beta_k).$$

This implies

$$\theta_k^\top T(x) + \alpha_k = \eta_k^\top T(x) + \beta_k \text{ for all } x \in \mathcal{S}_k.$$

We conclude $\theta_k = \eta_k$ and $\alpha_k = \beta_k$ because $T(\mathcal{S}_k)$ is p-dimensional, so there are p points $x_1, \ldots, x_p \in \mathcal{S}_k$ such that $T(x_1), \ldots, T(x_p) \in \mathbf{R}^p$ are linearly independent.

A.4 Proof of Theorem 4.6

For a probabilistic classifier $\eta : \mathcal{X} \to \Delta^K$ and the parameter $\xi = (\xi_1^\top, \ldots, \xi_K^\top)^\top$ we define the centered logit function $f : \mathcal{X} \to \mathbf{R}^K$ as $f_a(x) = \log\{\eta_a(x)\} - \frac{1}{K}\sum_{b=1}^{K}\log\{\eta_b(x)\}$. We define the functions $u_k(f, \xi) = \eta_a(x)\exp(\xi_k^\top S(x))$, $u_\cdot(f, \xi) = \sum_{k=1}^{K} u_k(f, \xi)$ and $v_k(f, \xi) = u_k(f, \xi)/u_\cdot(f, \xi)$, and notice that the objective is

$$\hat{L}(f, \xi) = \mathbf{E}_{\widehat{Q}_X}[\log\{u_\cdot(f, \xi)\}] - \log\{\mathbf{E}_{\widehat{P}}[\exp(\xi_y^\top S(x))]\}, \tag{A.2}$$

whereas the true objective is

$$L^\star(f, \xi) = \mathbf{E}_{Q_X}[\log\{u_\cdot(f, \xi)\}] - \log\{\mathbf{E}_{P}[\exp(\xi_y^\top S(x))]\}. \tag{A.3}$$

We see that the first order optimality conditions in estimating $\hat{\xi}$ are

$$\begin{aligned}
0 &= \partial_{\xi_a}\hat{L}(\hat{f}, \hat{\xi}) \\
&= \partial_{\xi_a}\left[\mathbf{E}_{\widehat{Q}_X}[\log\{u_\cdot(\hat{f}, \hat{\xi})\}] - \log\{\mathbf{E}_{\widehat{P}}[\exp(\hat{\xi}_y^\top S(x))]\}\right] \\
&= \mathbf{E}_{\widehat{Q}_X}\left[\partial_{\xi_a}\{u_\cdot(\hat{f}, \hat{\xi})\}/u_\cdot(\hat{f}, \xi)\right] - \frac{\partial_{\xi_a}\{\mathbf{E}_{\widehat{P}}[\exp(\hat{\xi}_y^\top S(x))]\}}{\mathbf{E}_{\widehat{P}}[\exp(\hat{\xi}_y^\top S(x))]} \\
&= \mathbf{E}_{\widehat{Q}_X}[S(x)v_a(\hat{f}, \hat{\xi})] - \mathbf{E}_{\widehat{P}}[S(x)\exp(\hat{\xi}_a^\top S(x))\mathbf{I}\{y = a\}]
\end{aligned} \tag{A.4}$$

where the last inequality holds because $\mathbf{E}_{\widehat{P}}[\exp(\hat{\xi}_y^\top S(x))] = 1$. Similarly, the first order optimality condition at truth (for $\xi^\star$) are

$$\begin{aligned}
0 &= \partial_{\xi_a} L^\star(f^\star, \xi^\star) \\
&= \mathbf{E}_{Q_X}[S(x)v_a(f^\star, \xi^\star)] - \mathbf{E}_{P}[S(x)\exp({\xi_a^\star}^\top S(x))\mathbf{I}\{y = a\}] \\
&= \mathbf{E}_{Q_X}[S(x)v_a(f^\star, \xi^\star)] - \mathbf{E}_{P_X}[S(x)\exp({\xi_a^\star}^\top S(x))\eta^\star_{P,a}(x)]
\end{aligned} \tag{A.5}$$

We decompose A.4 using the Taylor expansion and obtain:

$$0 = \partial_{\xi_a}\hat{L}(f^\star, \hat{\xi}) + \langle \hat{f} - f^\star, \partial_f\partial_{\xi_a}\hat{L}(\tilde{f}, \hat{\xi})\rangle \tag{A.6}$$

where $\tilde{f}$ is a function in the bracket $[f^\star, \hat{f}]$, *i.e.* for every x, $\tilde{f}(x)$ is a number between $\hat{f}(x)$ and $f^\star(x)$

Bound on $\langle \hat{f} - f^\star, \partial_f \partial_{\xi_a} \hat{L}(\tilde{f}, \hat{\xi})\rangle$: To bound the term we define $\zeta = \hat{f} - f^\star$ and notice that

$$\begin{aligned}\langle \zeta, \partial_f \partial_{\theta_a} \hat{L}(f, \hat{\xi})\rangle &= \textstyle\sum_b \langle \zeta_b, \partial_{f_b} \partial_{\theta_a} \hat{L}(f, \hat{\xi})\rangle \\ &= \textstyle\sum_b \langle \zeta_b, \partial_{f_b} \{\mathbf{E}_{\widehat{Q}_X}[S(x) v_a(f, \hat{\xi})]\}\rangle \\ &= \textstyle\sum_b \mathbf{E}_{\widehat{Q}_X}[S(x)\zeta_b(x)\partial_{f_b}\{v_a(f, \hat{\xi})\}] \\ &= \textstyle\sum_b \mathbf{E}_{\widehat{Q}_X}[S(x)\zeta_b(x) v_a(f, \hat{\xi})(\delta_{ab} - v_a(f, \hat{\xi}))]\,.\end{aligned}$$

The derivative in third equality in the above display is calculated in Lemma A.2. Here, from Assumption 4.4 we notice that $S(x)$ is bounded, *i.e.*, there exists a $c_1 > 0$ such that $\|S(x)\|_2 \le c_1$ for all $x \in \mathcal{X}$. This implies the followings: we have

$$\|\mathbf{E}_{\widehat{Q}_X}[S(x)\zeta_b(x) v_a(f, \xi)(\delta_{ab} - v_a(f, \xi))]\|_2 \le c_1 \mathbf{E}_{\widehat{Q}_X}[|\zeta_b(x)|]$$

since $0 \le v_a(f, \xi)(\delta_{ab} - v_a(f, \xi)) \le 1$ and we have

$$\|\langle \zeta, \partial_f \partial_{\theta_a} \hat{L}(f, \hat{\xi})\rangle\|_2 \le \textstyle\sum_b c_1 \mathbf{E}_{\widehat{Q}_X}[|\zeta_b(x)|] \le c_1\sqrt{K}\mathbf{E}_{\widehat{Q}_X}[\|\zeta(x)\|_2]\,.$$

It follows from Assumption 4.5: with probability at least $1 - \delta$ it holds $\sup_{i\in[n_Q]} \|\hat{f}(x_{Q,i}) - f^\star(x_{Q,i})\|_2 \le c_2 r_{n_P}\sqrt{\log(n_Q)\log\{1/\delta\}}$, we conclude that

$$|\langle \hat{f} - f^\star, \partial_f \partial_{\theta_a} \hat{L}(f, \theta)\rangle| \le \sqrt{K} c_1 c_2 r_{n_P}\sqrt{\log(n_Q)\log\{1/\delta\}} \tag{A.7}$$

holds with probability at least $1 - \delta$.

The term $\partial_{\xi_a} \hat{L}(f^\star, \hat{\xi})$ We use strong convexity 4.3 and convergence of the loss that

$$\sup_{\xi\in K} |\hat{L}(f^\star, \xi) - L^\star(f^\star, \xi)| \overset{n_P, n_Q\to\infty}{\longrightarrow} 0$$

for any compact K in van der Vaart & Wellner (2000, Corollary 3.2.3.) to conclude that $\hat{\xi} \to \xi^\star$ in probability and hence $\hat{\xi}$ is a consistent estimator for $\xi^\star$.

