1 Problem Statement

We aim to perform density ratio estimation: given samples from distributions $P_X^A(\cdot), P_X^B(\cdot)$, estimate

$$w(x) := \frac{P_X^B(x)}{P_X^A(x)}.$$

We want to do this because suppose you have training data, which consists of N^A samples $\{x_i^A, y_i^A\}$ drawn iid from the training distribution: $X_i^A, Y_i^A \sim P_{X,Y}^A(\cdot) = P_X^A(\cdot)P_{Y|X}(\cdot)$. You construct a predictive function $f(\cdot)$, and want to know the expected loss under the test distribution: $P_{X,Y}^B(\cdot) = P_X^B(\cdot)P_{Y|X}(\cdot)$. An unbiased estimate of the expected test loss is

$$L(f, P_{X,Y}^B) = \sum_{i=1}^{N^A} w(x_i^A) L(y_i^A, f(x_i^A)), \tag{1}$$

and so we need to find estimates $\hat{w}(x_i^A)$ of each $w(x_i^A)$, the true ratio at the covariates appearing in the training data. Note the train and test distributions are assumed to differ only in the marginal distributions: $P_X^A(\cdot) \neq P_X^B(\cdot)$, but $P_{Y|X}^A(\cdot) = P_{Y|X}^B(\cdot)$. This link between train/test distributions is called covariate shift.

2 Past Work in Density Ratio Estimation

2.1 Ratio Matching Methods

2.1.1 General Formulation

This section is mostly a summary of [3]. One line of work seeks to learn a function $\hat{w}(\cdot)$ minimizing $E_{X\sim P_X^A(\cdot)}[d_f(w(X), \hat{w}(X))]$, where $d_f(t, \hat{t}) := f(t) - (f(\hat{t}) + \nabla f(t - \hat{t}))$ is some Bregman divergence parameterized by f that quantifies the distance between true ratios and estimated ratios:

$$\begin{aligned}
E_{X\sim P_X^A(\cdot)}[d_f(w(X), \hat{w}(X))] &= E_{X\sim P_X^A(\cdot)}[f(w(X)) - f(\hat{w}(X)) - \nabla f(\hat{w}(X))(w(X) - \hat{w}(X))] &(2)\\
&= -E_{X\sim P_X^A(\cdot)}[f(\hat{w}(X))] - E_{X\sim P_X^A(\cdot)}[\nabla f(\hat{w}(X))w(X)] + E_{X\sim P_X^A(\cdot)}[\nabla f(\hat{w}(X))\hat{w}(X)] &(3)\\
&= -E_{X\sim P_X^A(\cdot)}[f(\hat{w}(X))] - E_{X\sim P_X^B(\cdot)}[\nabla f(\hat{w}(X))] + E_{X\sim P_X^A(\cdot)}[\nabla f(\hat{w}(X))\hat{w}(X)] &(4)
\end{aligned}$$

This expectation is unknown, but we can minimize over $\hat{w}(\cdot)$ an unbiased estimate, the empirical expectation:

$$\hat{E}_{X\sim P_X^A(\cdot)}[d_f(w(X), \hat{w}(X))] = -\tfrac{1}{N^A}\sum_i f(\hat{w}(x_i^A))] - \tfrac{1}{N^B}\sum_i \nabla f(\hat{w}(x_i^B))] + \tfrac{1}{N^B}\sum_i \nabla f(\hat{w}(x_i^A))\hat{w}(x_i^A)] \tag{5}$$

This empirical expectation can be calculated because we have samples from both $P_X^A(\cdot)$ and $P_X^B(\cdot)$, and the expression does not depend on the known true ratio function $w(\cdot)$.

