EE 377/STATS 311 Project Report: Learning from Different Domains

Farzan Farnia, Nishal Shah & Milind Rao
{farnia,nishalps,milind}@stanford.edu
Dept. of Electrical Engineering
Stanford University

1 Introduction

Most machine learning algorithms are theoretically designed and practically applied with the consideration that the training data and test data arise from a common distribution. In several applications in natural language processing ranging from information extraction or sentiment analysis, classification algorithms are trained using a limited set of documents and applied generally to works spanning different genres. Also, in genetics, most experiments are done and predictive methods developed for people from a particular geography but this is broadly applied to all humans whose genetic makeup arises from a different distribution. It is of interest to learn how efficiently algorithms trained on samples from a source distribution perform on test data arising from a target distribution. In [1], the authors investigate this problem of *domain adaptation*.

The authors first consider the problem of binary classification trained on labeled data from a source distribution. Classification error bounds are presented in terms of source domain error and divergence measures between the two distributions. Unlabeled data from either domain is then leveraged to estimate the divergence between domains to present a second bound. The third question the authors answer is learning with different amounts of labeled target and source data and they do this via a hypothesis that minimizes a convex combination of the error.

We aim to extend the results of the paper in a few directions. In Section 2, we introduce the problem setup. We then present a simple upper bound on the generalization error with different domains. The paper provides this result for only binary function with 0-1 loss. We extend it to arbitrary functions and a large class of loss functions. Seeking to make this bound tighter, we define a classifier based distance for a larger class of hypothesis and loss functions. We now extend the results by giving guarantees on an algorithm that seeks to minimize a convex combination of training error and test error. We attempt to use this bound to show what amount of regularization is appropriate. In Section 3, we present an alternate view of looking at the correct amount of regularization. Finally, we validate some of the theoretical results with simulations in Section 4. Conclusions are presented in Section 5.

2 Problem Setup and Theoretical Results

2.1 Problem Setup

Let us consider data $x \in \mathbb{R}^d$ which could arise from a source distribution D_S with pdf ϕ_S or a target distribution (D_T and ϕ_T). For data in the source distribution, We have $y = f_S(x)$ and for for data in the target distribution, $y = f_T(x)$. Here, f_S and f_T are deterministic multilabel functions which we aim to generalize further to random functions. For hypothesis $h \in \mathcal{H}$, consider bounded loss functions $|\ell(h, x, f)| \leq 1$. We suppose that there are n unlabelled points from each distribution and βm and $(1-\beta)m$ labelled points in training and test data sets respectively.

2.2 Generalized Results

2.2.1 A weak bound

The first theorem generalizes what is present in the paper and connects the estimation error $\mathbb{E}[\ell(h, X, f_T)]$ in the target distribution to what was learned on the training set.

Theorem 1. *With the set-up described above,*

$$\mathbb{E}_{D_T}[\ell(h,x,f_T)] \leq \mathbb{E}_{D_S}[\ell(h,x,f_S)] + \min\{\Pr_{D_S}(f_T \neq f_S), \Pr_{D_T}(f_T \neq f_S)\} + 2\|\phi_S - \phi_T\|_{TV}$$

Proof.

$$\begin{aligned}
\mathbb{E}_{D_T}[\ell(h,x,f_T)] &= \mathbb{E}_{D_S}[\ell(h,x,f_S)] + \mathbb{E}_{D_T}[\ell(h,x,f_T)] - \mathbb{E}_{D_T}[\ell(h,x,f_S)] + \mathbb{E}_{D_T}[\ell(h,x,f_S)] - \mathbb{E}_{D_S}[\ell(h,x,f_S)] \\
&\leq \mathbb{E}_{D_S}[\ell(h,x,f_S)] + |\int_x \phi_T(x)(\ell(h,x,f_T) - \ell(h,x,f_T))dx| + |\int_x (\phi_T(x) - \phi_S(x))\ell(h,x,f_S)dx| \\
&\leq \mathbb{E}_{D_S}[\ell(h,x,f_S)] + \mathbb{E}_{D_T}[\mathbf{1}(f_T \neq f_S)] + 2\|\phi_T - \phi_S\|_{TV}
\end{aligned}$$

We can also split the first equality differently leading to the other inequality. □