Following the consistency of $\hat{\xi}$ we see that for sufficiently large n_P, n_Q we have $\|\hat{\xi} - \xi^\star\|_2 \le \delta_\xi$ (δ_ξ is chosen according to Lemma A.1) with probability at least $1 - \delta$ and on the event it holds: $\|\hat{\xi}\|_2 \le \|\xi^\star\|_2 + \delta_\xi$. We define empirical process

$$Z_{a,n_P,n_Q} = \sup_{\|\xi\|_2 \le \|\xi^\star\|_2 + 1} \|\partial_{\xi_a} \hat{L}(f^\star, \xi) - \partial_{\xi_a} L^\star(f^\star, \xi)\|_2 \tag{A.8}$$

for which we shall provide a high probability upper bound. We denote $Z_{a,n_P,n_Q}(\xi) = \partial_{\xi_a} \hat{L}(f^\star, \xi) - \partial_{\xi_a} L^\star(f^\star, \xi)$ and notice that

$$\begin{aligned}&\partial_{\xi_a} \hat{L}(f^\star, \xi) - \partial_{\xi_a} L^\star(f^\star, \xi) \\ &= \underbrace{\mathbf{E}_{\widehat{Q}_X}[S(x) v_a(f^\star, \xi)] - \mathbf{E}_{Q_X}[S(x) v_a(f^\star, \xi)]}_{\triangleq \mathbf{A}(\xi)} \\ &\quad \underbrace{-\frac{\mathbf{E}_{\widehat{P}}[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}]}{\mathbf{E}_{\widehat{P}}[\exp(\xi_y^\top S(x))]} + \frac{\mathbf{E}_P[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}]}{\mathbf{E}_P[\exp(\xi_y^\top S(x))]}}_{\triangleq \mathbf{B}(\xi)} = \mathbf{A}(\xi) + \mathbf{B}(\xi)\end{aligned}$$

where to bound $\mathbf{A}(\xi)$ we notice that $S(x_{Q,i}) v_a(f^\star, \xi)$ are IID and bounded by c_1 ($\|S(x) v_a(f^\star, \xi)\| \le \|S(x)\|_2 \le c_1$ for all $x \in \mathcal{X}$) and hence sub-gaussian. We apply Hoeffding's concentration inequality for a sample mean of IID sub-gaussian random variables and obtain a constant $c_2 > 0$ such that for any $\delta > 0$ with probability at least $1 - \delta$ it holds

$$\mathbf{A}(\xi) = \mathbf{E}_{\widehat{Q}_X}[S(x) v_a(f^\star, \xi)] - \mathbf{E}_{Q_X}[S(x) v_a(f^\star, \xi)] \le c_1 c_2 \sqrt{\tfrac{\log(1/\delta)}{n_Q}}\,.$$

Using a chaining argument over an ℓ_2 ball of radius $\|\xi^\star\|_2 + \delta_\xi$ we obtain a uniform bound as the following: there exists a constant $c_2 > 0$ such that for any $\delta > 0$ with probability at least $1 - \delta$ it holds

$$\sup_{\xi:\|\xi\|_2 \le \|\xi^\star\|_2 + \delta_\xi} \mathbf{A}(\xi) \le c_1 c_3 \sqrt{\frac{K(p+1)\log(1/\delta)}{n_Q}} \,. \tag{A.9}$$

To bound $\mathbf{B}$ we first define

$$\begin{aligned}
\mathbf{B}.1(\xi, n_P) &\triangleq \mathbf{E}_{\widehat{P}}[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}] - \mathbf{E}_P[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}] \\
\mathbf{B}.2(\xi, n_P) &\triangleq \mathbf{E}_{\widehat{P}}[\exp(\xi_a^\top S(x))] - \mathbf{E}_P[\exp(\xi_a^\top S(x))]
\end{aligned}$$

and notice that both the random variables $\{S(x_{P,i})\exp(\xi_a^\top S(x_{P,i}))\mathbf{I}\{y_{P,i} = a\}\}_{i=1}^{n_P}$ and $\{\exp(\xi_a^\top S(x_{P,i}))\}_{i=1}^{n_P}$ are bounded for all $\|\xi\|_2 \le \|\xi^\star\|_2 + \delta_\xi$. Similarly as before we obtain constant $c_4, c_5 > 0$ such that the following hold with probability at least $1 - \delta$:

$$\begin{aligned}
\sup_{\xi:\|\xi\|_2 \le \|\xi^\star\|_2 + \delta_\xi} \big|\mathbf{B}.1(\xi, n_P)\big| &\le c_4 \sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}} \\
\sup_{\xi:\|\xi\|_2 \le \|\xi^\star\|_2 + \delta_\xi} \big|\mathbf{B}.2(\xi, n_P)\big| &\le c_5 \sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}} \,.
\end{aligned} \tag{A.10}$$

In Lemma A.1 we notice that

$$\inf_{\xi:\|\xi\|_2 \le \|\xi^\star\|_2 + \delta_\xi} \mathbf{E}_P[\exp(\xi_y^\top S(x))] \ge \frac{1}{2} \,. \tag{A.11}$$

Gathering all the inequalities in $\mathbf{B}$ we obtain

$$\begin{aligned}
\mathbf{B}(\xi) &= -\frac{\mathbf{E}_{\widehat{P}}[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}]}{\mathbf{E}_{\widehat{P}}[\exp(\xi_y^\top S(x))]} + \frac{\mathbf{E}_P[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}]}{\mathbf{E}_P[\exp(\xi_y^\top S(x))]} \\
&= -\frac{\mathbf{E}_P[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}] + \mathbf{B}.1(\xi, n_P)}{\mathbf{E}_P[\exp(\xi_y^\top S(x))] + \mathbf{B}.2(\xi, n_P)} + \frac{\mathbf{E}_P[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}]}{\mathbf{E}_P[\exp(\xi_y^\top S(x))]} \\
&= \frac{-\mathbf{B}.1(\xi, n_P)\mathbf{E}_P[\exp(\xi_a^\top S(x))] + \mathbf{B}.2(\xi, n_P)\mathbf{E}_P[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}]}{\Big\{\mathbf{E}_P[\exp(\xi_y^\top S(x))] + \mathbf{B}.2(\xi, n_P)\Big\}\mathbf{E}_P[\exp(\xi_y^\top S(x))]}
\end{aligned}$$

and this implies

$$|\mathbf{B}(\xi)| \le \frac{\|\mathbf{B}.1(\xi, n_P)\|_2 \mathbf{E}_P[\exp(\xi_a^\top S(x))] + |\mathbf{B}.2(\xi, n_P)| \|\mathbf{E}_P[S(x)\exp(\xi_a^\top S(x))\mathbf{I}\{y = a\}]\|_2}{\Big\{\mathbf{E}_P[\exp(\xi_y^\top S(x))] - |\mathbf{B}.2(\xi, n_P)|\Big\}\mathbf{E}_P[\exp(\xi_y^\top S(x))]}$$

where we use (A.10) and (A.11) to obtain a constant $c_6 > 0$ such that with probability at least $1 - \delta$ it holds

$$\sup_{\xi:\|\xi\|_2 \le \|\xi^\star\|_2 + \delta_\xi} |\mathbf{B}(\xi)| \le c_6 \sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}} \,. \tag{A.12}$$

We now combine (A.9) and (A.12) and obtain a constant $c_7 > 0$ such that with probability at least $1 - 2\delta$ we have

$$Z_{a,n_P,n_Q} \le c_7 \left\{ \sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}} + \sqrt{\frac{K(p+1)\log(1/\delta)}{n_Q}} \right\} \,. \tag{A.13}$$

Returning to the first order optimality condition (A.6) for estimating $\hat{\xi}$ we notice that

$$\begin{aligned}
0 &= \textstyle\sum_a (\hat{\xi}_a - \xi_a^\star)^\top \big\{\partial_{\xi_a} \hat{L}(f^\star, \hat{\xi}) + \langle \hat{f} - f^\star, \partial_f \partial_{\xi_a} \hat{L}(\tilde{f}, \hat{\xi})\rangle\big\} \\
&= \textstyle\sum_a (\hat{\xi}_a - \xi_a^\star)^\top \partial_{\xi_a} L^\star(f^\star, \hat{\xi}) + \sum_a (\hat{\xi}_a - \xi_a^\star)^\top \big\{Z_{a,n_P,n_Q}(\hat{\xi}) + \langle \hat{f} - f^\star, \partial_f \partial_{\xi_a} \hat{L}(\tilde{f}, \hat{\xi})\rangle\big\} \\
&= (\hat{\xi} - \xi^\star)^\top \partial_\xi L^\star(f^\star, \hat{\xi}) + \textstyle\sum_a (\hat{\xi}_a - \xi_a^\star)^\top \big\{Z_{a,n_P,n_Q}(\hat{\xi}) + \langle \hat{f} - f^\star, \partial_f \partial_{\xi_a} \hat{L}(\tilde{f}, \hat{\xi})\rangle\big\}
\end{aligned}$$

We combine it with the first ordder optimality condition for $\xi^\star$ (A.5) to obtain

$$\begin{aligned}
&(\hat{\xi} - \xi^\star)^\top \big\{\partial_\xi L^\star(f^\star, \hat{\xi}) - \partial_\xi L^\star(f^\star, \xi^\star)\big\} \\
&+ \textstyle\sum_a (\hat{\xi}_a - \xi_a^\star)^\top \big\{Z_{a,n_P,n_Q}(\hat{\xi}) + \langle \hat{f} - f^\star, \partial_f \partial_{\xi_a} \hat{L}(\tilde{f}, \hat{\xi})\rangle\big\} = 0
\end{aligned}$$

which can be rewritten as

$$(\hat{\xi}-\xi^\star)^\top\big\{\partial_\xi L^\star(f^\star,\hat{\xi})-\partial_\xi L^\star(f^\star,\xi^\star)\big\} = -\textstyle\sum_a(\hat{\xi}_a-\xi_a^\star)^\top\big\{Z_{a,n_P,n_Q}(\hat{\xi})+\langle\hat{f}-f^\star,\partial_f\partial_{\xi_a}\hat{L}(\tilde{f},\hat{\xi})\rangle\big\}\,. \tag{A.14}$$

Using the strong convexity assumption at $\xi^\star$ we obtain that the left hand side in the above equation is lower bounded as

$$(\hat{\xi}-\xi^\star)^\top\big\{\partial_\xi L^\star(f^\star,\hat{\xi})-\partial_\xi L^\star(f^\star,\xi^\star)\big\} \geq \mu\|\hat{\xi}-\xi^\star\|_2^2\,. \tag{A.15}$$

Let $\mathcal{E}$ be the event on which the following hold:

1. $\|\hat{\xi}-\xi^\star\|_2 \leq \delta_\xi$,
2. $|\langle\hat{f}-f^\star,\partial_f\partial_{\xi_a}\hat{L}(\tilde{f},\hat{\xi})\rangle| \leq \sqrt{K}c_1c_2r_{n_P}\sqrt{\log(n_Q)\log\{1/\delta\}}$ for all a,
3. $Z_{a,n_P,n_Q} \leq c_7\Big\{\sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_Q}}\Big\}$ for all a.

