REGULARIZED LEARNING FOR DOMAIN ADAPTATION UNDER LABEL SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose Regularized Learning under Label shifts (RLLS), a principled and practical approach to correct for shifts in the label distribution between the source and target domains, and is a special case of domain adaptation. We estimate the importance weights using both labeled source and unlabeled target data and then train a classifier on the weighted source samples. Depending on the number of source and target samples, we regularize the influence of these estimated weights and train a corrected classifier on the source set with the estimated weights. We derive a generalization bound for the loss of the classifier on the target set. This bound is independent of the (ambient) data dimensions and instead depends on the complexity of the function class. To the best of our knowledge, this is the first generalization bound for label-shift problems where the labels in the target domain are unknown. Based on this bound, we design a regularized estimator for importance weights that allows us to adapt to the uncertainty in the estimated weights for the small-sample regime. Experiments show that regularized RLLS significantly improves accuracy in the low (target) sample and large-shift regimes compared to previous methods.

1 INTRODUCTION

In supervised learning, the task to make predictions on a given unlabeled *target* set is usually solved by training a predictive model on a labeled *source* set. This is a sensible approach particularly when the source and target sets are expected have i.i.d. samples from the same distribution. However, this assumption does not hold in many real-world applications. For example, often times predictions for medical diagnostics are based on data from a particular population of patients and machines due to the limited variety of data that is available in the training phase.

To further expand on the example above, consider the following scenario. A classifier is trained to predict whether a patient has contracted a severe disease in country A based on the physical measurements and diagnosis (labeled source data). The disease is potentially even more prevalent in country B, where however no doctors are available to make a diagnosis which can only be reliably made using expert knowledge. The following questions are relevant in this scenario:

- If we send a group of volunteers to country B with the same equipment to record measurements of patients, how can the data in country A and the new unlabeled data from country B be used to update the classifier to give good diagnostic predictions for patients in country B?
- If the patients in country B come in slowly over time and the disease is deadly, can the classifier be updated after only few patients have been measured, so as to maximize the accuracy of the diagnosis for the following patients?

Similar problems arise in many other domains besides health care, but we will use this medical example throughout the paper for ease of presentation.

If we view the data in the source and target set, each consisting of an input $x \in \mathcal{X}$ and a label y in a finite domain $\mathcal{Y}$, as i.i.d. samples drawn from joint distributions $\mathbb{P}$ and $\mathbb{Q}$, we are dealing with the *distributional shift* problem when $\mathbb{P} \neq \mathbb{Q}$. There are various options to model the difference between $\mathbb{P}$ and $\mathbb{Q}$, depending on the use case at hand. Schölkopf et al. (2012) studies two main perspectives: (i)

causal learning where the explanatory input x causes the label y, and (ii) anti-causal setting, where the label y causes the input x.

For the causal setting, one usually makes the *covariate shift* assumption that the marginal distribution of the covariates x shifts from $\mathbb{P}$ to $\mathbb{Q}$, whereas the conditional probability mass function of y given x stays constant, which we denote as $p(y|x) = q(y|x)$ for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. Although this is a valid model in some situations, in many data science applications there is an inherent ground truth y that we want to learn given observations x, which can be "explained" by y. For example, the physical symptoms reported by a patient might be approximately similar among all sick patients. This simplified relation can be modeled e.g. by a directed graphical model $Y \to X$. It is now possible that the percentage of ill patients in the two countries A and B are very different, i.e. $\mathbb{P}(Y = i) \neq \mathbb{Q}(Y = i)$, or $p(y) \neq q(y)$ for short. Similarly, within one country, the number of ill patients may suddenly change because of an irregular event such as a natural catastrophe.

Covariate Shift	Label Shift
$p(x) \neq q(x)$	$p(y) \neq q(y)$
$p(y\|x) = q(y\|x)$	$p(x\|y) = q(x\|y)$

In all of the above cases, there is a shift in the distribution of the label y (sick or healthy), while given a fixed label, the conditional distribution of x given y is constant. By the Bayes rule, *label shift* causes the covariate shift condition to be violated, i.e. $p(y|x) \neq q(y|x)$. Despite being an intuitively natural scenario in some cases, even this simplified model has not yet been sufficiently studied in the literature. This paper aims to fill the gap in terms of both theoretical understanding and practical methods for the label shift setting and thereby move a step closer towards more complicated and realistic distributional shift models.

Lipton et al. (2018) introduced a provably good importance-weight estimator called BBSE, which has access to the labeled source data set and a large batch of unlabeled target data, and demonstrated promising performance on real data. There are a two natural follow-up questions which remain unanswered by their work: What are the generalization guarantees for the final predictor which uses the weighted samples? How do we deal with the uncertainty of the weight estimation when only few samples are available and how can we update the classifier?

In this work, we propose an efficient method to estimate the importance weights and relax the problem-dependent warm-up sample complexity requirement for BBSE (which is unknown and can be arbitrary large). We further improve the estimation error in the finite-sample setting by a factor of $k \log(k)$ over the prior work, where k is the number of classes. We deploy this estimator and propose a label-shift correcting estimator and provide generalization guarantees. Similar to Lipton et al. (2018), we first use a black-box classifier, which can be biased, uncalibrated and inaccurate, to estimate importance weights for the source samples. Depending on the number of source and target samples, we regularize the influence of these estimated weights and train a corrected classifier on the source set with the estimated weights. We then derive a dimension-independent generalization bound for the final Regularized Learning under Label Shift (RLLS) classifier.

We empirically study the performance of RLLS and show weight estimation as well as prediction accuracy for a variety of shifts, including shifting source data and shifting target data with various parameters, different sample sizes and regularization parameters. We achieve an order of magnitude smaller weight estimation error than baseline methods in large shift cases and enjoy 20% higher accuracy or F-1 score in corresponding predictive tasks. We also see at least 10% accuracy improvement using our regularized estimator when faced with small target sample sizes.

2 REGULARIZED LEARNING OF LABEL SHIFTS (RLLS)

Let us formally define the short hand for the marginal probability mass functions of Y with respect to $\mathbb{P}, \mathbb{Q}$ as $p, q : [k] \to [0, 1]$ with $p(i) = \mathbb{P}(Y = i)$, and $q(i) = \mathbb{Q}(Y = i)$ for all $i \in [k]$, representable by vectors in $\mathbb{R}_+^k$ which sum to one. In the label shift setting, we define the importance weight vector $w \in \mathbb{R}^k$ between these two domains as $w(i) = \frac{q(i)}{p(i)}$. We quantify the shift using the exponent of the infinite and second order Renyi divergence as follows

$$d_\infty(q||p) := \sup_i \frac{q(i)}{p(i)} \quad , \quad \textit{and} \quad d(q||p) := \mathbb{E}_{(Y)\sim\mathbb{Q}}\left[w(Y)Y\right] = \sum_i^k q(i)\frac{q(i)}{p(i)}.$$

Given a hypothesis class $\mathcal{H}$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ our aim is to find the hypothesis $h \in \mathcal{H}$ which minimizes

$$\mathcal{L}(h) = \mathbb{E}_{(X,Y)\sim\mathbb{Q}}\left[\ell(Y, h(X))\right] = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[w(Y)\ell(Y, h(X))\right]$$