We can extend the proof to continuous bounded labelling functions as well. Consider a natural Lipschitz type constraint on the loss function - $|\ell(h,x,f) - \ell(h,x,f')| \leq L|f(x) - f(x')|$. This holds for a wide variety of loss functions such as $\ell(h,x,f) = |h(x) - f(x)|$. Now, recognize that if $|\ell(h,x,f_S) - \ell(h,x,f_T)| \geq \delta \Rightarrow |f_S(x) - f_T(x)| \geq \delta/L$. using these in the previous proof for any $\delta > 0$,

$$\begin{aligned}
\mathbb{E}_{D_T}[\ell(h,x,f_T)] &\leq \mathbb{E}_{D_S}[\ell(h,x,f_S)] + \delta + |\int_x \phi_T(x)\mathbf{1}(|f_S(x) - f_T(x)| \geq \delta/L)dx| + |\int_x (\phi_T(x) - \phi_S(x))\ell(h,x,f_S)dx| \\
&\leq \mathbb{E}_{D_S}[\ell(h,x,f_S)] + \delta + \min\{\Pr_{D_S}(|f_S - f_T| \geq \delta/L), \Pr_{D_T}(|f_S - f_T| \geq \delta/L)\} + 2\|\Phi_S - \Phi_T\|_{TV}
\end{aligned}$$

If there is no difference in labelling functions between the two data sets, the error is bounded by the total-variation distance between the two datasets. This is a very loose bound as it does not take into account the loss function. Here is an example where it is loose. $\mathcal{H} = \{x \mapsto \mathbf{1}(1 \leq b) : b \in \mathbb{R}_-\}, \ell(h,x,f) = |h(x) - f(x)| \forall h, f \in \mathcal{H}$. Now consider source and target distributions on real line that differ on $x \geq 0$. In this case, there should be no error but the total variation distance is non-zero. Now we give an example where this bound is tight. Consider $x \in \{0,1\}, f_S = f_T = \mathbf{1}_{\{1\}}, h = \mathbf{1}_{\{0\}}, \Phi_S = \mathcal{B}(\rho_1), \Phi_T = \mathcal{B}(\rho_2), \ell(h,x,f) = \mathbf{1}(h(x) \geq f(x)) - \mathbf{1}(h(x) \leq f(x))$. In this case, the bound of $2\|\Phi_S - \Phi_T\|_{TV} = 2(\rho_2 - \rho_1)$ holds tightly.

2.2.2 Towards stronger bounds

We now aim to modify results in the paper to include bounds with Rademacher complexity instead of VC dimensions to give tighter bounds in some conditions. We present here an extension to the lemma giving performance bounds on the estimate of a general classifier based distance measure from empirical estimates. This distance metric is defined in [2]. It is smaller than the total variation distance as it restricts the subsets over which measures are taken by the hypothesis class.

We have,

$$\begin{aligned}
d_{\mathcal{G}}(D, D') &= \sup_{g \in \mathcal{G}} |\mathbb{E}_D[g(x)] - \mathbb{E}_{D'}[g(x')]| \\
&= \sup_{g \in \mathcal{G}} |\mathbb{E}[g'(z)]|,
\end{aligned}$$

where $z = (x, x')$, $x \overset{\text{iid}}{\sim} D$, $x' \overset{\text{iid}}{\sim} D'$, $g'(z) = g(x) - g(x')$. Similarly, the empirical estimate with n samples from D, D' is,

$$\hat{d}_{\mathcal{G}}(z_1^n) = \sup_{g \in \mathcal{G}} |\frac{1}{n} \sum_{i \in [n]} g'(z_i)|$$

Now observe that $|d_{\mathcal{G}} - \hat{d}_{\mathcal{G}}(z_1^n)| \geq \epsilon$ implies $\exists g \in \mathcal{G} ||\mathbb{E}[g'(z)]| - |\frac{1}{n}\sum_{i=1}^n g'(z_i)|| \geq \epsilon$. This is seen for the g that either maximizes the real distance measure or the empirical one. This in turn implies by triangular inequality that $\sup_g |\mathbb{E}[g'(z)] - \frac{1}{n}\sum_{i=1}^n g'(z_i)| \geq \epsilon$. Denote $G_n(Z = z_1^n) = \sup_g \mathbb{E}[g'(z)] - \frac{1}{n}\sum_{i=1}^n g'(z_i)$. We observe bounded differences if one of the parameter changes and hence by McDiarmid inequality,

$$\begin{aligned}|G_n(Z) - G_n(Z_{\setminus i}, z_i')| &\leq \frac{2}{n} \\ \Pr(G_n \geq \mathbb{E}[G_n] + \epsilon) &\leq \exp\left(\frac{-n\epsilon^2}{2}\right)\end{aligned}$$

Here, $Z_{\setminus i}$ indicates all components of Z except the i^{th} one.