We notice that the event $\mathcal{E}$ has probability $1-(2K+1)\delta$. Under the event there exists a $c_9>0$ such that the right hand side in (A.14) is upper bounded as

$$\begin{aligned}
&\Big|-\textstyle\sum_a(\hat{\xi}_a-\xi_a^\star)^\top\big\{Z_{a,n_P,n_Q}(\hat{\xi})+\langle\hat{f}-f^\star,\partial_f\partial_{\xi_a}\hat{L}(\tilde{f},\hat{\xi})\rangle\big\}\Big| \\
&\leq \textstyle\sum_a\|\hat{\xi}_a-\xi_a^\star\|_2\Big\{\|Z_{a,n_P,n_Q}(\hat{\xi})\|_2+\big\|\langle\hat{f}-f^\star,\partial_f\partial_{\xi_a}\hat{L}(\tilde{f},\hat{\xi})\rangle\big\|_2\Big\} \\
&\leq \textstyle\sum_a\|\hat{\xi}_a-\xi_a^\star\|_2\Big\{Z_{a,n_P,n_Q}+\big\|\langle\hat{f}-f^\star,\partial_f\partial_{\xi_a}\hat{L}(\tilde{f},\hat{\xi})\rangle\big\|_2\Big\} \\
&\leq \textstyle\sum_a\|\hat{\xi}_a-\xi_a^\star\|_2c_9\Big\{r_{n_P}\sqrt{\log(n_Q)\log\{1/\delta\}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_Q}}\Big\} \\
&\leq c_9\Big\{r_{n_P}\sqrt{\log\{n_Q/\delta\}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_Q}}\Big\}\sqrt{K}\|\hat{\xi}-\xi^\star\|_2
\end{aligned} \tag{A.16}$$

Combining the bounds (A.15) and (A.16) for left and right hand sides we obtain a $c_{10}>0$ such that on the event $\mathcal{E}$ it holds

$$\|\hat{\xi}-\xi^\star\|_2 \leq c_{10}\Big\{r_{n_P}\sqrt{\log(n_Q)\log\{1/\delta\}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_Q}}\Big\}\,.$$

Having a concentration on $\hat{\xi}$ we now notice that

$$\begin{aligned}
&\|\hat{\omega}-\omega^\star\|_{1,P} \\
&= \textstyle\int\big|\exp(\hat{\xi}_y^\top S(x))-\exp({\xi_y^\star}^\top S(x))\big|p_X(x)dx \\
&= \textstyle\int|(\hat{\xi}_y-\xi_y^\star)^\top S(x)|\exp(\xi_x)p_X(x)dx
\end{aligned}$$

where ξ_x is a number between $\hat{\xi}_y^\top S(x)$ and ${\xi_y^\star}^\top S(x)$. On the event $\mathcal{E}$ we notice that $\|\hat{\xi}\|_2 \leq \|\xi^\star\|_2+\delta_\xi$ and hence it holds: $|\hat{\xi}_y^\top S(x)| \leq \|\hat{\xi}_y\|_2\|S(x)\|_2 \leq \|\hat{\xi}\|_2\|S(x)\|_2 \leq c_1(\|\xi^\star\|_2+\delta_\xi)$. Furthermore, we notice that $|{\xi_y^\star}^\top S(x)| \leq \|\xi^\star\|_2\|S(x)\|_2 \leq c_1\|\xi^\star\|_2$, which implies $|\xi_x| \leq c_1(\|\xi^\star\|_2+\delta_\xi)$. Returning to the integral we obtain that on the event $\mathcal{E}$ it holds:

$$\begin{aligned}
&\textstyle\int|(\hat{\xi}_y-\xi_y^\star)^\top S(x)|\exp(\xi_x)p_X(x)dx \\
&\leq \|\hat{\xi}_y-\xi_y^\star\|_2\textstyle\int\|S(x)\|_2\exp\{c_1(\|\xi^\star\|_2+\delta_\xi)\}p_X(x)dx \\
&\leq \|\hat{\xi}-\xi^\star\|_2c_1\exp\{c_1(\|\xi^\star\|_2+\delta_\xi)\} \\
&\leq c_{11}\Big\{r_{n_P}\sqrt{\log(n_Q)\log\{1/\delta\}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_P}}+\sqrt{\frac{K(p+1)\log(1/\delta)}{n_Q}}\Big\}
\end{aligned}$$

for some $c_{11}>0$, which holds with probability at least $1-(2K+1)\delta$.

Lemma A.1. *There exists $\delta_\xi>0$ such that*

$$\inf_{\xi:\|\xi\|_2\leq\|\xi^\star\|_2+\delta_\xi}\mathbf{E}_P[\mathit{exp}(\xi_y^\top S(x))] \geq \frac{1}{2}\,.$$

Proof of Lemma A.1. To establish a bound on $\mathbf{E}_P[\exp(\xi_y^\top S(x))] - \mathbf{E}_P[\exp(\xi_y^{\star\top} S(x))]$ for any $\|\xi - \xi^\star\|_2 \leq \delta_\xi$ we notice that

$$\begin{aligned}
&|\mathbf{E}_P[\exp(\xi_y^\top S(x))] - \mathbf{E}_P[\exp(\xi_y^{\star\top} S(x))]| \\
&= \int \left|\exp(\xi_y^\top S(x)) - \exp(\xi_y^{\star\top} S(x))\right| p_X(x)dx \\
&= \int |(\xi_y - \xi_y^\star)^\top S(x)| \exp(\xi_x) p_X(x)dx
\end{aligned}$$

where ξ_x is a number between $\xi_y^\top S(x)$ and $\xi_y^{\star\top} S(x)$. We notice that $\|\xi\|_2 \leq \|\xi^\star\|_2 + \delta_\xi$ and hence it holds: $|\xi_y^\top S(x)| \leq \|\xi_y\|_2 \|S(x)\|_2 \leq \|\xi\|_2 \|S(x)\|_2 \leq c_1(\|\xi^\star\|_2 + \delta_\xi)$. Furthermore, we notice that $|\xi_y^{\star\top} S(x)| \leq \|\xi^\star\|_2 \|S(x)\|_2 \leq c_1 \|\xi^\star\|_2$, which implies $|\xi_x| \leq c_1(\|\xi^\star\|_2 + \delta_\xi)$. Returning to the integral we obtain

$$\begin{aligned}
&\int |(\xi_y - \xi_y^\star)^\top S(x)| \exp(\xi_x) p_X(x)dx \\
&\leq \|\xi_y - \xi_y^\star\|_2 \int \|S(x)\|_2 \exp\{c_1(\|\xi\|_2 + \delta_\xi)\} p_X(x)dx \\
&\leq \delta_\xi c_1 \exp\{c_1(\|\xi^\star\|_2 + \delta_\xi)\}\,.
\end{aligned}$$

We choose $\delta_\xi > 0$ small enough such that $\delta_\xi c_1 \exp\{c_1(\|\xi^\star\|_2 + \delta_\xi)\} \leq 1/2$. Since $\mathbf{E}_P[\exp(\xi_y^{\star\top} S(x))] = 1$ we obtain that for any ξ with $\|\xi - \xi^\star\|_2 \leq 1/2$ we have

$$\mathbf{E}_P[\exp(\xi_y^\top S(x))] \geq \mathbf{E}_P[\exp(\xi_y^{\star\top} S(x))] - |\mathbf{E}_P[\exp(\xi_y^\top S(x))] - \mathbf{E}_P[\exp(\xi_y^{\star\top} S(x))]| \geq \tfrac{1}{2}\,.$$

This implies the lemma.