In the usual finite sample setting however, $\mathcal{L}$ unknown and we are given samples $\{(x_j, y_j)\}_{j=1}^n$ from $\mathbb{P}$ instead. If we are given the vector of importance weights w we could then minimize the empirical loss with importance weighted samples defined as

$$\mathcal{L}_n(h) = \frac{1}{n}\sum_{j=1}^{n} w(y_j)\ell(y_j, h(x_j))$$

where n is the number of available observations drawn from $\mathbb{P}$ used to learn the classifier h. As w is unknown in practice, we have to find the minimizer of the empirical loss with *estimated* importance weights

$$\mathcal{L}_n(h; \widehat{w}) = \frac{1}{n}\sum_{j=1}^{n} \widehat{w}(y_j)\ell(y_j, h(x_j)) \tag{1}$$

where $\widehat{w}$ are estimates of w. Given a set D_p of n_p samples from the source distribution $\mathbb{P}$, we first divide the data set into two where we use $(1-\beta)n_p$ samples in set D_p^{weight} to compute the estimate $\widehat{w}$ and the remaining $n = \beta n_p$ in the set D_p^{class} to find the classifier which minimizes the loss (1), i.e. $\widehat{h}_{\widehat{w}} = \arg\min_{h\in\mathcal{H}} \mathcal{L}_n(h; \widehat{w})$. In the following, we describe how to estimate the weights $\widehat{w}$ and provide guarantees for the resulting estimator $\widehat{h}_{\widehat{w}}$.

Plug-in weight estimation The following simple correlation between the label distributions p, q was noted in Lipton et al. (2018): for a fixed hypothesis, if for all $y \in \mathcal{Y}$ it holds that $q(y) \geq 0 \implies p(y) \geq 0$ $(*)$, we have

$$\begin{aligned} q_h(Y = i) &= \sum_j \mathbb{Q}(h(X) = i|Y = j)q(j) = \sum_j \mathbb{P}(h(X) = i|Y = j)q(j) \\ &= \sum_j \mathbb{P}(h(X) = i, Y = j)\frac{q(j)}{p(j)} = \sum_j \mathbb{P}(h(X) = i, Y = j)w_j \end{aligned}$$

for all $i, j \in \mathcal{Y}$, which can equivalently be written in matrix vector notation (by slight abuse of notation) as

$$q_h = C_h w \tag{2}$$

where C_h is the confusion matrix, that is $[C_h]_{i,j} = \mathbb{P}(h(X) = i, Y = j)$ and q_h is the vector which represents the probability mass function of $h(X)$ under distribution $\mathbb{Q}$. The assumption $(*)$ is a reasonable condition since without any prior knowledge, there is no way to properly reason about a class in the target domain that is not represented in the source domain.

In reality, both q_h and C_h can only be estimated by the corresponding finite sample averages $\widehat{q}_h, \widehat{C}_h$. Lipton et al. (2018) simply compute the inverse of the estimated confusion matrix $\widehat{C}_h$ in order to estimate the importance weight, i.e. $\widehat{w} = \widehat{C}_h^{-1}\widehat{q}_h$. While $C_h^{-1}\widehat{q}_h$ is a statistically efficient estimator, $\widehat{w}$ with estimated $\widehat{C}_h^{-1}$ can be arbitrarily bad especially for small sample sizes and small minimum singular values. Intuitively, when there are very few samples, the weight estimation will have high variance in which case it might be better to avoid importance weighting altogether. Furthermore, even when the sample complexity in Lipton et al. (2018), unknown in practice, is met, the resulting error of this estimator is linear in k which is problematic for large k.

We therefore aim to address these shortcomings by proposing the following two-step procedure to compute importance weights: Given a "decent" black box estimator which we denote by h_0, we

1. calculate the measurement error adjusted $\widehat{\theta}$ as described in Section 2.1 for h_0
2. compute the regularized weight $\widehat{w} = \mathbf{1} + \lambda\widehat{\theta}$ where λ depends on the sample size $(1-\beta)n_p$

By "decent" we refer to a classifier h_0 which yields a full rank confusion matrix C_{h_0} An example for such a classifier h_0 is one that always outputs a fixed class. As it does not capture any characteristics of the the data, there is no hope to gain any statistical information.

2.1 ESTIMATOR CORRECTING FOR FINITE SAMPLE ERRORS

Both the confusion matrix C_{h_0} and the label distribution q_{h_0} on the target for the black box hypothesis h_0 are unknown and we are instead only gave access to finite sample estimates $\widehat{C}_{h_0}, \widehat{q}_{h_0}$. In what follows all empirical and population confusion matrices as well as label distributions are defined with respect to the hypothesis $h = h_0$. For notational simplicity, we thus drop the subscript h_0 in what follows.

In the unweighted case we have $w = \mathbf{1}$ so that define the amount of weight shift as $\theta = w - \mathbf{1}$. The reparameterized linear model (2) with respect to θ then reads

$$b := q - C\mathbf{1} = C\theta$$

with corresponding finite sample quantity $\widehat{b} = \widehat{q} - \widehat{C}\mathbf{1}$. When $\widehat{C}$ is near singular, the estimation of θ becomes unstable. Furthermore, large values in the true shift θ result in large variances. We address this problem by adding a regularizing ℓ_2 penalty term to the usual loss and thus push the amount of shift towards 0, a method that has been proposed in (Pires & Szepesvári, 2012). In particular, we compute

$$\widehat{\theta} = \underset{\widehat{\theta}_\rho : \rho \in \{2^{i+1}\Delta_C \Delta_b, i \in \mathbb{N}\}}{\arg\min} \|\widehat{C}\widehat{\theta}_\rho - \widehat{b}\|_2 + 3\Delta_C \|\widehat{\theta}_\rho\|_2 \tag{3}$$

where $\widehat{\theta}_\rho$ is defined by

$$\widehat{\theta}_\rho = \arg\min_{\theta'} \|\widehat{C}\theta' - \widehat{b}\|_2^2 + \rho\|\theta'\|_2.$$

The proof can be found in Section A.1. Here, Δ_C, Δ_b are parameters which will eventually be high probability upper bounds for $\|\widehat{C} - C\|_2$ and $\|\widehat{b} - b\|_2$ respectively. Let $\sigma_{\min}$ be the minimum singular value of the true confusion matrix C. For the estimate $\widehat{\theta}$, we then have the following upper bound

Lemma 1 *For $\widehat{\theta}$ as defined in equation* (3)*, we have with probability at least $1 - \delta$ that*

$$\|\widehat{\theta} - \theta\|_2 \leq \mathcal{O}\left(\frac{1}{\sigma_{min}}\left(\|\theta\|_2\sqrt{\frac{\log(k/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(1/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(1/\delta)}{n_q}}\right)\right)$$

where $\delta < 0.5$[1]*.*

A couple of remarks are in order at this point. First of all, the upper bound in Lemma 1 includes $\sigma_{\min}(C)$ as a parameter, apart from the norm of the true weight shift θ. Recall that these quantities are all heavily dependent on the black box estimator h_0 that we are given a priori. In particular, if h_0 is an ideal estimator, and the source set is balanced, C is the unit matrix with $\sigma_{\min} = 1/k$. In contrast, when the model h_0 is uncertain, the singular value $\sigma_{\min}$ is close to zero. For least square problems under the Gaussian assumption and with measurement errors in the matrix, it is standard to use regularized total least squares approaches which requires a single value decomposition. Our choice for the alternative estimator in Eq. 3 with norm instead of norm squared regularization is motivated by the label shift setting, where true shifts θ with a large norm may shrink the estimate $\widehat{\theta}$ too much, away from the true θ.