Now, we use Rademacher complexity to bound

$$\begin{aligned}\mathbb{E}[G_n] &\leq 2\mathbb{E}\left[\sup_g \frac{1}{n}\sum_{i=1}^n \sigma_i g'(z)\right] \\ &\leq 2\mathcal{R}_n(\mathcal{G})\end{aligned}$$

Combining these, we get a lemma which tells us how quickly we can learn the classifier based distance from sample values.

Lemma 1. *With probability $\geq 1-\delta$, we get the following bound on the estimate of the binary classifier based distance,*

$$d_{\mathcal{G}} \leq \hat{d}_{\mathcal{G}}(z_1^n) + 2\mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{2\log(2/\delta)}{n}}$$

We apply Lemma 1 to the symmetric difference hypothesis class defined as,

$$g_{h,h'} \in \mathcal{G} = \mathcal{H}\Delta\mathcal{H} \Rightarrow g_{h,h'}(x) = \ell(h, x, h'). \tag{1}$$

We obtain the following lemma, an extension of that in the paper to continuous labelling functions.

Lemma 2. *From the definition of the classifier based distance*

$$|\mathbb{E}_{D_S}[\ell(h, X, h')] - \mathbb{E}_{D_T}[\ell(h, X, h')]| \leq d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$$

We focus on loss functions and hypothesis classes that satisfy certain properties:

1. Loss functions:
$$\ell(h_1, x, h_2) \leq \ell(h_1, x, h_3) + \ell(h_3, x, h_2)$$
Examples include $\mathbf{1}(h_1 \neq h_2), |h_1 - h_2|$.

2. There is jointly optimum hypothesis h^*:
$$\lambda = \inf_{h \in \mathcal{H}} \mathbb{E}_{D_S}[\ell(f_S, x, h)] + \mathbb{E}_{D_T}[\ell(h, x, f_T)]$$

Theorem 2. *With the definitions and conditions holding from above,*

$$\mathbb{E}_{D_T}[\ell(h, x, f_T)] \leq \mathbb{E}_{D_S}[\ell(h, x, f_S)] + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(z_1^n) + 2\mathcal{R}_n(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{2\log(2/\delta)}{n}} + \lambda$$

Proof. Above conditions are applied to obtain

$$\begin{aligned}\mathbb{E}_{D_T}[\ell(h, x, f_T)] &\leq \mathbb{E}_{D_T}[\ell(h, x, h^*)] + \mathbb{E}_{D_T}[\ell(h^*, x, f_T)] \\ &\leq \mathbb{E}_{D_T}[\ell(h^*, x, f_T)] + \mathbb{E}_{D_S}[\ell(h, x, h^*)] + |\mathbb{E}_{D_T}[\ell(h, x, h^*)] - \mathbb{E}_{D_S}[\ell(h, x, h^*)]| \\ &\leq \mathbb{E}_{D_T}[\ell(h^*, x, f_T)] + \mathbb{E}_{D_S}[\ell(f_S, x, h^*)] + \mathbb{E}_{D_S}[\ell(h, x, f_S)] + d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \\ &\leq \mathbb{E}_{D_S}[\ell(h, x, f_S)] + \lambda + d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)\end{aligned}$$

□

Now given some labeled data from source $((1-\beta)m)$ and target $((\beta)m)$, we aim to solve find the best hypothesis by minimizing an α combination of the empirical errors (denoted by $\hat{E}$).

$$\hat{E}_\alpha(h) = \alpha \hat{E}_T(h,x,y) + (1-\alpha)\hat{E}_S(h,x,y)$$
$$\mathbb{E}_\alpha(h) = \alpha \mathbb{E}_{D_S}[\ell(h,x,f_S)] + (1-\alpha)\mathbb{E}_{D_T}[\ell(h,x,f_T)]$$

It can be easily observed from the previous result and theorem that

$$|\mathbb{E}_\alpha(h) - \mathbb{E}_{D_T}[h,x,f_T]| \leq (1-\alpha)|d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T) + \lambda| \tag{2}$$