□

Lemma A.2 (Derivatives)**.** *The following holds:*

1. $\partial_{\theta_b} u_a(f,\theta) = T(x) u_a(f,\theta)\delta_{a,b}$, $\partial_{\theta_b} u.(f,\theta) = T(x) u_b(f,\theta)$ *and* $\partial_{\theta_b} v_a(f,\theta) = T(x) v_a(f,\theta)\{\delta_{a,b} - v_b(f,\theta)\}$.
2. $\partial_{f_b}\eta_a = \eta_a(\delta_{a,b} - \eta_b)$, $\partial_{f_b}\{u_a(f,\theta)\} = (\delta_{a,b} - \eta_b) u_a(f,\theta)$, $\partial_{f_b}\{u.(f,\theta)\} = u_b(f,\theta) - \eta_b u.(f,\theta)$ *and* $\partial_{f_b}\{v_a(f,\theta)\} = v_a(f,\theta)(\delta_{a,b} - v_b(f,\theta))$.

Proof of lemma A.2. We calculate the derivatives one by one.

1: We notice that

$$\begin{aligned}
\partial_{\theta_b} u_a(f,\theta) &= \partial_{\theta_b}\{\eta_a(x)\exp(\theta_a^\top T(x))\} \\
&= \eta_a(x)\exp(\theta_a^\top T(x)) T(x)\delta_{a,b} = T(x) u_a(f,\theta)\delta_{a,b}
\end{aligned}$$

and that

$$\partial_{\theta_b} u.(f,\theta) = \textstyle\sum_{a=1}^K T(x) u_a(f,\theta)\delta_{a,b} = T(x) u_b(f,\theta)$$

which finally implies

$$\begin{aligned}
\partial_{\theta_b} v_a(f,\theta) &= \tfrac{\partial_{\theta_b}\{u_a(f,\theta)\} u.(f,\theta) - \partial_{\theta_b}\{u.(f,\theta)\} u_a(f,\theta)}{\{u.(f,\theta)\}^2} \\
&= \tfrac{T(x) u_a(f,\theta)\delta_{a,b} u.(f,\theta) - T(x) u_b(f,\theta) u_a(f,\theta)}{\{u.(f,\theta)\}^2} \\
&= T(x)\{u_a/u.\}\{\delta_{a,b} - u_a/u.\} = T(x) v_a(\delta_{a,b} - v_a)
\end{aligned}$$

2: Here

$$\partial_{f_b}\eta_a = \partial_{f_b}\left\{\tfrac{e^{f_a}}{\sum_j e^{f_j}}\right\} = \tfrac{\delta_{a,b} e^{f_a} \sum_j e^{f_j} - e^{f_a} e^{f_b}}{\left\{\sum_j e^{f_j}\right\}^2} = \eta_a(\delta_{a,b} - \eta_b)\,,$$

$$\begin{aligned}
\partial_{f_b}\{u_a(f,\theta)\} &= \partial_{f_b}\{\eta_a\}\exp(\theta_a^\top T(x)) \\
&= \eta_a(\delta_{a,b} - \eta_b)\exp(\theta_a^\top T(x)) = (\delta_{a,b} - \eta_b) u_a(f,\theta)\,, \\
\partial_{f_b}\{u.(f,\theta)\} &= \textstyle\sum_a (\delta_{a,b} - \eta_b) u_a(f,\theta) = u_b(f,\theta) - \eta_b u.(f,\theta)
\end{aligned}$$

and finally,

$$\begin{aligned}\partial_{f_b}\{v_a(f,\theta)\} &= \frac{\{\partial_{f_b}u_a(f,\theta)\}u.(f,\theta) - \{\partial_{f_b}u.(f,\theta)\}u_a(f,\theta)}{\{u.(f,\theta)\}^2} \\ &= \frac{(\delta_{a,b}-\eta_b)u_a(f,\theta)u.(f,\theta) - \{u_b(f,\theta)-\eta_b u.(f,\theta)\}u_a(f,\theta)}{\{u.(f,\theta)\}^2} \\ &= (u_a/u.)\{\delta_{a,b}u. - \cancel{\eta_b u.} - u_a + \cancel{\eta_b u.}\}/u. \\ &= (u_a/u.)\{\delta_{a,b} - (u_a/u.)\} = v_a\{\delta_{a,b} - v_a\}\end{aligned}$$

□

A.5 Proof of Lemma 4.8

We start by decomposing the loss difference in the left hand side of equation (4.1).

$$\begin{aligned}\mathcal{L}_Q(\hat{f}_{\hat{\omega}}) - \mathcal{L}_Q(f^\star) &= \mathcal{L}_P(\hat{f},\omega^\star) - \mathcal{L}_P(f^\star,\omega^\star) \\ = \underbrace{\mathcal{L}_P(\hat{f},\omega^\star) - \mathcal{L}_{\widehat{P}}(\hat{f},\omega^\star)}_{(a)} &+ \underbrace{\mathcal{L}_{\widehat{P}}(\hat{f},\omega^\star) - \mathcal{L}_{\widehat{P}}(\hat{f},\hat{\omega})}_{(b)} + \underbrace{\mathcal{L}_{\widehat{P}}(\hat{f},\hat{\omega}) - \mathcal{L}_{\widehat{P}}(f^*,\hat{\omega})}_{\leq 0} \\ &+ \underbrace{\mathcal{L}_{\widehat{P}}(f^*,\hat{\omega}) - \mathcal{L}_{\widehat{P}}(f^*,\omega^\star)}_{(c)} + \underbrace{\mathcal{L}_{\widehat{P}}(f^*,\omega^\star) - \mathcal{L}_P(f^*,\omega^\star)}_{(d)},\end{aligned} \tag{A.17}$$

where we write $\hat{f}_{\hat{\omega}} \equiv \hat{f}$.

Uniform bound on (a) To control (a) in (4.1) we establish a concentration bound on the following generalization error

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left\{\mathcal{L}_P(f,\omega^\star) - \mathcal{L}_{\widehat{P}}(f,\omega^\star)\right\} \\ &= \sup_{f\in\mathcal{F}}\left\{\mathbf{E}\big[g_f(X,Y)\big] - \tfrac{1}{n_P}\textstyle\sum_{i=1}^{n_P} g_f(X_{P,i},Y_{P,i})\right\} \triangleq F(z_{1:n_P})\,,\end{aligned}$$

where, for $i \geq 1$ we denote $z_{1:i} = (z_1,\ldots,z_i)$ and $z_i = (X_{P,i}, Y_{P,i})$. First, we use a modification of McDiarmid concentration inequality to bound $F(z_{1:n_P})$ in terms of its expectation and a $O(1/\sqrt{n_P})$ term, as elucidated in the following lemma.

Lemma A.3. *There exists a constant $c_1 > 0$ such that with probability at least $1-\delta$ the following holds*

$$F(z_{1:n_P}) \leq \mathbf{E}\big[F(z_{1:n_P})\big] + c_1\sqrt{\tfrac{\log(1/\delta)}{n_P}}\,. \tag{A.18}$$

Next, we use a symmetrization argument (see (Wellner et al., 2013, Chapter 2, Lemma 2.3.1)) to bound the expectation $\mathbf{E}\big[F(z_{1:n_P})\big]$ by the Rademacher complexity of the hypothesis class $\mathcal{G}$, *i.e.*,

$$\mathbf{E}\big[F(z_{1:n_P})\big] \leq 2\mathcal{R}_{n_P}(\mathcal{G})\,. \tag{A.19}$$

Combining (A.18) and (A.19) we obtain

$$(a) = \mathcal{L}_P(\hat{f},\omega^\star) - \mathcal{L}_{\widehat{P}}(\hat{f},\omega^\star) \leq 2\mathcal{R}_{n_P}(\mathcal{G}) + c_1\sqrt{\tfrac{\log(1/\delta)}{n_P}}\,, \tag{A.20}$$

with probability at least $1-\delta$.

Bound on (b) and (c) Denoting $z_i = (X_{P,i}, Y_{P,i})$ and $\ell_f(z_i) = \ell(f(X_{P,i}), Y_{P,i})$ we notice that for any $f \in \mathcal{F}$ we have

$$\begin{aligned}\left|\mathcal{L}_{\widehat{P}}(f,\omega^\star) - \mathcal{L}_{\widehat{P}}(f,\hat{\omega})\right| &= \left|\tfrac{1}{n_P}\textstyle\sum_{i=1}^{n_P}\left\{\hat{\omega}(z_i) - \omega^\star(z_i)\right\}\ell_f(z_i)\right| \\ &\leq \tfrac{1}{n_P}\textstyle\sum_{i=1}^{n_P}\left|\left\{\hat{\omega}(z_i) - \omega^\star(z_i)\right\}\ell_f(z_i)\right| \leq \tfrac{\|\ell\|_\infty}{n_P}\textstyle\sum_{i=1}^{n_P}\left|\hat{\omega}(z_i) - \omega^\star(z_i)\right|.\end{aligned}$$

Since $\hat{\omega}(z) - \omega^\star(z)$ is a sub-gaussian random variable, we use sub-gaussian concentration to establish that for some constant $c_2 > 0$

$$\text{for any } f\in\mathcal{F},\ \left|\mathcal{L}_{\widehat{P}}(f,\omega^\star) - \mathcal{L}_{\widehat{P}}(f,\hat{\omega})\right| \leq \|\ell\|_\infty\left\{\mathbf{E}_{z_1}\left[|\hat{w}(z_1) - w^\star(z_1)|\right] + c_2\sqrt{\tfrac{\log(1/\delta)}{n_P}}\right\} \tag{A.21}$$

with probability at least $1-\delta$. This provides a simultaneous bound (on the same probability event) for both (b) and (c) with $f = \hat{f}$ and $f = f^\star$.