2.2 REGULARIZED ESTIMATOR AND GENERALIZATION BOUND

When the number of samples in the target set is very small or the label shift is mild, the estimated weights might be too uncertain to be applied. We therefore propose a regularized estimator defined as follows

$$\widehat{w} = \mathbf{1} + \lambda\widehat{\theta}. \tag{4}$$

Note that $\widehat{w}$ implicitly depends on λ, β. By rewriting $\widehat{w} = (1-\lambda)\mathbf{1} + \lambda(\mathbf{1} + \widehat{\theta})$, we see that intuitively λ should be closer to 1 the more reason there is to believe that $\mathbf{1} + \widehat{\theta}$ is in fact the true weight.

Let us define the set $\mathcal{G}(\ell, \mathcal{H}) = \{g_h(x, y) = w(y)\ell(h(x), y) : h \in \mathcal{H}\}$ for a loss function ℓ with $\|\ell\|_\infty \leq 1$ and hypothesis class $\mathcal{H}$ and the commonly used Rademacher complexity measure

[1] Throughout the paper, $\mathcal{O}$ hides universal constant factors. Furthermore, we use $\mathcal{O}(\cdot + \cdot)$ for short to denote $\mathcal{O}(\cdot) + \mathcal{O}(\cdot)$.

Algorithm 1 RLLS

1: Input: source set D_p, D_q, $\theta_{\max}$, black box estimator h_0, model class $\mathcal{H}$
2: Determine optimal split ratio $\beta^\star$ and regularizer $\lambda^\star$ by minimizing the RHS of Eq. (6)
3: Randomly partition source set D_p into $D_p^{\text{class}}, D_p^{\text{weight}}$ such that $|D_p^{\text{class}}| = \beta^\star n_p =: n$
4: Compute $\widehat{\theta}$ using Eq. (3) and $\widehat{w} := 1 + \lambda^\star \widehat{\theta}$
5: Minimize the importance weighted empirical loss to obtain the weighted estimator

$$\widehat{h}_{\widehat{w}} = \arg\min_{h \in \mathcal{H}} \mathcal{L}_n(h; \widehat{w}), \qquad \textit{where} \qquad \mathcal{L}_n(h; \widehat{w}) = \frac{1}{n} \sum_{(x,y) \in D_p^{\text{class}}} \widehat{w}(y) \ell(y, h(x))$$

6: Deploy $\widehat{h}_{\widehat{w}}$ if the risk is acceptable

$\mathcal{R}_n(\mathcal{G}) := \mathbb{E}_{X(i),Y(i)\sim\mathbb{P}:i\in[n]} \left[\mathbb{E}_{\sigma_i:i\in[n]} \frac{1}{n} \left[\sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i w(y_i) \ell(X_i, h(Y_i))\right]\right]$ (see e.g. Bartlett & Mendelson (2002)). We can now state a generalization bound for the classifier $\widehat{h}_{\widehat{w}}$ in a general hypothesis class $\mathcal{H}$, which is trained on source data with the estimated weights defined in equation (4).

Theorem 1 (Generalization bound for $\widehat{h}_{\widehat{w}}$) *Given n_p samples from the source data set and n_q samples from the target set, a hypothesis class $\mathcal{H}$ and loss function ℓ, the following upper bound*

$$|\mathcal{L}(\widehat{h}_{\widehat{w}}) - \mathcal{L}(h^*)| \leq 2\mathcal{R}_n(\mathcal{G}) + d_\infty(q||p)\sqrt{\frac{\log(1/\delta)}{\beta n_p}} + (1-\lambda)\,\|\theta\|_2 + \frac{\lambda}{\sigma_{\min}} \mathcal{O}\left(\|\theta\|_2 \sqrt{\frac{\log(k/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(1/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(1/\delta)}{n_q}}\right) \tag{5}$$

holds with probability at least $1-\delta$ where $\delta \leq 0.5$.

The proof can be found in Section A.2 as well a tighter bound of $\mathcal{O}(\frac{d_\infty(q||p)}{n_q} + \sqrt{\frac{d(q||p)}{n_q}})$ in terms of dependence on the quantities $d(q||p)$ for finite hypothesis classes in Section A.3. The size of $\mathcal{R}_n(\mathcal{G})$ is determined by the structure of the function class $\mathcal{H}$ and the loss ℓ. For example for the $0/1$ loss, the VC dimension of $\mathcal{H}$ can be deployed to upper bound the Rademacher complexity.

In order to use Theorem 1 in practice and exploit the possibility of choosing parameters λ, β depending on the sample sizes, we need to assume some prior knowledge about the shift θ and $\sigma_{\min}$ of the possible shifts we might encounter. In particular, we assume that it is known that the label shift is not too big, i.e. $\|\theta\|_2 \leq \theta_{\max}$ (on the other hand it is unreasonable to have knowledge about the shift estimator, i.e. $\sigma_{\min}$) and we consider the following upper bound instead for practical implementation

$$|\mathcal{L}(\widehat{h}_{\widehat{w}}) - \mathcal{L}(h^*)| \leq 2\mathcal{R}_n(\mathcal{G}) + d_\infty(q||p)\sqrt{\frac{\log(2/\delta)}{\beta n_p}} + (1-\lambda)\,\theta_{\max} + \frac{\lambda}{\sigma_{\min}} \mathcal{O}\left(\theta_{\max} \sqrt{\frac{\log(k/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(3/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(3/\delta)}{n_q}}\right). \tag{6}$$

The bound in equation (6) suggests to use algorithm 1, in particular when only very few target samples have been observed before an update of the classifier is necessary. The bound (6) gives us guidance for a good choice of λ as a function of the number of samples in practice. Whenever $n_q \geq \frac{1}{\theta_{\max}^2 (\sigma_{\min} - \frac{1}{\sqrt{n_p}})^2}$ (hereby neglecting the log factors), the best choice is $\lambda = 1$ and 0 else. In Figure 2.2 we show curves of fixed $\sigma_{\min}$ in parameter space, under which λ should be chosen as 0 and above which it should be 1. When the confusion matrix has small singular values, we see that the estimated weights should only be trusted for rather high n_q and high believed shifts $\theta_{\max}$.

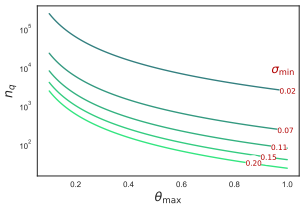


Figure 1: Given a $\sigma_{\min}$ and $\theta_{\max}$, λ switches from 0 to 1 at a particular n_q. n_p is kept fixed.

However, since $\theta_{\max}$ is just an upper bound on the true amount of shift $\|\theta\|_2$, when

$$\frac{1}{\theta_{\max}^2(\sigma_{\min} - \frac{1}{\sqrt{n_q}})^2} \leq n_q \leq \frac{1}{\|\theta\|_2(\sigma_{\min} - \frac{1}{\sqrt{n_q}})^2}$$

λ should in fact ideally be 0. Thus for n_q that are a little bit below the threshold (depending on the certainty of the belief on $\theta_{\max}$), it could be sensible to use an intermediate value $\lambda \in (0, 1)$ which we explore in the experimental section 3.3.