We now present a bound on how far $\hat{E}_\alpha$ is from $\mathbb{E}_\alpha$. We can write the difference as,

$$\hat{E}_\alpha(h) - \mathbb{E}_\alpha(h) = \frac{\alpha}{\beta m}\sum_{i=1}^{\beta m}(\ell(h,x_i,f_S) - \mathbb{E}_{D_S}(h,x,f_S)) + \frac{1-\alpha}{(1-\beta)m}\sum_{i=\beta m+1}^{m} \ell(h,x_i,f_T) - \mathbb{E}[\ell(h,x_i,f_T)]$$

Similar to proof of Lemma 1, let $G_n = \sup_h \hat{E}_\alpha(h) - \mathbb{E}_\alpha(h)$. We have bounded differences for G_n with the bound being $\frac{2\alpha}{\beta m}$ when $i \leq \beta m$ variable is changed and $\frac{2(1-\alpha)}{(1-\beta)m}$ otherwise. From McDiarmid inequality,

$$\Pr(G_n - \mathbb{E}[G_n] \geq \epsilon) \leq \exp\left(-\frac{2m\epsilon^2}{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}}\right)$$

Also, we use mechanism of Rademacher complexity to conclude that,

$$\mathbb{E}[G_n] \leq 2\mathcal{R}_m(\mathcal{H}\Delta\mathcal{H}).$$

Thus, with probability $\geq 1-\delta$, we have

$$|\hat{E}_\alpha(h) - \mathbb{E}_\alpha(h)| \leq 2\mathcal{R}_m(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} \tag{3}$$

Theorem 3. *Suppose we have loss functions and hypothesis classes that satisfy constraints, we have n unlabelled points in each points, βm labelled source distribution points and $(1-\beta)m$ target distribution points and we find the empirical risk minimizer of an alpha combination $(\hat{h})$ and obtain with probability $\geq 1-2\delta$,*

$$\mathbb{E}_{D_T}[\ell(\hat{h},x,f_T)] \leq \mathbb{E}_{D_T}[\ell(h^*,x,f_T)] + 4\mathcal{R}_m(\mathcal{H}\Delta\mathcal{H}) + 2\sqrt{\frac{\log(2/\delta)}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)}$$
$$+ 2(1-\alpha)(\hat{d}_{\mathcal{H}\Delta\mathcal{H}} + 2\mathcal{R}_n(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{2\log(2/\delta)}{n}} + \lambda)$$

Proof. The proof employs a combination of all previous lemmas and theorems. We step through it here,

$$\begin{aligned}
\mathbb{E}_{D_T}[\ell(\hat{h},x,f_T)] &\leq \mathbb{E}_\alpha(\hat{h}) + (1-\alpha)(d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T) + \lambda) \\
&\leq \hat{E}_\alpha(\hat{h}) + 2\mathcal{R}_m(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} + (1-\alpha)(d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T) + \lambda) \\
&\leq \hat{E}_\alpha(h^*) + 2\mathcal{R}_m(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} + (1-\alpha)(d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T) + \lambda) \\
&\leq \mathbb{E}_\alpha(h^*) + 4\mathcal{R}_m(\mathcal{H}\Delta\mathcal{H}) + 2\sqrt{\frac{\log(2/\delta)}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} + (1-\alpha)(d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T) + \lambda) \\
&\leq \mathbb{E}_{D_T}[\ell(h^*,x,f_T)] + 4\mathcal{R}_m(\mathcal{H}\Delta\mathcal{H}) + 2\sqrt{\frac{\log(2/\delta)}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} \\
&\quad + 2(1-\alpha)(\hat{d}_{\mathcal{H}\Delta\mathcal{H}} + 2\mathcal{R}_n(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2n}} + \lambda)
\end{aligned}$$