Bound on (d) We note that

$$\mathcal{L}_{\widehat{P}}(f^*,\omega^\star) - \mathcal{L}_P(f^*,\omega^\star) = \tfrac{1}{n_P}\sum_{i=1}^{n_P}\omega^\star(z_i)\ell_{f^\star}(z_i) - \mathbf{E}_P\big\{\omega^\star(z_1)\ell_{f^\star}(z_1)\big\}\,,$$

where $\{\omega^\star(z_i)\ell_{f^\star}(z_i)\}_{i=1}^{n_P}$ are IID sub-gaussian random variables. Using Hoeffding concentration bound we conclude that there exists a constant $c_3 > 0$ such that for any $\delta > 0$ the following holds with probability at least $1-\delta$

$$\tfrac{1}{n_P}\sum_{i=1}^{n_P}\omega^\star(z_i)\ell_{f^\star}(z_i) - \mathbf{E}_P\big\{\omega^\star(z_1)\ell_{f^\star}(z_1)\big\} \le c_3\sqrt{\tfrac{\log(1/\delta)}{n_P}}\,. \tag{A.22}$$

Finally, using (A.20) on (a) (which is true on an event of probability $\ge 1-\delta$), (A.21) on (b) and (c) (simultaneously true on an event of probability $\ge 1-\delta$), and (A.22) on (d) (holds on an event of probability $\ge 1-\delta$) we conclude that with probability at least $1-3\delta$ the following holds

$$\mathcal{L}_Q(\hat{f}_{\hat{w}}) - \mathcal{L}_Q(f^\star) \le 2\mathcal{R}_n(\mathcal{G}) + \|\ell\|_\infty \cdot \mathbf{E}_{z_1}\big[|\hat{\omega}(z_1) - \omega^\star(z_1)|\big] + c_4\sqrt{\tfrac{\log(1/\delta)}{n_P}} \tag{A.23}$$

where $c_4 = c_1 + \|\ell\|_\infty c_2 + c_3$.

Proof of Lemma A.3. For the simplicity of notations we drop the subscript from n_P and denote the sample size simply by n. For $i \le n$ we define $\mathbf{E}_{i:n}$ as the expectation with respect to the random variables $z_i,\dots,z_n$, and for $i > n$ we define $\mathbf{E}_{i:n}\big[F(z_{1:n})\big] = F(z_{1:n})$ and notice that

$$F(z_{1:n}) - \mathbf{E}\big[F(z_{1:n})\big] = \sum_{i=1}^n\Big\{\mathbf{E}_{(i+1):n}\big[F(z_{1:n})\big] - \mathbf{E}_{i:n}\big[F(z_{1:n})\big]\Big\}\,. \tag{A.24}$$

Here,

$$\begin{aligned}
&\mathbf{E}_{(i+1):n}\big[F(z_{1:n})\big] - \mathbf{E}_{i:n}\big[F(z_{1:n})\big]\\
&= \mathbf{E}_{(i+1):n}\Big\{F(z_{1:n}) - \mathbf{E}_{z_i}\big[F(z_{1:n})\big]\Big\}\\
&= \mathbf{E}_{(i+1):n}\Big\{F(z_1,\dots,z_{i-1},z_i,z_{i+1},\dots,z_n) - \mathbf{E}_{z_i'}\big[F(z_1,\dots,z_{i-1},z_i',z_{i+1},\dots,z_n)\big]\Big\}\\
&= \mathbf{E}_{(i+1):n}\mathbf{E}_{z_i'}\Big\{F(z_1,\dots,z_{i-1},z_i,z_{i+1},\dots,z_n) - F(z_1,\dots,z_{i-1},z_i',z_{i+1},\dots,z_n)\Big\}
\end{aligned} \tag{A.25}$$

where, z_i' is an IID copy of z_i. We notice that

$$\begin{aligned}
&F(z_1,\dots,z_{i-1},z_i,z_{i+1},\dots,z_n) - F(z_1,\dots,z_{i-1},z_i',z_{i+1},\dots,z_n)\\
&= \sup\nolimits_{f\in\mathcal{F}}\Big\{\mathbf{E}\big[g_f(z_1)\big] - \tfrac{1}{n}\sum_{i=1}^n g_f(z_i)\Big\}\\
&\qquad - \sup_{f\in\mathcal{F}}\Big\{\mathbf{E}\big[g_f(z_1)\big] - \frac{1}{n}\sum_{i=1}^n g_f(z_i) + \frac{1}{n}g_f(z_i') - \frac{1}{n}g_f(z_i)\Big\}\\
&\le \sup\nolimits_{f\in\mathcal{F}}\Big\{-\tfrac{1}{n}g_f(z_i') + \tfrac{1}{n}g_f(z_i)\Big\}
\end{aligned} \tag{A.26}$$

where the last inequality is obtained by setting $A_f = \mathbf{E}\big[g_f(z_1)\big] - \frac{1}{n}\sum_{i=1}^n g_f(z_i)$ and $B_f = \frac{1}{n}g_f(z_i') - \frac{1}{n}g_f(z_i)$ in the following stream of inequalities

$$\begin{aligned}
\sup_{f\in\mathcal{F}}\{A_f\} - \sup_{f\in\mathcal{F}}\{A_f + B_f\} &= \sup\nolimits_{f\in\mathcal{F}}\{A_f + B_f - B_f\} - \sup\nolimits_{f\in\mathcal{F}}\{A_f + B_f\}\\
&\le \sup\nolimits_{f\in\mathcal{F}}\{A_f + B_f\} + \sup\nolimits_{f\in\mathcal{F}}\{-B_f\} - \sup\nolimits_{f\in\mathcal{F}}\{A_f + B_f\}\\
&= \sup_{f\in\mathcal{F}}\{-B_f\}\,,
\end{aligned}$$

and that

$$\begin{aligned}
\sup\nolimits_{f\in\mathcal{F}}\Big\{-\tfrac{1}{n}g_f(z_i') + \tfrac{1}{n}g_f(z_i)\Big\} &\le \tfrac{1}{n}\Big\{\sup\nolimits_{f\in\mathcal{F}}|g_f(z_i')| + \sup\nolimits_{f\in\mathcal{F}}|g_f(z_i)|\Big\}\\
&= \tfrac{1}{n}\Big\{w^\star(z_i')\sup\nolimits_{f\in\mathcal{F}}|\ell(z_i')| + w^\star(z_i)\sup\nolimits_{f\in\mathcal{F}}|\ell(z_i')|\Big\}\\
&\le \tfrac{\|\ell\|_\infty}{n}\big\{w^\star(z_i') + w^\star(z_i)\big\}\,,
\end{aligned} \tag{A.27}$$

where $\ell(z) = \ell(f(x), y)$. We use inequalities (A.26) and (A.27) in (A.25) and get the following

$$\mathbf{E}_{(i+1):n}\big[F(z_{1:n})\big] - \mathbf{E}_{i:n}\big[F(z_{1:n})\big] \leq \frac{\|\ell\|_\infty}{n}\big\{\mathbf{E}_{z_i'}[w^\star(z_i')] + w^\star(z_i)\big\}\,. \tag{A.28}$$

Now, we use (A.24) and (A.28) to bound the moment generating function of $F(z_{1:n}) - \mathbf{E}[F(z_{1:n})]$ as seen in the following inequalities. For $\lambda > 0$

$$\begin{aligned}
&\mathbf{E}\Big\{\exp\Big(\lambda\big\{F(z_{1:n}) - \mathbf{E}[F(z_{1:n})]\big\}\Big)\Big\} \\
&= \mathbf{E}\Big\{\exp\Big(\lambda\sum_{i=1}^n\big\{\mathbf{E}_{(i+1):n}\big[F(z_{1:n})\big] - \mathbf{E}_{i:n}\big[F(z_{1:n})\big]\big\}\Big)\Big\} \\
&\leq \mathbf{E}\Big\{\exp\Big(\lambda\sum_{i=1}^n\frac{\|\ell\|_\infty}{n}\big\{\mathbf{E}[w^\star(z_i')] + w^\star(z_i)\big\}\Big)\Big\} \\
&\leq \mathbf{E}\Big\{\exp\Big(\lambda\sum_{i=1}^n\frac{\|\ell\|_\infty}{n}\big\{w^\star(z_i') + w^\star(z_i)\big\}\Big)\Big\},\ \text{ since }\ e^{\mathbf{E}\{X\}} \leq \mathbf{E}\big\{e^X\big\} \\
&= \prod_{i=1}^n\mathbf{E}\Big\{\exp\Big(\frac{\lambda\|\ell\|_\infty}{n}w^\star(z_i)\Big)\Big\}\mathbf{E}\Big\{\exp\Big(\frac{\lambda\|\ell\|_\infty}{n}w^\star(z_i')\Big)\Big\} \\
&\leq \prod_{i=1}^n\exp\Big(\frac{2c\lambda^2\|\ell\|_\infty^2}{n^2}\Big) = \exp\Big(\frac{2c\lambda^2\|\ell\|_\infty^2}{n}\Big)\,.
\end{aligned} \tag{A.29}$$