3 EXPERIMENTS

In this section we illustrate our theoretical bounds by running our RLLS estimation procedure on a variety of artificially generated shifts on the MNIST (LeCun & Cortes, 2010) and CIFAR10 (Krizhevsky & Hinton, 2009) datasets. In order to find $\widehat{\omega} = \mathbf{1} + \widehat{\theta}$ in Eq. (3), performing a grid search over ρ on the entire set of natural numbers is not feasible. Instead, we call a built-in solver to directly solve the low dimensional problem $\min_\theta \|\widehat{C}\theta - \widehat{b}\|_2^2 + \rho 3\Delta_C \|\theta\|_2$ where the choice of $\rho = 0.01$ is observed to yield a low loss in the first line of Eq. (3) on various levels of label shift.

We follow three different kinds of sampling strategies to generate a label shift between the source and target set: *Tweak-one Shift*, *Dirichlet shift*, and *Minority-class shift*. For the *Tweak-one Shift*, we randomly choose one class, e.g. i and set $p(y = i) = \rho$ while all other classes are distributed evenly. For the *Dirichlet shift*, we draw a probability vector p from the Dirichlet distribution with concentration parameter set to α for all classes, before including sample points which correspond to the multinomial label variable according to p. In *Minority-Class Shift*, we set a certain number of classes to have probability $p = 0.001$, while the distribution over the rest of the classes is uniform. One of these sampling strategies is used to sample either the source or the target set, while the other set is sampled with a uniform distribution.

After artificially shifting the label distribution in one of the source and target sets, we then follow algorithm 1, where we choose the black box predictor h_0 to be a two-layer fully connected neural network trained on the (unweighted) source dataset. We then evaluate the mean square estimation error (MSE) $\mathbb{E}\|\widehat{w} - w\|^2$, predictive accuracy on the target set as well as F-1 score, $2(\text{precision} * \text{recall})/(\text{precision} + \text{recall})$, and compare them with the performance of the black box shift estimation method (BBSE) in Lipton et al. (2018). Here precision is the the proportion of true predicted labels in all the predicted labels, while recall is the proportion of true predicted labels in all the true labels. Theoretically, KMM (Zhang et al., 2013) is also a possible baseline to compare with. However, as mentioned in Lipton et al. (2018), it is not scalable to large sample sizes for weight estimation.

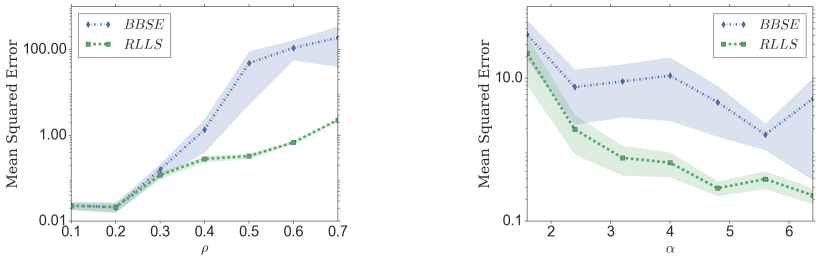
3.1 WEIGHT ESTIMATION COMPARISON

In this set of experiments on the CIFAR10 dataset, we illustrate our weight estimation performance using a simple two-layer neural network as the black box estimator h_0 by comparing it with BBSE from Lipton et al. (2018).

The precise procedure to artificially create a particular shift is as follows: we sample target data uniformly and sample data points from the source set according to a probability vector $p(y)$ that corresponds to different shift parameters ρ in tweak-one Shift and α in Dirichlet Shift for 20 times. Hereby, we fix the number of source and target data points to 10000 for the Dirichlet shift and tweak-one shift for $\rho \leq 0.5$. For bigger shifts, i.e. $\rho \geq 0.5$, we keep the maximum possible number of samples. In Figure 2, we compare the proposed RLLS weight estimation to the BBSE from (Lipton et al., 2018). When the shift is small (ρ is small), the two methods perform similarly, while RLLS outperforms BBSE dramatically when the shift is large.

3.2 PREDICTIVE TASKS WITH LARGE SAMPLE SIZE

In this section, we compare the predictive performances in terms of accuracy and F-1 score between a classifier trained on unweighted source data and the classifiers trained on weighted loss obtained by using the RLLS and BBSE procedure (which is called BBSC).

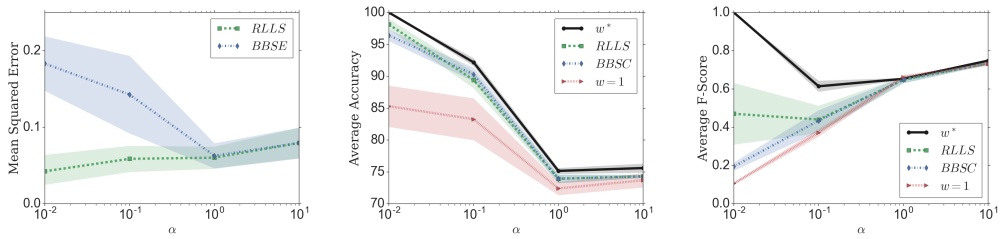


(a) (b)

Figure 2: Comparing MSE of estimated weights using BBSE and RLLS on CIFAR10 with (a) tweak-one shift on source and uniform target, and (b) Dirichlet shift on source and uniform target

When n_q is large, λ is set to one and thus the intermediate weight estimate $\widehat{\theta}$ is "fully" used, i.e. corresponds to $\widehat{w} = \mathbf{1} + \widehat{\theta}$. In this case, better weight estimation should directly lead to better performance of the estimator. One subtlety arises when we compare shifts on the target vs. shifts on the source domain. When the target is seriously biased for smaller α, the predictive task becomes rather easy, such that even the majority class vote would give high accuracy, though it would have zero accuracy on all but one class. Therefore, metrics that account for class imbalance and other types of errors in confusion matrix, like the F-1 score, are in order to allow for a more comprehensive performance evaluation for the prediction.

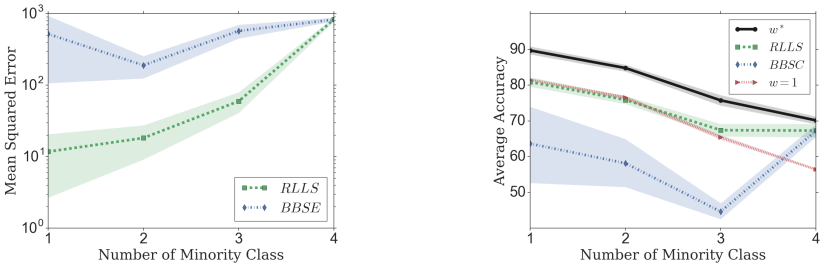
Figure 3 depicts the MSE of the weight estimation (a), the corresponding performance comparison on accuracy (b) and F-1 score (c) when the target data on CIFAR10 is shifted using the Dirichlet Shift. We use the Dirichlet parameters $\alpha = [0.01, 0.1, 1, 10]$ to bias the target set and run 20 different experiments for each parameter. We use the ResNet-18 (He et al., 2016) deep neural net as our hypotheis class $\mathcal{H}$. We can see that in the large shift case when $\alpha = 0.01$, the F-1 score for BBSC and unweighted classifier can be very bad while the accuracy is still high. This is due to classifiers making wrong predictions on minority classes in the target data.