2.1.2 Formulation with KL loss

Particular ratio matching methods differ on their choice of divergence d_f. If $f(t) = t\log t - t$, then $\nabla f(t) = \log t$ and the quantity to be minimized is:

$$\hat{E}_{X\sim P_X^A(\cdot)}[d_f(w(X), \hat{w}(X))] = \tfrac{1}{N^A}\sum_i \hat{w}(x_i^A) - \tfrac{1}{N^B}\sum_i \log \hat{w}(x_i^B) \tag{6}$$

Furthermore, $E_{X\sim P_X^A(\cdot)}[w(X)] = \int_x P_X^A(x)\frac{P_X^B(x)}{P_X^A(x)}dx = \int_x P_X^B(x)dx = 1$, and thus $\hat{w}(x)$ should satisfy the empirical version of this constraint: $\frac{1}{N^B}\sum_i \hat{w}(x_i^B) = 1$. Combining the objective function of Equation 11 with this constraint and a non-negativity constraint on the weight ratios gives the following problem:

$$\begin{aligned}
\min_{\{\hat{w}(x_i^B)\}_{i=1}^{N^B}} -\tfrac{1}{N^B}\sum_i \log \hat{w}(x_i^B) \quad &\text{subject to} &(7)\\
\tfrac{1}{N^B}\sum_i \hat{w}(x_i^B) = 1 &&(8)\\
\hat{w}(x_i^B) > 0 &&(9)
\end{aligned}$$

Note that this minimization only gives the length N^B vector of ratios at the *test* points $\{x_i^B\}$, whereas in Equation 1, one needs the ratios at the *training* points $\{x_i^A\}$, which are not available, because in the current formulation, we are not estimating the *entire* function $\hat{w}(\cdot)$, but rather just it's value at the finite set of testing points. To obtain the ratios at the training points, we need to learn the entire function $w(\cdot)$. One approach would be to pre-specify a set of K basis functions $\{\phi_k(\cdot)\}$ and assume

$$\hat{w}(\cdot) = \sum_k \alpha_k \phi_k(\cdot). \tag{10}$$

The optimization problem is then to learn the weights parameterizing the ratio function:

$$\min_{\{\alpha_k\}} -\tfrac{1}{N^B} \sum_i \log \sum_k \alpha_k \phi_k(x_i^B) \;\text{ subject to} \tag{11}$$

$$\tfrac{1}{N^B} \sum_i \sum_k \alpha_k \phi_k(x_i^B) = 1 \tag{12}$$

$$\alpha_k \geq 0 \tag{13}$$

This is a convex problem that scales with the number of basis functions, not the number of data points.

2.1.3 Formulation with squared loss

We can also consider the ratio matching method if the f parameterizing the divergence d_f is chosen to be $f(t) = \frac{1}{2}t^2$. Then, $\nabla f(t) = t$, and according to Equation 5, the objective to be minimized is:

$$\hat{E}_{X \sim P_X^A(\cdot)}[d_f(w(X), \hat{w}(X))] = \tfrac{1}{2N^A} \sum_i \hat{w}(x_i^A)^2 - \tfrac{1}{N^B} \sum_i \hat{w}(x_i^B) \tag{14}$$

Assuming the basis representation of Equation 10, and using the notation $\phi(x) := (\phi_1(x), \ldots, \phi_K(x))'$ and $\alpha := (\alpha_1, \ldots, \alpha_K)$, the optimization problem becomes:

$$\min_{\alpha} \tfrac{1}{2}\alpha' \Big(\tfrac{1}{N^A} \sum_i \phi(x_i^A)\phi(x_i^A)' \Big)\alpha - \Big(\tfrac{1}{N^B} \sum_i \phi(x_i^B) \Big)' \alpha \;\text{ subject to} \tag{15}$$

$$\alpha \geq 0 \tag{16}$$

Thus the squared loss formulation leads to a quadratic program which should be able to be solved more easily than the KL formulation (though both are convex). Regularization on the basis function coefficients α can be added in both the formulations. Under the lasso loss for α with penalty λ, the squared loss formulation has the advantage that the optimal α is piecewise linear as a function of λ, so that the optimal α for all possible λ can be enumerated, and model selection can be performed efficiently.