□

2.3 Applications and Examples

Consider a regression setting. Let $\mathcal{H} = \{x \mapsto w^\intercal x : w \in \mathbb{R}^d, \|w\|_2 \leq B_w\}$ for constant $B_w \leq 1/2$. x in this case is bounded as $\mathbb{E}[\|x\|_2^2] \leq B^2$. Let $\ell(h_w, x, h_{w'}) = |h_w(x) - h_{w'}(x)|/B$ for $h_w, h_{w'} \in \mathcal{H}$. It can be seen that $\mathcal{R}_n(\mathcal{H}\Delta\mathcal{H}) \leq 2B_w/\sqrt{n}$. Let $f_s = h_{w_S}, f_T = h_{w_T}$. Let distributions be $D_S = \mathcal{N}(0, B^2/dI), D_T = \mathcal{N}(0, B^2/4d)$. We can calculate

$$\begin{aligned}
\lambda &= \inf_w \mathbb{E}_{D_S}[|(w - w_S)^\intercal x|/B] + \mathbb{E}_{D_T}[|(w - w_T)^\intercal x|/B] \\
&\leq \inf_w \sqrt{\mathbb{E}_{D_S}[((w - w_S)^\intercal x)^2]}/B + \sqrt{\mathbb{E}_{D_T}[|(w - w_T)^\intercal/Bx|^2]} \\
&\leq \inf_w \frac{\|w - w_S\|_2}{\sqrt{d}} + \frac{\|w - (w_T)\|_2}{2\sqrt{d}} \\
&\leq \frac{\|w_S - w_T\|_2}{3\sqrt{d}}
\end{aligned}$$

The bound becomes

$$\mathbb{E}_{D_T}[\ell(\hat{h}, x, f_T)] \leq \mathbb{E}_{D_T}[\ell(h^*, x, f_T)] + 4\frac{B_w}{\sqrt{m}} + 2\sqrt{\frac{\log(2/\delta)}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} +$$

$$2(1-\alpha)(\hat{d}_{\mathcal{H}\Delta\mathcal{H}} + 2\frac{B_w}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{2n}} + \frac{\|w_S - w_T\|_2}{3\sqrt{d}})$$

We can see that if increase B_w which is the regularization term, $\hat{d}_{\mathcal{H}\Delta\mathcal{H}} + \mathbb{E}_{D_T}[\ell(h^*, x, f_T)]$ goes down but the rademacher complexity term goes up. There is an appropriate regularization to use. Depending on β, the bound can be optimized to find the right mixing ratio α which also determines the appropriate regularization to use. As the number of labelled or unlabelled data points increase, the error falls. As the difference between the labelling functions decrease, the error bound falls as expected. This bound is much tighter than the one in [1] because there is no dependence on the dimension.

3 Regularization in different domains

In this section, we see another view to analyse the amount of regularization to use with different domains. For simplicity, consider a ridge-regression problem where we are interested in finding $\boldsymbol{\beta}^*$ minimizing

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \|\mathbf{X}\boldsymbol{\beta} - Y\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{4}$$

To see how the effect of regularization in reducing variance could be different in the multiple domains case consider the following two scenarios:

1. $y_i = \mathbf{x}_i\gamma + \epsilon_i$, $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ for both source and test domains, we have N samples in total. Here ϵ_i is drawn i.i.d. according to $\mathcal{N}(0, 1)$.

2. $y_i = \mathbf{x}_i\gamma + \epsilon_i$, $\mathbf{x}_S \sim \mathcal{N}(\mu, \Sigma_1)$, $\mathbf{x}_T \sim \mathcal{N}(-\mu, \Sigma_2)$ we have $N/2$ samples of each. Here ϵ_i is drawn i.i.d. according to $\mathcal{N}(0, 1)$.

It can be seen that if τ_i is the ith eigenvalue of $\mathbf{X}^T\mathbf{X}$, then we can formulate the bias and variance of (4) as follows

$$\text{Bias}^2 = \sum_{i=1}^{d} \frac{\tau_i\lambda^2\gamma_i^2}{(\tau_i + \lambda)^2} \tag{5}$$

$$\text{Var} = \frac{\sigma^2}{n}\sum_{i=1}^{d} \frac{\tau_i^2}{(\tau_i + \lambda)^2}. \tag{6}$$

To see the effect of multiple domains, assume in Scenario 2, every sample in $\mathbf{X}_S$ is orthogonal to samples in $\mathbf{X}_T$. Then eigenvectors of $\mathbf{X}_S^T\mathbf{X}_S$ and $\mathbf{X}_T^T\mathbf{X}_T$ are orthogonal to each other meaning that in a high-dimensional space both Variance and Bias-square are roughly doubled, that needs higher coefficient of regularization to reach a better bias-variance trade off, compared to Scenario 1.