(a) (b) (c)

Figure 3: (a) Mean squared error in estimated weights, (b) accuracy and (c) F-1 score on CIFAR10 for uniform source and Dirichlet shifted target.

A harder problem which does not allow "cheating" is when the shift is in fact on the source set and the target set has uniform label class distribution. Figure 4 shows results using Minority-Class shifted source data and uniform target data from MNIST. For this set of experiments, we first randomly separate the original dataset into source and target before randomly sampling 5000 data points from the target set. The number of minority class ranges from 1 to 4 to bias the source set and run 20 different experiments for each parameter. To better see the difference between methods, we create relatively large shift with one or more class of data being minority. We also use a fixed black-box classifier that is trained on biased source data. We use two-layer fully connected neural networks for both weight estimation and predictive tasks. We can see that the MSE in weight estimation is relatively large. As the error goes up, the performance for all methods goes down, except for BBSC approaching RLLS due to similar weight estimation in the extreme case. Moreover, RLLS always outperforms BBSC whenever the weight estimation error is lower. Notice that generalizing from an imbalanced training set is a common data science task where reweighting the samples in the source set is also standard practice (and it is a priori assumed that the target set is uniform).



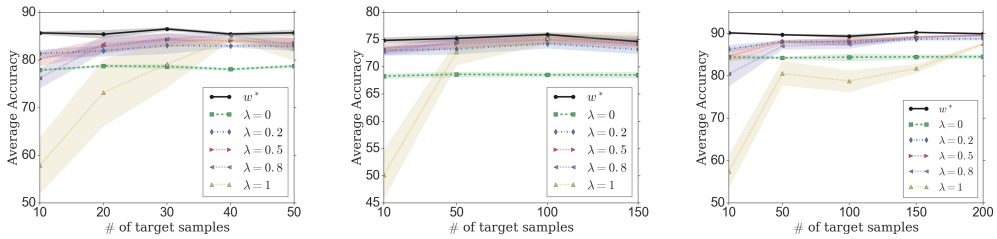
(a) (b)

Figure 4: (a) Mean squared error in estimated weights and (b) accuracy on MNIST for minority-class shifted source and uniform target.

3.3 Label Shift Predictive Tasks in the low sample regime

As discussed in Section 2.2, λ in equation (4) may be chosen according to the believed upper and lower bounds on $\theta_{\max}, \sigma_{\min}$. In the following, we present simulation results of our procedure in the low sample regime in Figure 5 where we choose different values of λ and vary the number of target samples n_q. Here we fix the sample size in source as 10000 and investigate a Dirichlet Shift with fixed $\alpha = 0.5$ on the source data. Figure 5 suggests that our method with $\lambda \in (0, 1)$ outperforms the fully weighted case, i.e. $\lambda = 1$, for small sample sizes. This shows that we are able to adjust to uncertainty on importance weights and generalize well for a varying range of target sample sizes.

We also show a comparison of using different black-box predictors h_0 for weight estimation and observe that fully weighted methods with $\lambda = 1$ achieve the best performance faster with a better trained black-box classifier (a), while it takes longer for it to improve with corrupted one (c). Here we train the black-box classifier with full training data (a), tweak-one with $\rho = 0.2$ (b) and tweak-one with $\rho = 0.6$ (c) to obtain different predictors. This reflects the relation between eigenvalue of confusion matrix $\sigma_{\min}$ and target sample size n_q in Theorem 1. In other words, we need more samples from the target data to compensate a bad predictor in weight estimation. So the generalization error decreases faster with an increasing number of samples for good predictors.



(a) (b) (c)

Figure 5: Performance on MNIST for Dirichlet shifted source and uniform target with various target sample size and λ using (a) better predictor, (b) neutral predictor and (c) corrupted predictor.

4 Related Work

The task of being robust against distributional shifts has been widely studied in the literature. One line of work focuses on distributional robustness (see e.g. Esfahani & Kuhn; Namkoong & Duchi (2016)) without explicitly modeling the shift by making assumptions on the distribution. However, these methods achieve robustness only against tiny changes in the distribution and are often impractical for large-scale high dimensional data.

Another line of work has addressed the causal setting where the conditional distribution of Y given X is assumed to be constant. This assumption is sensible when there is reason to believe that there is a true optimal mapping from X to Y which does not change if the distribution of X changes. In our example, if a given set of symptoms correctly predicts a disease for the "right reasons", the

predictive function should be the same, no matter which country you are moving to. Among the various methods to correct for covariate shift, the majority uses the concept of importance weights $q(x)/p(x)$ (Zadrozny, 2004; Cortes et al., 2010; Cortes & Mohri, 2014; Shimodaira, 2000), which are unknown but can be estimated for example via kernel embeddings (Huang et al., 2007; Gretton et al., 2009; 2012; Zhang et al., 2013; Zaremba et al., 2013) or by learning a binary discriminative classifier between source and target (Lopez-Paz & Oquab, 2016; Liu et al., 2017). Common issues shared by these methods is that they may result in a massive computational burden for large scale problems. Moreover, inaccurate estimates make classifiers suffer from high variance.

Despite the wide applicability of label shift, approaches with global guarantees in high dimensional data regimes remain under-explored. The correction of label shift mainly requires to estimate the importance weights $q(y)/p(y)$ over the labels which typically live in a very low-dimensional space. Bayesian approaches are studied when a prior over the marginal label distribution is assumed (Storkey, 2009; Chan & Ng, 2005). However, these methods often require explicitly computing the posterior distribution of y and suffer from the curse of dimensionality. Recently, for the label shift setting, Lipton et al. (2018) proposed Black Box Shift Estimation (BBSE), where a black box classifier is used to estimate the importance weights $q(y)/p(y)$ which are then used to train a classifier on the source data. This approach is related to Buck et al. (1966); Forman (2008); Saerens et al. (2002) in the low dimensional setting but lacks guarantees for the excess risk.

Generalization theory, which is usually used to bound the excess loss, has historically been developed for the case when $\mathbb{P} = \mathbb{Q}$ (see e.g. (Vapnik, 1999; Bartlett & Mendelson, 2002; Kakade et al., 2009; Wainwright, 2019)). Ben-David et al. (2010) provides theoretical analysis and generalization guarantee for domain adaption when the H-divergecne between join distributions is considered, while Crammer et al. (2008) shows the learning from multiple sources. For the covariate shift setting, Cortes et al. (2010) provides a generalizing bound when $q(x)/p(x)$ is known which however does not apply in practice.

5 Discussion

In this work, we establish the first generalization guarantee for the label shift setting where no prior knowledge of $q(y)/p(y)$ is required. Although RLLS is inspired by BBSE, it carries a more robust importance weight estimator as well as generalization guarantees for both the large and small sample regime. RLLS is also equipped with a sample-size-dependent regularization technique and further improves the classifier in both regimes.