2.1.4 Cross-validation

One of the advantages of ratio matching methods is that cross validation can be used to choose the hyperparameters, $\{\phi_k(\cdot)\}$ and λ, governing the ratio-learning procedure. Cross validation is possible firstly because the learning procedure is *inductive*: an entire ratio function is learned using a training set, so that weight estimates can be obtained for a *separate* test set. Secondly, the quality of these test set weight estimates can be evaluated *directly*. As we shall see, neither of these is possible with kernel mean matching.

Both the ratio-learning procedure and the weight estimate evaluation step require samples from both $P_X^A(\cdot)$ and $P_X^B(\cdot)$. Thus, to evaluate the out-of-sample performance of a given hyperparameter setting, one can divide the data into k folds, where each fold contains samples from both $P_X^A(\cdot)$ and $P_X^B(\cdot)$, and repeatedly evaluate the ratio estimates on 1 fold obtained from the ratio function learned from the remaining folds.

2.1.5 Formulation with dimension reduction

Density ratio estimation is hard in high dimensions, and the aforementioned methods have been adapted to project X into a lower dimensional subspace (spanned by the columns of) $U \in \mathbb{R}^{D \times d}$, and estimate the ratio of the projected distributions. The key is then to decide what U should be.

To elaborate, let $V = U^{\perp}$, and let (overloading notation) $U(\cdot), V(\cdot)$ denote projection functions onto the subspaces U, V. Note that one can write x uniquely as $x = U(x) + V(x)$. $P_X^A(\cdot)$ then can be written as $P_X^A(x) =$

$P^A_{U(X)}(U(x))P^A_{V(X)|U(X)}(V(x))$, and similarly for $P^B_X(\cdot)$. Under the assumption

$$P^A_{V(X)|U(X)}(\cdot) = P^B_{V(X)|U(X)}(\cdot), \text{ then} \tag{17}$$

$$w(x) = \frac{P^B_{U(X)}(U(x))P^B_{V(X)|U(X)}(V(x))}{P^A_{U(X)}(U(x))P^A_{V(X)|U(X)}(V(x))} = \frac{P^B_{U(X)}(U(x))}{P^A_{U(X)}(U(x))} := w_{U(\cdot)}(U(x)), \tag{18}$$

where $w_{U(\cdot)}(\cdot)$ is defined to be the ratio of the projected densities.

The goal is then to find the (smallest) subspace U such that $P^A_{U(X)}(\cdot) \neq P^B_{U(\cdot)}(\cdot)$. In practice, d is fixed beforehand, and a search for the d-dimensional subspace U maximizing some dissimilarity measure between $P^A_{U(\cdot)}(\cdot)$ and $P^B_{U(\cdot)}(\cdot)$ is sought. One approach is to view samples from the 2 distributions as belonging to different classes, and apply linear discriminant analysis to find a subspace maximizing the ratio of total between-class distance over total within-class distances. Another disimilarity is Pearson Divergence (PD):

$$PD(P^A_{U(X)}(\cdot), P^B_{U(X)}(\cdot)) = E_{U(X)\sim P^A_{U(X)}(\cdot)}\left[(\frac{P^B_{U(X)}(U(X))}{P^A_{U(X)}(U(X))} - 1)^2\right] \tag{19}$$