Following the bound on moment generating function in (A.29) we get

$$\begin{aligned}
&\mathbf{P}\big\{F(z_{1:n}) - \mathbf{E}[F(z_{1:n})] > t\big\} \\
&\leq e^{-\lambda t}\mathbf{E}\Big\{\exp\Big(\lambda\big\{F(z_{1:n}) - \mathbf{E}[F(z_{1:n})]\big\}\Big)\Big\} \\
&= \exp\Big(-\lambda t + \frac{2c\lambda^2\|\ell\|_\infty^2}{n}\Big)\,,
\end{aligned}$$

where letting $\lambda = nt/(4c\|\ell\|_\infty^2)$ we obtain

$$\mathbf{P}\big\{F(z_{1:n}) - \mathbf{E}[F(z_{1:n})] > t\big\} \leq \exp\Big(-\frac{nt^2}{8c\|\ell\|_\infty^2}\Big)\,,$$

and letting $t = \|\ell\|_\infty\sqrt{8c\log(1/\delta)/n}$ we establish the lemma with $C = \|\ell\|_\infty\sqrt{8c}$.

□

B EXPERIMENT DETAILS

B.1 DATA DETAILS

B.1.1 WATERBIRDS

Data The training data has 4795 sample points with group-wise sample sizes $\{0 : 3498, 1 : 184, 2 : 56, 3 : 1057\}$. We combine the test and the validation data to create test data which has 8192 sample points, and the group vise sample sizes are $\{0 : 3189, 1 : 3187, 2 : 908, 3 : 908\}$. The images are embedded into 512 dimensional feature vectors using ResNet18 (He et al., 2016) pre-trainined on Imagenet (Deng et al., 2009), which we use as covariates. Recent work suggests that using pre-trained features without additional fine-tuning of the feature extractor is beneficial for out-of-distribution generalization (Kumar et al., 2022; Rosenfeld et al., 2022).

Source and target domains For the source domain we use the original training set of images. We consider five different target domains: (1) the target domain with all the groups $g \in \{0, 1, 2, 3\}$ from test data, (2) with groups $g \in \{0, 3\}$ *i.e.*, landbirds on land backgrounds and waterbirds on water backgrounds, (3) with groups $g \in \{0, 2\}$, *i.e.*, landbirds on land backgrounds and waterbirds on land backgrounds, (4) with groups $g \in \{1, 3\}$, and (5) with groups $g \in \{1, 2\}$. Note that all of the target domains have landbirds and waterbirds.

B.1.2 BREEDS

Data BREEDS (Santurkar et al., 2020) is a subpopulation shift benchmark derived from ImageNet (Deng et al., 2009). It uses the class hierarchy to define groups within classes. For example, in the Entity-30 task considered in this experiment, class fruit is represented by strawberry, pineapple, jackfruit, Granny Smith in the source and buckeye, corn, ear, acorn in the target. Each source and target datasets are split into training and test datasets. In the source domain the training (resp. test) data has 159037 (resp. 6200) sample points, whereas in the target domain the sample sizes are 148791 (resp. 5800) for training (resp. test) data. There are 30 different classes in both source and target domains. The highest (resp. lowest) class proportion in source training data is 4.9% (resp. 1.58%) and in target training data is 4.9% (resp. 1.53%). Here, the images are embedded using SwAV (Caron et al., 2020). The embedding is of dimension 2048, which we consider as covariates for our analysis. As in the WATERBIRDS experiment, we do not fine-tune the embedder.

Source and target domains In our BREEDS case study we mix a small amount of labeled target samples into the source data. We mix π proportion of labeled target samples into the source domain and evaluate the performance of our method for several mixing proportions (π). Below we describe the step by step procedure for creating mixed source and target datasets:

1. In both the source and target domains we combine the training and test datasets.
2. Let m be the sample size of the combined source data. We add $\lfloor m\pi \rfloor$ many labeled target samples into the source data.
3. Resulting source and target datasets are then split to create training (80%) and test (20%) data.

B.2 MODEL DETAILS

B.2.1 IMPLEMENTATION FOR EXTRA

We describe the implementation details for ExTRA weights in Algorithm 1.

The normalization regularizer λ is required to control the value of the normalizer $\hat{N}_t$. It makes sure that the value of $\hat{N}_t$ remains close to 1, as the function $x + x^{-1}, x > 0$ is minimized at $x = 1$. The regularizer is particularly important when the feature distribution between source and target data has very little overlap (happens in BREEDS case study).

Algorithm 1 Exponential Tilt Reweighting Alignment (ExTRA)

Require:
- **Dataset:** labeled source data $\{(X_{P,i}, Y_{P,i})\}_{i=1}^{n_P}$ and unlabeled target data $\{X_{Q,i}\}_{i=1}^{n_Q}$.
- **Hyperparameters:** learning rate $\eta > 0$, batch size $B \in \mathbb{N}$, normalization regularizer $\lambda > 0$.
- **Probabilistic source classifier:** $\widehat{\eta}_P : \mathcal{X} \to \Delta^K$.
- **Initial values**: $\{(\hat{\theta}_{k,0}, \hat{\beta}_{k,0})\}_{k=1}^K$

Initialize $\hat{\theta}_0$ at some value.
repeat $t \geq 0$
  sample minibatchs $(X_{P,1}, Y_{P,1}), \ldots, (X_{P,B}, Y_{P,B}) \sim \widehat{P}$, and $X_{Q,1}, \ldots, X_{Q,B} \sim \widehat{Q}_X$
  Compute loss $\hat{L}_t = \frac{1}{B}\sum_{i=1}^B \log\left\{\sum_{k=1}^K \widehat{\eta}_{P,k}(X_{Q,i})\exp(\hat{\theta}_{k,t}^\top T(X_{Q,i}) + \hat{\beta}_k\right\}$ and normalizer $\hat{N}_t = \frac{1}{B}\sum_{i=1}^B \exp\{\hat{\theta}_{Y_{P,i},t}^\top T(X_{P,i}) + \hat{\beta}_{Y_{P,i},t}\}$
  Objective $\hat{O}_t = -\hat{L}_t + \log(\hat{N}_t) + \lambda\hat{N}_t + \lambda\hat{N}_t^{-1}$
  Update $\hat{\theta}_{k,t+1} \leftarrow \hat{\theta}_{k,t} - \eta\partial_{\theta_k}\hat{O}_t\{(\hat{\theta}_{k,t}, \hat{\beta}_{k,t}), k = 1, \ldots, K\}$ and $\hat{\beta}_{k,t+1} \leftarrow \hat{\beta}_{k,t} - \eta\partial_{\beta_k}\hat{O}_t\{(\hat{\theta}_{k,t}, \hat{\beta}_{k,t}), k = 1, \ldots, K\}$
until converges
Estimated value $\{(\hat{\theta}_k, \hat{\beta}_k)\}_{k=1}^K$
$\hat{\alpha}_k \leftarrow \hat{\beta}_k - \log\hat{N}(\{(\hat{\theta}_k, \hat{\beta}_k)\}_{k=1}^K)$
return parameters $\{(\hat{\theta}_k, \hat{\alpha}_k)\}_{k=1}^K$ and the weight function $\omega(x, y) = \exp(\hat{\theta}_y T(x) + \hat{\alpha}_y)$

B.2.2 WATERBIRDS

Source classifier $\widehat{\eta}_P$ is a logistic regression model fitted on source data using `sklearn.linear_model.LogisticRegression` module with the parameters {solver = 'lbfgs', C = 0.1, tol = 1e-6, max_iter=500} and the rest set at their default values. We also use several calibration techniques for the source classifier (Shrikumar & Kundaje, 2019): (1) no model calibration (none), (2) temperature scaling (TS), (3) bias corrected temperature scaling (BCTS), and (4) vector scaling (VS). For TS, BCTS and VS we use the implementation in Shrikumar & Kundaje (2019).

ExTRA importance weights In each iterations of ExTRA importance weight calculations (Algorithm 1) for WATERBIRDS data we fix an initialization of the parameters and compute the parameters for several values of the hyperparameters learning rate $\eta \in \{5 \times 10^{-4}, 4 \times 10^{-5}\}$, batch size $B = 500$, epochs $E \in \{100, 200, 400\}$ and source model calibrations {none, TS, BCTS, VS}. The details can be found in the supplementary codes. Since there are significant overlap between source and target feature distributions we set $\lambda = 0$. We select the hyperparameter setup that produces the lowest value of the objective $-\hat{L} + \log(\hat{N})$ over the full data $\{(X_{P,i}, Y_{P,i})\}_{i=1}^{n_P}$ and $\{X_{Q,i}\}_{i=1}^{n_Q}$.