Note that we consider this a necessary first step in the direction of solving shifts of this type, although the label shift assumption might not hold in the real world. In particular, x in practice cannot be solely explained by the wanted label y, but also depends on attributes z which might not be observable. For example, the symptoms might not only depend on the disease but also the country and living conditions, in the example of disease prediction. In such a case, the label shift assumption only holds in a slightly modified sense, i.e. that $\mathbb{P}(X|Y = y, Z = z) = \mathbb{Q}(X|Y = y, Z = z)$. If the attributes Z are observed, then our framework can readily be used to perform importance weights. In the future work, we plan to study the setting where the label shift assumption is slightly violated.

Furthermore, it is not clear whether the final predictor is in fact "better" or more robust to shifts just because it achieves a better target accuracy than a vanilla unweighted estimator. In fact, there is a reason to believe that under certain shift scenarios, the predictor might learn to use spurious correlations to boost accuracy. Finding a procedure which can both learn a robust model and achieve high accuracies on new target sets remains to be an ongoing challenge. Moreover, the current choice of regularization depends on the number of samples rather than date driven regularization is desired.

An interesting direction toward active learning for the same disease-symptoms problem is when in the country B we also have a practitioner whom we can ask to diagnose some of the patients. Now the question is which patient we would diagnose? Also, consider the case of high risk, we might be able to send some of the patients to the country A for further medical diagnosis or treatment, up to some varying cost. We plan to extend the current framework to active learning setting where we actively query the label of certain x's (Beygelzimer et al., 2009) as well as the cost-sensitive setting where we also consider the cost of querying labels (Krishnamurthy et al., 2017).

REFERENCES

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pp. 33–1, 2012.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. *arXiv preprint arXiv:1602.07764*, 2016.

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.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 49–56. ACM, 2009.

AA Buck, JJ Gart, et al. Comparison of a screening test and a reference test in epidemiologic studies. ii. a probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology*, 83(3):593–602, 1966.

Yee Seng Chan and Hwee Tou Ng. Word sense disambiguation with distribution estimation. In *IJCAI*, volume 5, pp. 1010–5, 2005.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pp. 442–450, 2010.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, pp. 1–52.

George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.

Arthur Gretton, Alexander J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pp. 601–608, 2007.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pp. 793–800, 2009.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daume III, and John Langford. Active learning for cost-sensitive classification. *arXiv preprint arXiv:1703.01014*, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916*, 2018.

Song Liu, Akiko Takeda, Taiji Suzuki, and Kenji Fukumizu. Trimmed density ratio estimation. In *Advances in Neural Information Processing Systems*, pp. 4518–4528, 2017.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, pp. 2208–2216, 2016.

Bernardo Avila Pires and Csaba Szepesvári. Statistical linear estimation with penalized estimators: an application to reinforcement learning. *arXiv preprint arXiv:1206.6444*, 2012.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pp. 3–28, 2009.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 114. ACM, 2004.

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in neural information processing systems*, pp. 755–763, 2013.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827, 2013.

A PROOFS

A.1 PROOF OF LEMMA 1

From Thm. 3.7 in (Pires & Szepesvári, 2012) we know that for $\widetilde{w}$ as defined in equation (3), if with probability at least $1-\delta$, $\|\widehat{C}-C\|_2 \leq \Delta_C$ and $\|\widehat{b}-b\|_2 \leq \Delta_b$ hold simultaneously, then

$$\Upsilon(\widehat{\theta}) \leq \inf_{\theta' \in \mathcal{R}^k} \{\Upsilon(\theta') + 3\Delta_C \|\theta'\|_\alpha\} + 3\Delta_b. \tag{7}$$

where we use the shorthand $\Upsilon(\theta) = \|C\theta - b\|_2$.

We can get an upper bound on the right hand side of (7) is the infimum by simply choosing a feasible $\theta' = \theta$. We then have $\|C\theta - b\|_2 = 0$ and hence

$$\inf_{\theta'} \{\Upsilon(\theta') + 3\Delta_C \|\theta'\|_2\} \leq 3\Delta_C \|\theta\|_2$$

as a consequence,

$$\Upsilon(\widehat{\theta}) = \|C\widehat{\theta} - b\|_2 = \|C\left(\widehat{\theta} - \theta\right)\|_2 \leq 3\Delta_C \|\theta\|_2 + 3\Delta_b$$

Since $\|C\left(\widehat{\theta} - \theta\right)\|_2 \geq \sigma_{\min}(C)\|\widehat{\theta} - \theta\|_2$ by definition of the minimum singular value, we thus have

$$\|\widehat{\theta} - \theta\|_2 \leq \frac{1}{\sigma_{min}} \left(3\Delta_C \|\theta\|_2 + 3\Delta_b\right)$$

Let us first notice that

$$b_h = q_h - C\mathbf{1} = q_h - p_h$$

The mathematical definition of the finite sample estimates $\widehat{C}_h, \widehat{b}_h$ (in matrix and vector representation) with respect to some hypothesis h are as follows

$$[\widehat{C}_h]_{ij} = \frac{1}{(1-\beta)n_p} \sum_{(x,y)\in D_p} \mathbb{I}_{h(x)=i, y=j}$$

$$\widehat{[b_h]}(i) = \frac{1}{m} \sum_{(x,y)\in D_q} \mathbb{I}_{h(x)=i} - \frac{1}{(1-\beta)n_p} \sum_{(x,y)\in D_p^{\text{weight}}} \mathbb{I}_{h(x)=i}$$

where $m = |D_q|$ and $\mathbb{I}$ is the indicator function. C_h, b_h can equivalently be expressed with the population over $\mathbb{P}$ for C_h and over $\mathbb{Q}$ for b_h respectively. We now use the following concentration Lemmas to bound the estimation errors of $\widehat{C}, \widehat{b}$ where we drop the subscript h for ease of notation.

Lemma 2 (Concentration of measurement matrix $\widehat{C}$) *For finite sample estimate $\widehat{C}$ we have*

$$\|\widehat{C} - C\|_2 \leq \frac{2\log(2k/\delta)}{3(1-\beta)n_p} + \sqrt{\frac{2\log(2k/\delta)}{(1-\beta)n_p}}$$

with probability at least $(1-\delta)$.

Lemma 3 (Concentration of label measurements) *For the finite sample estimate $\widehat{b}$ with respect to any hypothesis h it holds that*

$$\|\widehat{b} - b\|_2 \leq 2\sqrt{\log(1/\delta)} \left(\frac{1}{\sqrt{n_q}} + \frac{1}{\sqrt{(1-\beta)n_p}}\right)$$

with probability at least $1-\delta$ for $\delta < 0.5$.

By Lemma. 2 for concentration of C and Lemma. 3 for concentration of b we now have with probability at least $1-\delta$

$$\|\widehat{\theta} - \theta\|_2 \leq \frac{1}{\sigma_{min}} \Bigg(\frac{2\|\theta\|_2 \log(2k/\delta)}{(1-\beta)n_p} + \|\theta\|_2 \sqrt{\frac{18\log(4k/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{36\log(2/\delta)}{n_q}} + \sqrt{\frac{36\log(2/\delta)}{(1-\beta)n_p}} \Bigg).$$

which, considering that $O(\frac{1}{\sqrt{n}})$ dominates $O(\frac{1}{n})$, yields the statement of the Lemma 1.

A.1.1 PROOF OF LEMMA 2

We prove this lemma using the theorem 1.4[Matrix Bernstein] and Dilations technique from Tropp (2012).