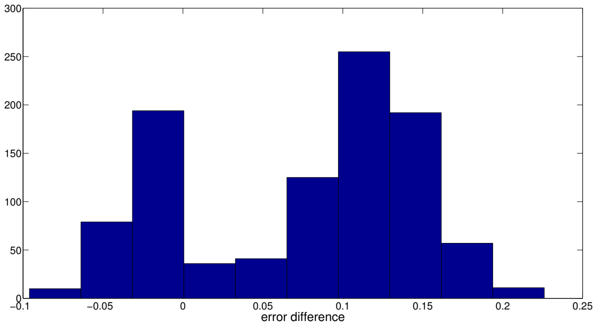

Figure 1: histogram of $e_{\text{unmerged}} - e_{\text{merged}}$ for $q_1 = 0.4, q_2 = 0.6, \beta_1 = \beta_2$

4 Numerical Results

To evaluate the main idea of the paper, we did several numerical experiments. Here, we sought to test the idea that we can learn on data drawn from a distribution D_S and then predict on data points drawn from a different distribution D_T.

A simple idea is to perturb the predictors' distribution, i.e the marginal $D_{\mathbf{X}}$ because if we apply a discriminative learning approach like logistic regression the goal would be to learn the conditional $D_{Y|\mathbf{X}}$ that is independent from the marginal $D_{\mathbf{X}}$. To experiment this idea, first we generated n_1 data points $(\mathbf{x}_i, y_i)_{i=1}^{n_1}$, with $\mathbf{X} \in \{0,1\}^p$ drawn i.i.d. from a Bernoulli distribution with parameter $P(X_i = 1) = q_1$ and

$$P(Y_i = 1) = \frac{1}{1 + \exp(-\beta_1^* \mathbf{x}_i)}. \tag{7}$$

Also we draw n_2 datapoints $(\mathbf{x}_i, y_i)_{i=n_1+1}^{n_1+n_2}$, with $\mathbf{X} \in \{0,1\}^p$ drawn i.i.d. from a Bernoulli distribution with parameter $P(X_i = 1) = q_2$, and

$$P(Y_i = 1) = \frac{1}{1 + \exp(-\beta_2^* \mathbf{x}_i)}. \tag{8}$$

We expected that if $\beta_1^* \approx \beta_2^*$ then even if q_1 and q_2 are very different we would learn better when we merge the two groups of data points together.

To test this hypothesis, we took $\beta_1^* = \beta_2^* \sim \mathcal{N}(\mathbf{0}, I_p)$ and $q_1 = 0.6$ and $q_2 = 0.4$. We used the built-in Glmfit function in Matlab and averaged the results over 1000 Monte Carlo runs. As expected, the test accuracy rate raised by 0.078 with merging the two datasets together. In Figure 1, the histogram indicates that the test error when we separately learn over datasets is usually larger than the test error when we merge the data points and learn over the merged dataset. In order to evaluate the effect of the distance $q_1 - q_2$, the second time we set $q_1 = 0.9$ and $q_2 = 0.1$ and did the experiment again but this time the test accuracy dropped by 0.02. In Figure 2, one can observe how the accuracy rate drops by increasing $q_1 - q_2$.

We also did another experiment to understand how raising the ratio of samples coming from source distribution can affect the performance of the trained model over the target distribution. To this end, we trained a logistic regression model given 1000 samples with 50 features. The underlying distribution for both target and source samples is i.i.d. Gaussian with identity covariance matrix, with random Normally-distributed mean vectors. For each fixed proportion we averaged the results over 100 Monte Carlo runs. In Figure 3, we plotted the drop in error-rate, compared to the case all samples coming from target distribution, versus the proportion of samples drawn from the target distribution. We marked different mean vector distances with different colors. Observe how the error difference increases to 0 while increasing the proportion of samples drawn from target distribution. Also notice how raising the distance makes the difference larger.