2.2 Kernel Mean Matching

2.2.1 Formulation

KMM [2] is motivated by the following theorem: let $\phi : \mathcal{X} \to \mathcal{F}$ be a feature map on covariate $x \in \mathcal{X}$, such that the induced kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is "universal". Given distributions $P^X_A(\cdot), P^X_B(\cdot)$, the solution to the following optimization problem is the function $w(\cdot) : x \to \frac{P^B_X(x)}{P^A_X(x)}$:

$$\min_{w(\cdot)} |E_{X\sim P^B_X(\cdot)}[\phi(X)] - E_{X\sim P^A_X(\cdot)}[w(X)\phi(X)]| \text{ subject to} \tag{20}$$

$$w(x) \geq 0 \tag{21}$$

$$E_{X\sim P^A_X(\cdot)}[w(x)] = 1 \tag{22}$$

In practice, we have samples $x^A_i \sim P^A_X(\cdot)$, $x^B_i \sim P^B_X(\cdot)$, and minimize empirical loss to get estimate $\hat{w}(\cdot)$:

$$\min_{B(\cdot)} |\tfrac{1}{N^B}\sum_i \phi(x^B_i) - \tfrac{1}{N^A}\sum_i w(x^A_i)\phi(x^A_i)| \text{ subject to} \tag{23}$$

$$w(x^A_i) \in [0, W_{\max}] \tag{24}$$

$$|\tfrac{1}{N^A}\sum_i w(x^A_i)]| \leq 1 - \epsilon \tag{25}$$

One can provide probabilistic bounds on the suboptimality of estimated $\hat{w}(\cdot)$ under the true loss of Equation 20. These bounds depend on upper bound $W_{\max}$ such that $w(x) \leq W_{\max}$, which is unknown, but assumed in the optimization. Furthermore, for the true $w(\cdot)$, the empirical expectation $|\frac{1}{N^A}\sum_i w(x^A_i)]|$ will differ from its true expectation of 1, and so in trying to find that true $w(\cdot)$, we allow the empirical expectation some deviation error.

2.2.2 Drawbacks

Although the optimization of Equation 23 is technically over the space of ratio functions, returning an estimated $\hat{w}(\cdot)$, we only actually have access to the ratio for the samples from $P^A_X(\cdot)$: $\{\hat{w}(x^A_i)\}$. Thus the ratio learning procedure is not *inductive*, and we cannot estimate ratios for a separate validation set, and thus do not have a way to evaluate out-of-sample ratio estimation performance. This means there is no clear way to perform cross validation to choose the hyperparameters of the ratio learning procedure: $\phi(\cdot), W_{\max}, \epsilon$. Furthermore, even if out-of-sample ratio estimates are available, the difference in feature means objective function is not directly minimizing accuracy of ratios, so perhaps "performance" on a validation would not be measuring the right thing anyways. Though, there are bounds showing that achieving small difference in (weighted) feature means (Equation 23) implies the loss on the reweighted training set is close with high probability to the actual loss on the test set.

2.3 Classifier-based methods

2.3.1 Standalone Approaches

With classifier methods, we assume a joint distribution over Z, X. Z is Bernoulli - if $Z = 1$, X is drawn from $P_X^B(\cdot)$. Otherwise, X is drawn from $P_X^A(\cdot)$. Once $P_{Z,X}(\cdot)$ is learned,

$$w(x) = \frac{P_X^B(x)}{P_X^A(x)} = \frac{P_{X|Z}(x|z=1)}{P_{X|Z}(x|z=0)} = \frac{\frac{P_{Z|X}(z=1|x)P_X(x)}{P_Z(z=1)}}{\frac{P_{Z|X}(z=0|x)P_X(x)}{P_Z(z=0)}} = \frac{P_{Z|X}(z=1|x)}{P_{Z|X}(z=0|x)}\frac{P_Z(z=0)}{P_Z(z=1)} \tag{26}$$

$P_Z(\cdot)$ is estimated simply as $P_Z(z=1) = \frac{N^B}{N^A+N^B}$. $P_{Z|X}(\cdot)$ can be learned using *any* classifier, by labeling samples with $Z = 1$ if they came from $P_X^B(\cdot)$, else $Z = 0$. This approach is equivalent to propensity scores.