We can rewrite $C_h = \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[e_{h(x)}e_y^\top\right]$ where $e(i)$ is the one-hot-encoding of index i.

Consider a finite sequence $\{\Psi(i)\}$ of independent random matrices with dimension k. By dilations, lets construct another sequence of self-adjoint random matrices of $\{\widetilde{\Psi}(i)\}$ of dimension $2k$, such that for all i

$$\widetilde{\Psi}(i) = \begin{bmatrix} \mathbf{0} & \Psi(i) \\ \Psi(i)^\top & \mathbf{0} \end{bmatrix}$$

therefore,

$$\widetilde{\Psi}(i)^2 = \begin{bmatrix} \Psi(i)\Psi(i)^\top & \mathbf{0} \\ \mathbf{0} & \Psi(i)^\top\Psi(i) \end{bmatrix} \tag{8}$$

which results in $\|\widetilde{\Psi}(i)\|_2 = \|\Psi(i)\|_2$. The dilation technique translates the initial sequence of random matrices to the sequence of random self-adjoint matrices where we can apply the Matrix Bernstein theorem which states that, for a finite sequence of i.i.d. self-adjoint matrices $\widetilde{\Psi}(i)$, such that, almost surely $\forall i,\ \mathbb{E}\left[\widetilde{\Psi}(i)\right] = 0$ and $\|\widetilde{\Psi}(i)\| \leq R$, then for all $t \geq 0$,

$$\|\frac{1}{t}\sum_{i=1}^{t}\widetilde{\Psi}(i)\| \leq \frac{R\log(2k/\delta)}{3t} + \sqrt{\frac{2\varrho^2\log(2k/\delta)}{t}}$$

with probability at least $1-\delta$ where $\varrho^2 := \|\mathbb{E}\left[\widetilde{\Psi}^2(i)\right]\|_2,\ \forall i$ which is also $\varrho^2 = \|\mathbb{E}\left[\Psi^2(i)\right]\|_2,\ \forall i$ due to Eq. 8. Therefore, thanks to dilation trick,

$$\|\frac{1}{t}\sum_{i=1}^{t}\Psi(i)\| \leq \frac{R\log(2k/\delta)}{3t} + \sqrt{\frac{2\varrho^2\log(2k/\delta)}{t}}$$

with probability at least $1-\delta$.

Now, by plugging in $\Psi(i) = e_{h(x(i))}e_{y(i)}^\top - C$, we have $\mathbb{E}\left[\widetilde{\Psi}(i)\right] = 0$. Together with $\|\widetilde{\Psi}(i)\| \leq 2$ as well as $\varrho^2 = \|\mathbb{E}\left[\Psi^2(i)\right]\|_2 = 1$ and setting $t = n$, we have

$$\|\widehat{C} - C\|_2 \leq \frac{2\log(2k/\delta)}{3(1-\beta)n} + \sqrt{\frac{2\log(2k/\delta)}{(1-\beta)n}}$$

A.1.2 PROOF OF LEMMA 3

The proof of this lemma is mainly based on a special case of and appreared at proposition 6 in Azizzadenesheli et al. (2016), Lemma F.1 in Anandkumar et al. (2012) and Proposition 19 of Hsu et al. (2012).

Analogous to the previous section we can rewrite $b_h = \mathbb{E}_{(x,y)\in\mathbb{Q}}[e_{h(x)}] - \mathbb{E}_{(x,y)\in\mathbb{P}}[e_{h(x)}]$ where $e(i)$ is the one-hot-encoding of index i. Note that (dropping the subscript h) we have

$$\|\widehat{b}_h - b_h\|_2 \leq \|\widehat{q}_h - q_h\|_2 + \|\widehat{p}_h - p_h\|_2$$

We now bound both estimates of probability vectors separately.

Consider a fixed multinomial distribution characterized with probability vector of $\overline{\varsigma} \in \Delta_{k-1}$ where Δ_{k-1} is a $k-1$ dimensional simplex. Further, consider t realization of this multinomial distribution $\{\varsigma(i)\}_{i=1}^t$ where $\varsigma(i)$ is the one-hot-encoding of the i'th sample. Consider the empirical estimate mean of this distribution through empirical average of the samples; $\widehat{\varsigma} = \frac{1}{t}\sum(i)^t\varsigma(i)$, then

$$\|\widehat{\varsigma} - \overline{\varsigma}\| \leq \frac{1}{\sqrt{t}} + \sqrt{\frac{\log(1/\delta)}{t}}$$

with probability at least $1-\delta$.

By plugging in $\overline{\varsigma} = q_h, \widehat{\varsigma} = \widehat{q}_h$ with $t = n_q$ and finally $\{\varsigma(i)\}_{i=1}^{n_q} = \{e_{h(x(i))}\}(i)^{n_q}$ and equivalently for p_h we obtain for $\delta < 0.5$ that

$$\|\widehat{b}_h - b_h\|_2 \leq 2\sqrt{\log(1/\delta)}\left(\frac{1}{\sqrt{(1-\beta)n_p}} + \frac{1}{\sqrt{n_q}}\right)$$

we get the statement in Lemma. 3

A.2 Proof of Theorem 1

We want to ultimately bound $|\mathcal{L}(\widehat{h}_{\widehat{w}}) - \mathcal{L}(h^\star)|$. By addition and subtraction we have

$$\begin{aligned}\mathcal{L}(\widehat{h}_{\widehat{w}}) - \mathcal{L}(h) = &\underbrace{\mathcal{L}(\widehat{h}_{\widehat{w}}) - \mathcal{L}_n(\widehat{h}_{\widehat{w}})}_{(b)} + \underbrace{\mathcal{L}_n(\widehat{h}_{\widehat{w}}) - \mathcal{L}_n(\widehat{h}_{\widehat{w}}; \widehat{w})}_{(a)} \\ &+ \underbrace{\mathcal{L}_n(\widehat{h}_{\widehat{w}}; \widehat{w}) - \mathcal{L}_n(h^\star; \widehat{w})}_{\leq 0} + \underbrace{\mathcal{L}_n(h^\star; \widehat{w}) - \mathcal{L}_n(h^\star)}_{(a)} + \underbrace{\mathcal{L}_n(h^\star) - \mathcal{L}(h^\star)}_{(b)} \qquad (9)\end{aligned}$$

where $n = \beta n_p$ and we used optimality of $\widehat{h}_{\widehat{w}}$. Here (a) is the weight estimation error and (b) is the finite sample error.