Runtime ExTRA algorithm requires solving a simple stochastic optimization problem. In the WATERBIRDS case-study, 100 epochs of the Adam (Kingma & Ba, 2017) optimizer took 11.92 ± 0.48 seconds.

Model selection We considered 120 different models which are logistic regression models of three categories:

1. A vanilla model.
2. A model fitted on weighted data to balance the class proportion on source data.
3. A model fitted on weighted data to balance the group proportions on the source data.

Each of these models are fitted with scikit-learn logistic regression module where we use 2 different regularizers, ℓ_1 and ℓ_2 and 20 different regularization strengths (`numpy.logspace(-4, -1, 20)`) and we use `liblinear` solver to fit the models. Rest of the parameters are set at their default values.

For model selection experiments the source and target data are each split into equal parts to create source and target training and test datasets. The models are fitted using test data on the source domain. We then calculate its (1) ScrVal accuracy using the source training data, (2) ACT-NE accuracy using labeled training data on source and unlabeled training data on target, (3) ExTRA accuracy on training data on source, and finally (4) oracle target accuracy on test data on target. We then summarize the accuracies in (1) oracle *target accuracies* for the models chosen according to the best ScrVal, ACT-NE and ExTRA accuracies on the source domain, and (2) the *rank correlations* between the ScrVal, ACT-NE, and ExTRA accuracies with the corresponding oracle target accuracies.

B.2.3 BREEDS

Source classifier $\widehat{\eta}_P$ The source classifier $\widehat{\eta}_P$ used in BREEDS case study is similar to the one in WATERBIRDS. It uses same model and parameters to obtain a probabilistic classifier $\widehat{\eta}_P$. We use bias corrected temperature scaling (BCTS) (Shrikumar & Kundaje, 2019) for calibrating $\widehat{\eta}_P$.

ExTRA importance weights We set the hyperparameters at some fixed values and obtain our weights for several random initializations of the tilt parameters. The hyperparameter values are: (1) learning rate $\eta = 10^{-4}$, (2) batch size $B = 1500$, (3) number of epochs epochs $= 500$, and (4) regularization strength for normalizer $\lambda = 10^{-6}$. Rest of the setup is same as in WATERBIRDS.

B.3 ADDITIONAL RESULTS

B.3.1 WATERBIRDS

Precision and recall We report the precision and recall for the weights. They are defined as following: within an x proportion of samples with the highest weights (call this set A)

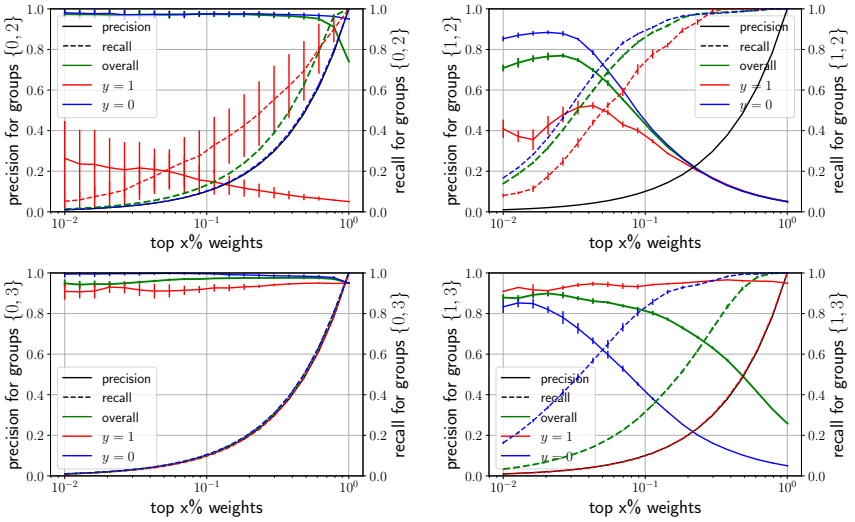

Figure 4: ExTRA precision and recall on WATERBIRDS for different targets. The black solid line refers to a baseline for the recall curve when the weights are completely non-informative of the target domain.

1. precision is the proportion of sample points from the groups comprising the target domain in A, *i.e.*
$$\frac{\#\{\text{sample points in } A \text{ with } g \in \{\text{target groups}\}\}}{|A|},$$
and,
2. recall is the ratio between the number of sample points in A that are from the target groups and the total number of points in source data that are from the target groups, *i.e.*
$$\text{precision} = \frac{\#\{\text{sample points in } A \text{ with } g \in \{\text{target groups}\}\}}{\#\{\text{sample points in source data with } g \in \{\text{target groups}\}\}}.$$

Target domains consisting of a majority and a minority group, *e.g.* $\{0, 2\}$, are noticeably imbalanced in the source data, thus we also report precision and recall conditioned on the class label (*i.e.* treating source and target as consisting of either only landbirds ($y = 0$) or only waterbirds ($y = 1$) in the aforementioned precision and recall definitions).

We report results for four target domains in Figure 4. Overall, the ExTRA weights are informative (precision curves have downward trends and the recall curves are above the non-informative baseline, *i.e.* solid black lines). We note that for class $y = 0$ for target domain $\{0, 2\}$ and for class $y = 1$ for target domain $\{1, 3\}$ we see that the ExTRA weights are almost non-informative (precision curve is almost flat and the recall curve is almost aligned with the baseline). This is due to the group imbalance within a class. In the example of class $y = 0$ with target domain $\{0, 2\}$, the groups with $y = 0$ are $g = 0$ and $g = 1$. Since the sample size for $g = 1$ is very small compared to $g = 0$, most of the samples in $y = 0$ class are from the correct group $g = 0$ when we consider $\{0, 2\}$ as our target domain, and any weights would have precision-recall curves that are close to the non-informative baseline. Similar behavior is observed for the other example.

Upweighted images We visualize images from the WATERBIRDS dataset corresponding to the 16 largest ExTRA weights for the $\{1, 2\}$ target domain consisting of the two minority groups. Among these 16 images, 12 correspond to the correct groups, *i.e.* either landbirds on water or waterbirds on land. We emphasize that in the source domain there are only 5% of images corresponding to groups $\{1, 2\}$ and ExTRA weights upweigh them as desired. The 4 images from the other groups (highlighted with red border) are: (i) 2nd row, 3rd column (waterbird on water); (ii) 3rd row, 2nd column (waterbird on water); (iii) 4th row, 3rd column (waterbird on water); (iv) 4th row, 4th column (landbird on land). Arguably, the background in (i) is easy to confuse with the land background and

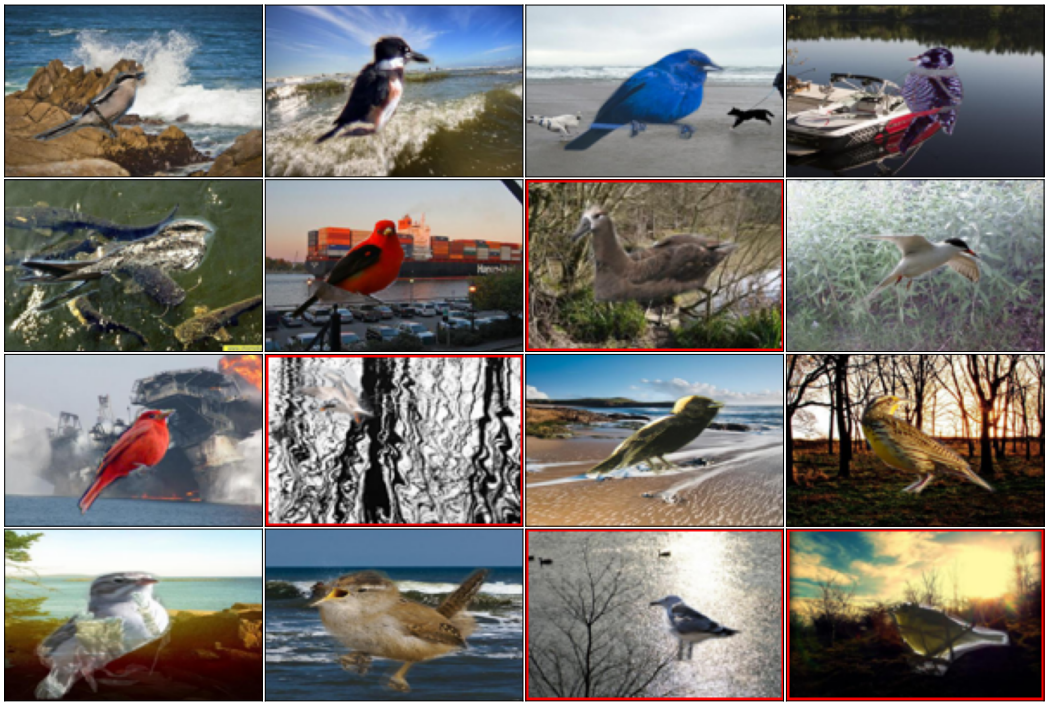

Figure 5: WATERBIRDS images with 16 largest ExTRA weights for $\{1, 2\}$ as the target domain. The four images highlighted with red border are from other groups.