2.3.2 Joint Approaches

So far, the task of learning a predictive model for the test distribution $P_X^B(\cdot)$ has proceeded in 2 steps:

1. Given training and test samples $\{x_i^A\}, \{x_i^B\}$, estimate the weight ratio function:
   find $w^*(\cdot) = \text{argmin}_{w(\cdot)} L_w(w(\cdot); \{x_i^A\}, \{x_i^B\})$, where L_w is some loss function measuring how well weight function $w(\cdot)$ estimates the ratios, for example Equation 5.
2. Learn a predictive model $f^*(\cdot)$ minimizing (estimated) test-set loss:
   find $f^*(\cdot) = \text{argmin}_{f(\cdot)} \sum_i w^*(x_i^A) L_f(f(x_i^A), y_i^a)$, using training set covariates x_i^A and labels y_i^A, where L_f is per-sample predictive loss, e.g. squared or logistic loss.

[1] propose to solve these 2 optimization problems jointly: that is, find:

$$\text{argmin}_{w(\cdot),f(\cdot)} L_w(w(\cdot); \{x_i^A\}, \{x_i^B\}) + \sum_i w(x_i^A) L_f(f(x_i^A), y_i^a) \tag{27}$$

They assume the predictive task to be classification, and assume $f(\cdot)$ to be a logistic regression classifier, with $L_f(\cdot,\cdot)$ thus being logistic loss. They assume $w(\cdot)$ to be of the form of Equation 26, with $P_{Z|X}(\cdot)$ modelled by another logistic regression. $w(\cdot)$ appears in both terms, so that the joint optimization is different from the sequential optimization. $w(\cdot)$ is the information shared between the 2 tasks.

There are some drawbacks: their method is only applicable when the prediction task is classification, due to the assumption of $f(\cdot)$ being a logistic regression. Secondly, it may be sufficient to estimate $w_{U(\cdot)}(\cdot)$ for some subspace/projection U, for example, if $Y \perp X|U(X)$ where $X, Y \sim P_{X,Y}^B(\cdot)$. Intuitively, the prediction task (which requires weight estimates) contains information of what U might be, which can then be used to refine the weight estimates that the prediction task depended on in the first place. Thus, U should be shared between the tasks.

3 Past Dependency Measures

This work will perform ratio estimation of projected densities into a subspace U that is "useful" for prediction as measured by some dependence measure between $U(X)$ and Y. We review some candidate measures.

3.1 f-Divergence Based Measures

Given a joint distribution $P_{X,Y}(\cdot,\cdot)$ this class of dependency measure quantifies the dependence between X and Y to be the f-divergence between $P_{X,Y}(\cdot,\cdot)$ and $P_X(\cdot)P_Y(\cdot)$, where P_X, P_Y are the marginals of $P_{X,Y}$.

3.1.1 Definitions

Given convex function $f : \mathbb{R} \to \mathbb{R}$, distributions $p(\cdot), q(\cdot)$, the *f-divergence* between p, q is defined to be:

$$D_f(p||q) = \int_x f(\tfrac{p(x)}{q(x)}) p(x) dx \tag{28}$$

The f-divergence generalizes the KL-divergence. Indeed, suppose $f = KL$ where

$$KL(u) = \begin{cases} -\log(u) \text{ if } u > 0 \\ +\infty \text{ otherwise.} \end{cases} \tag{29}$$

Then

$$D_{KL}(p||q) = E_{X\sim p}[-\log \tfrac{p(X)}{q(X)}]. \tag{30}$$

Given f, the corresponding *f-information* between X and Y in the joint distribution P_{XY} is defined to be:

$$I_f(X,Y) = D_f(P_X P_Y || P_{XY}) \tag{31}$$

where P_X, P_Y are the marginals of P_{XY}. Note that $I_{KL}(X,Y)$ is simply the mutual information between X and Y, and that all the below statements about $D_f(p||q)$ hold for $I_f(X,Y)$, with $p \leftarrow P_X P_Y$ and $q \leftarrow P_{XY}$.