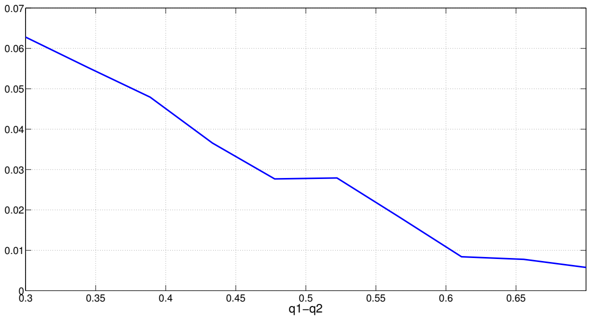

Figure 2: graph of accuracy rate change vs. $q_1 - q_2$

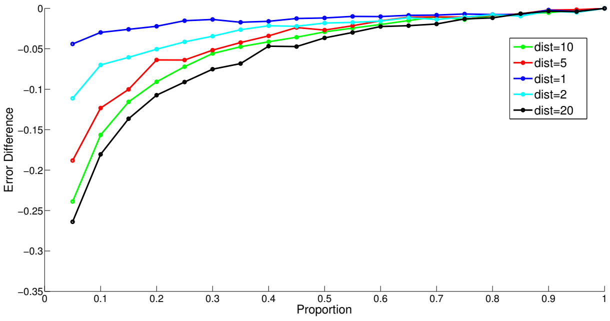

Figure 3: Error difference vs. target sample proportion

Model complexity and domain adaptation

The paper [1] bounds the error rate in target domain by bounding the error rate in terms which decrease when we make the hypothesis class more complex (corresponds to better classification between source and target domain unlabelled data) and another term which increases when we make the hypothesis class more complex.

It is known that regularization helps when the amount of data is small compared to the parameter dimensions by restricting the learnt parameters to belong to small hypothesis classes. But does regularization help when the training and testing distributions are different?

We study this behavior in the context of ridge regression, where we learn a regularized estimator when data is generated by $Y = \theta^T X + \epsilon$, with $X \sim \mathcal{N}(0, I)$ for source distribution, and $\epsilon \sim \mathcal{N}(0, I)$. We vary the target test distribution to be $X \sim \mathcal{N}(\mu e_d, I)$, where e_d is the d dimensional vector of all ones. We choose $d = 10$, and vary the number of training samples from source distribution to estimate θ with different regularization parameter.

As we can see from Figure 4, when θ is estimated from limited number of source samples, there exists some values of regularization coefficients which have much lower test error on very different target distributions compared to unregularized ($\lambda = 0$) or over regularized estimators. In other words, regularization is very important for test domains which are very different from source domains. Note that the 'best' regularization amount is same for all target domains considered. However, once we have enough data in source domain, the unregularized estimate performs well.

Hence, controlling model complexity by regularization makes estimator robust to changes test domain relative to training domain.

5 Conclusions

In this paper, we have considered the problem of domain adaptation where a learning algorithm may receive training and test data algorithms from different distributions. In particular, we have looked at

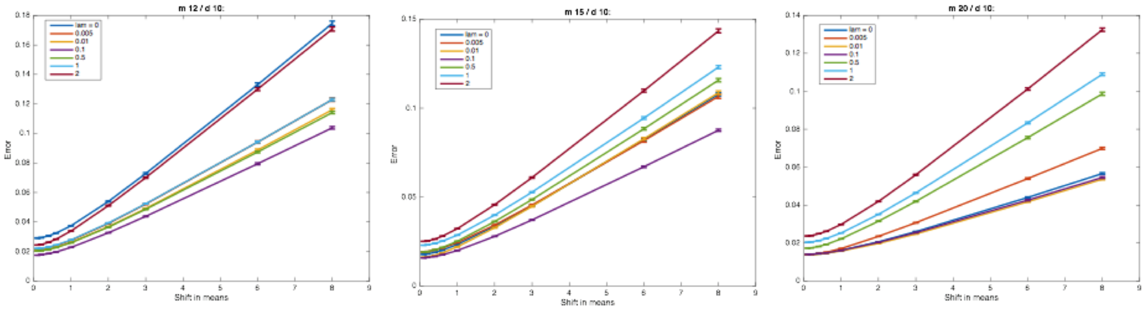

Figure 4: Test error on target distribution when the parameter is estimated with different number of source samples 'm' and distance between the source and target distribution is varied. Different lines correspond to different regularization amounts.

algorithms to minimize error with labelled and unlabelled data of different sizes from each domain. We have extended the results of [1] to to arbitrary hypothesis and loss classes while analysing the error of an algorithm which minimizes a convex combination of training and test error. The contribution of the paper was to find out the optimal way to perform regularization in the domain adaptation problem. Regularization reduces error to a larger extent when we have few data points and the domain and target distributions widely vary.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[2] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting change in data streams," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004, pp. 180–191.