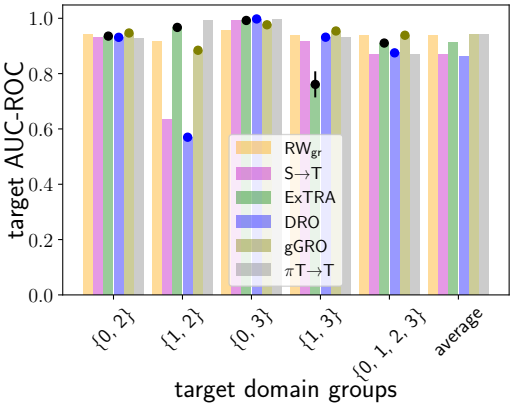

Figure 6: Area under the curve for target receiver operating characteristic curves for WATERBIRDS.

the blue sky in (iv) is easy to confuse with the water background, suggesting that these images might be representative of the target domain of interest despite belonging to different groups.

Area under the Receiver Operating Characteristic curve Here we compare the area under the curve for target receiver operating characteristic curves (we call it target AUC-ROC). The behaviors for target AUC-ROC's are similar to target accuracies (Figure 2).

B.3.2 BREEDS

We present precision and recall curves for the target samples identified within the source samples with larger ExTRA weights (analogous to the corresponding WATERBIRDS experiment) in Figure 7 for varying mixing proportion π. In comparison to WATERBIRDS, we note that both precision and recall are lower, which we think is due to a larger amount of the original source samples representa-

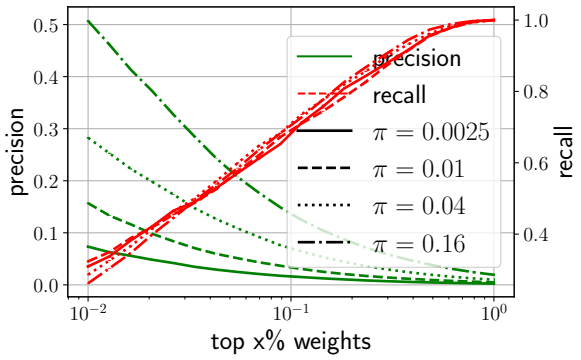

Figure 7: ExTRA precision and recall on BREEDS.

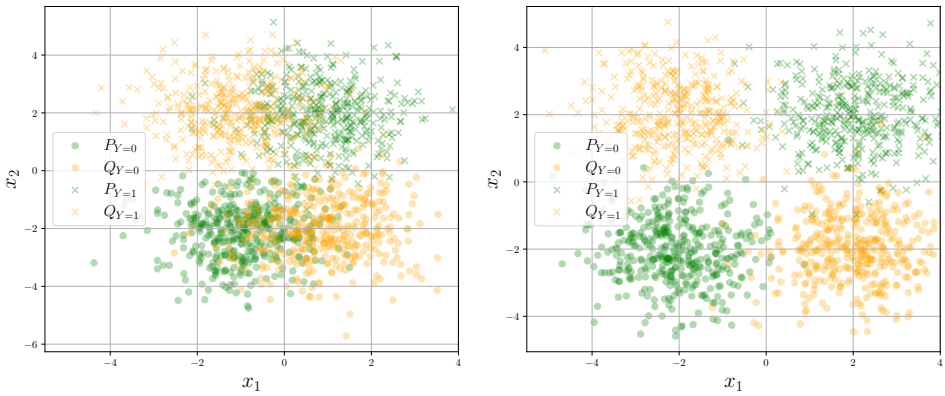

Figure 9: Source and target data with $\delta = 1$ (left) and $\delta = 2$ (right).

tive of the target domain distribution as can be seen from the improved performance of the ExTRA fine-tuned model in Figure 3 even when $\pi = 0$.

C SYNTHETIC EXPERIMENT: NORMAL MIXTURE

We demonstrate utility of the ExTRA weights for reweighing source data in the following synthetic experiment. As seen in Figure 9 both the source and target distributions are associated with binary classification tasks and their class conditionals are normally distributed in $\mathbf{R}^2$. More specifically, the class conditionals in source distribution have means $(-\delta, -2)^\top$ and $(\delta, 2)^\top$ for classes 1 and 0 resp., whereas the means in target distributions are $(\delta, -2)^\top$ and $(-\delta, 2)$. All the class conditionals have variances $\mathbf{I}_2$. This is an instance of concept drift which does not fall under subpopulation shift (considering classes as subpopulations).

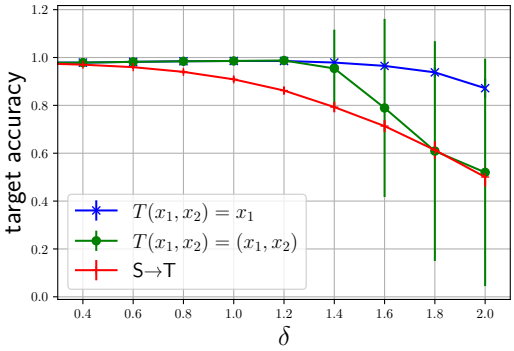

Figure 8: Target accuracies for different models and for different δ values.

The distribution shift in this example satisfies exponential tilt assumption with sufficient statistic $T(x_1, x_2) = x_1$. In our experiment we use two sufficient statistics: (1) $T(x_1, x_2) = x_1$ - an ideal sufficient statistic for this example and (2) $T(x_1, x_2) = (x_1, x_2)$ - a simple choice for a sufficient statistic. Here, we vary δ to control the overlap between the class conditionals in corresponding classes. For a small δ the source and target class conditionals for class 0 (or 1) have overlapping support, satisfying the anchor-set assumption in Definition 4.1.

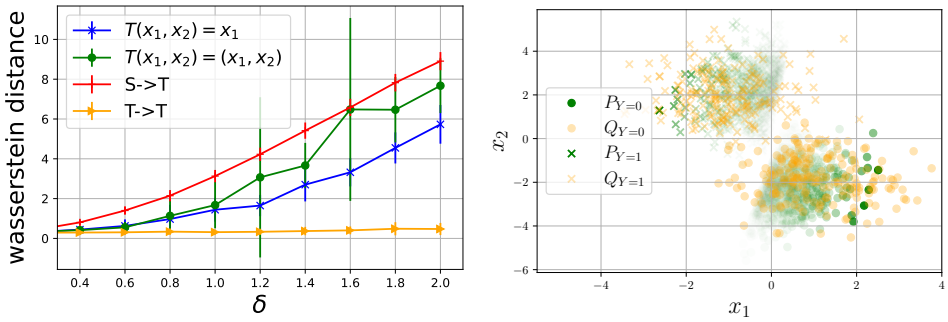

Figure 10: *Left*: Wasserstein distances between source vs target, weighted source vs target and target vs target data. *Right*: Scatter plot for weighted source (highlighted according to the ExTRA weights) and target data.

We compare the target accuracies of three different models trained on: (1) source data (S->T), (2) ExTRA weighted source data using $T(x_1, x_2) = x_1$, and (3) ExTRA weighted source data using $T(x_1, x_2) = (x_1, x_2)$. As we see in Figure 8, the ideal sufficient statistic $T(x_1, x_2) = x_1$ produces better target accuracy than the other two models. We also observe that $T(x_1, x_2) = (x_1, x_2)$ has better target accuracy than S->T for small values of δ. The reason for such observation goes back to the anchor set condition in Definition 4.1. As we see in Figure 9, left plot, a small δ ($\delta = 1$) results in support overlap between source and target class conditionals for class 0 (and 1), and the overlapped region of support works as anchor set. On contrary, large δ's (Figure 9, right plot) has very little overlap, resulting in violation in the anchor set assumption and leading to a non-identifiability in the exponential tilt model for sufficient statistic $T(x_1, x_2) = (x_1, x_2)$. Nevertheless, the ideal sufficient statistic $T(x_1, x_2) = x_1$ continues to work even for large δ.

We also investigate whether reweighed source data with ExTRA weights better approximates the target data compared to the uniformly weighted source data. We first compare the following Wasserstein distances between target and: (1) source (S->T), (2) target (T->T), (3) ExTRA weighted source with $T(x_1, x_2) = (x_1, x_2)$, and (4) ExTRA weighted source with $T(x_1, x_2) = x_1$ for several values of δ. As we see in Figure 10 left plot, ExTRA weighted source data is closer to the target data compared to the uniformly weighted source data. Similar to the target accuracy inspection (Figure 8), the sufficient statistic $T(x_1, x_2) = x_1$ shows better performance in adapting to the target distribution than $T(x_1, x_2) = (x_1, x_2)$. In Figure 10 right panel, we compare the weighted source data for $\delta = 1$ when the ExTRA weights are calculated with sufficient statistic $T(x_1, x_2) = (x_1, x_2)$. Compared to the uniformly weighted source data (see Figure 9, left plot) we see that ExTRA weighted source data is also closer to the target one qualitatively.