Uniform law for bounding (b) We bound (b) using standard results for uniform laws for uniformly bounded functions which holds since $\|w\|_\infty \leq d_\infty(q||p)$ and $\|\ell\|_\infty \leq 1$. Since $|w(y)\ell(h(x), y)| \leq d_\infty(q||p),\ \forall x, y \in \mathcal{X} \times \mathcal{Y}$, by deploying the McDiarmid's inequality we then obtain that

$$\sup_{h\in\mathcal{H}} |\mathcal{L}_n(h) - \mathcal{L}(h)| \leq 2\mathcal{R}_n(\mathcal{G}(\ell, \mathcal{H})) + d_\infty(q||p)\sqrt{\frac{\log\frac{2}{\delta}}{n}}$$

where $\mathcal{G}(\ell, \mathcal{H}) = \{g_h(x, y) = w(y)\ell(h(x), y) : h \in \mathcal{H}\}$ and the Rademacher complexity is defined as $\mathcal{R}_n(\mathcal{G}) := \mathbb{E}_{X(i),Y(i)\sim\mathbb{P}:i\in[n]}\left[\mathbb{E}_{\sigma_i:i\in[n]}\frac{1}{n}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n \sigma_i w(y_i)\ell(x_i, h(y_i))\right]\right]$.

of the hypothesis class $\mathcal{H}$ (see for example Percy Liang notes on Statistical Learning Theory and chapter 4 in Wainwright (2019))

Bounding term (a) Remember that $k = |\mathcal{Y}|$ is the cardinality of the finite domain of Y, or the number of classes. Let us define $\widetilde{\ell} \in \mathcal{R}^k$ with $\widetilde{\ell}_j = \sum_{i=1}^n \mathbb{I}_{y(i)=j}\ell(y(i), h(x(i)))$. Notice that by definition $\|\widetilde{\ell}\|_1 \leq n$ and $\|\widetilde{\ell}\|_\infty \leq n$ from which it follows by Hoelder's inequality that $\|\widetilde{\ell}\|_2 \leq n$. Furthermore, we slightly abuse notation and use w to denote the k-dimensional vector with $w(i) = w(i)$. Therefore, for all h we have via the Cauchy Schwarz inequality that

$$\begin{aligned}|\mathcal{L}_n(h; w) - \mathcal{L}_n(h; \widehat{w})| &= |\frac{1}{n}\sum_{i=1}^n (w(y(i)) - \widehat{w}(y(i)))\ell(h(x(i)), y(i))| \\ &\leq |\frac{1}{n}\sum_{j=1}^k (w(j) - \widehat{w}(j))\widetilde{\ell}(j)| \\ &\leq \frac{1}{n}\|\widehat{w} - w\|_2\|\widetilde{\ell}\|_2 \leq \|\widehat{w} - w\|_2 \\ &\leq \|\lambda\widehat{\theta} - \theta\|_2 \leq (1-\lambda)\|\theta\|_2 + \lambda\|\widehat{\theta} - \theta\|_2\end{aligned}$$

It then follows by Lemma 1 that

$$\sup_{h\in\mathcal{H}} |\mathcal{L}_n(h; w) - \mathcal{L}_n(h; \widehat{w})| \leq (1-\lambda)\|\theta\|_2$$
$$O\left(\frac{\lambda}{\sigma_{min}}((\|\theta\|_2)\sqrt{\frac{\log(k/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(1/\delta)}{(1-\beta)n_p}}\sqrt{\frac{\log(1/\delta)}{n_q}})\right)$$

Plugging both bounds back into equation (9) concludes the proof of the theorem.

A.3 GENERALIZATION FOR FINITE HYPOTHESIS CLASSES

For finite hypothesis classes, one may bound (b) in (9) using Bernstein's inequality.

We bound (b) by first noting that $w(Y)\ell(Y, h(X))$ satisfies the Bernstein conditions so that Bernstein's inequality holds

$$\mathbb{E}_{\mathbb{P}}[w(Y)] = 1, \quad \mathbb{E}_{\mathbb{P}}[w(Y)^2] = d(q||p), \quad \sigma^2_{\mathbb{P}}(w(Y)) = d(q||p) - 1 \tag{10}$$

by definition. Because we assume $\ell \leq 1$, we directly have

$$\mathbb{E}_{\mathbb{P}}\left[w(Y)^2 l^2(Y, h(X))\right] \leq \mathbb{E}_{\mathbb{P}} w(Y)^2 = d(q||p) \tag{11}$$

Since we have a bound on the second moment of weighted loss while its first moment is $L(h)$ we can apply Bernstein's inequality to obtain for any fixed h that

$$|\mathcal{L}_n(h) - \mathcal{L}(h)| \leq \frac{d_\infty(q||p)\log(\frac{2}{\delta})}{3n} + \sqrt{\frac{2\left(d(q||p) - \mathcal{L}(h)^2\right)\log(\frac{2}{\delta})}{n}}$$

For the uniform law for finite hypothesis classes we can just do union bound on top of Bernstein and could get a slightly better dependence on $d_\infty(q||p)$.

The second moment of the importance weighted loss $\mathbb{E}_{\mathbb{P}}\left[\omega(Y)^2\ell^2(h(X), Y)\right]$, given a $h \in \mathcal{H}'$ can be bounded for general $\alpha \geq 0$, potentially leading to smaller values than $d(q||p)$:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\left[\omega_Y^2 \ell^2(Y, h(X))\right] \\
&\quad = \sum_i^k p(i)\frac{q^2(i)}{p^2(i)}\mathbb{E}_{X\sim p(X|Y=i)}\left[\ell^2(h(X), X)\right] \\
&\quad = \sum_i^k q(i)^{\frac{1}{\alpha}}\frac{q(i)}{p(i)}q(i)^{\frac{\alpha-1}{\alpha}}\mathbb{E}_{X\sim p(X|Y=i)}\left[\ell^2(h(X), i)\right] \\
&\quad \leq \left(\sum_i^k q(i)\frac{q(i)^\alpha}{p(i)^\alpha}\right)^{\frac{1}{\alpha}}\left(\sum_i^k q(i)\mathbb{E}_{X\sim p(X|Y=i)}\left[\ell^2(h(X), i)\right]^{\frac{2\alpha}{\alpha-1}}\right)^{\frac{\alpha-1}{\alpha}} \\
&\quad = \left(\sum_i^k q(i)\frac{q(i)^\alpha}{p(i)^\alpha}\right)^{\frac{1}{\alpha}}\left(\sum_i^k q(i)\mathbb{E}_{X\sim p(X|Y=i)}\left[l^2(h(X), i)\right]\mathbb{E}_{X\sim p(X|Y=i)}\left[\ell^2(h(X), i)\right]^{\frac{\alpha+1}{\alpha-1}}\right)^{\frac{\alpha-1}{\alpha}} \\
&\quad \leq \left(\sum_i^k q(i)\frac{q(i)^\alpha}{p(i)^\alpha}\right)^{\frac{1}{\alpha}}\sum_i^k q(i)\mathbb{E}_{X\sim p(X|Y=i)}\left[\ell^2(h(X), i)\right]^{1-\frac{1}{\alpha}}\mathbb{E}_{X\sim p(X|Y=i)}\left[\ell^2(h(X), i)\right]^{1+\frac{1}{\alpha}}
\end{aligned} \tag{12}$$

where the first inequality follows from Hölder's inequality, the second one follows from Jensen's inequality and the fact that the loss is in $[0, 1]$ as well as the fact that the exponentiation function is convex in this region. Moreover, since $1 + \frac{1}{\alpha} \geq 1$ and upper bound for the loss square, $l(\cdot, \cdot)^2 \leq 1$, then;

$$\mathbb{E}_{\mathbb{P}}\left[w(Y)^2\ell^2(h(X), Y)\right] \leq \left(\sum_i^k q(i)\frac{q(i)^\alpha}{p(i)^\alpha}\right)^{\frac{1}{\alpha}}\sum_i^k q(i)\mathbb{E}_{X\sim\mathbb{P}|Y=i}\left[\ell^2(h(X), i)\right]^{1-\frac{1}{\alpha}}$$

which gives bound on the second moment of weighted loss.