3.1.2 Variational Characterization of f-Divergence

A variational characterization of $D_f(p||q)$ motivates optimization based lower bound approximations:

$$D_f(p||q) = \sup_g E_{X\sim q}[g(X)] - E_{X\sim p}[f^*(g(X))] \tag{32}$$

where f^* denotes the convex conjugate of f and the sup is over all functions $g : \mathcal{X} \to \mathbb{R}$.

Furthermore, it can shown that [4] if g attains the sup, then

$$g(x) \in \delta f(\tfrac{p(x)}{q(x)}) \text{ for all } x. \tag{33}$$

For example, in the case of KL-divergence, $f^*(v) = -1 - \log(-v)$ for $v < 0$ and $+\infty$ for $v \geq 0$. Then,

$$D_{KL}(p||q) = \sup_{g<0} E_{X\sim q}[g(X)] - E_{X\sim p}[-1 - \log(-g(X))] \tag{34}$$

$$= \sup_{g>0} E_{X\sim p}[\log g(X)] - E_{X\sim q}[g(X)] + 1 \tag{35}$$

$$g_{KL}(x) = \tfrac{q(x)}{p(x)} \tag{36}$$

Thus, evaluating $D_f(p||q)$ via the variational approach of Equation 32 gives not only $D_f(p||q)$, but also a function with a 1-to-1 correspondence with density ratio function $\frac{p(x)}{q(x)}$ (the reciprocal function $\frac{q(x)}{p(x)}$, in the case of KL-divergence). This is like how in the variational characterization of a Bayesian model's evidence, we get not only (a lower bound on) the evidence, but also the posterior density function over latent variables.

3.1.3 Estimation of f-Divergences

Given N^p samples $x_i^p \sim p$, N^q samples $x_i^q \sim q$, the empirical estimation of $D_f(p||q)$ solves the optimization of Equation 32 but with 2 approximations: empirical expectations replace true expectations, and the sup is restricted to be over a chosen function class $\mathcal{G}$, by choosing a set of K basis functions $\{\phi_k(\cdot)\}$ and letting

$$\mathcal{G} = \{g(\cdot) : x \to \langle \alpha, \Phi(x)\rangle; \alpha \in \mathbb{R}^K\} \tag{37}$$

where $\Phi(x) = (\phi_1(x), \ldots, \phi_K(x))$. For example, assuming $\phi_k(\cdot) > 0$, estimation of $D_{KL}(p||q)$, combining Equation 35 and 37, becomes:

$$\tilde{D}_{KL}(p||q) = \max_{\alpha\geq 0} \tfrac{1}{N^p} \sum_i \log\langle \alpha, \Phi(x_i^p)\rangle - \tfrac{1}{N^q} \sum_i \langle \alpha, \Phi(x_i^q)\rangle + 1, \text{ with, based on Equation 36} \tag{38}$$

$$\hat{g}_{KL}(x) = \langle \hat{\alpha}, \Phi(x)\rangle \approx \tfrac{p(x)}{q(x)} \tag{39}$$

being an estimate of density ratio function $\frac{p(x)}{q(x)}$, where $\hat{\alpha}$ is the argmax of Equation 38.

Another interesting choice of f due to the tractability of calculating the corresponding D_f is when $f = SQ$, where

$$SQ(u) = \begin{cases} \frac{1}{2}(u-1)^2 \text{ if } u > 0 \\ +\infty \text{ otherwise.} \end{cases} \tag{40}$$

D_{SQ} is the *Pearson Divergence*. Furthermore, letting $\overline{SQ}(u) = \frac{1}{2}u^2$ if $u > 0$, $+\infty$ otherwise, the following identity follows from Equation 32:

$$D_{SQ}(p||q) = D_{\overline{SQ}}(p||q) - \tfrac{1}{2} \tag{41}$$

Thus to estimate $D_{SQ}(p||q)$, it suffices to estimate $D_{\overline{SQ}}(p||q)$, which is computationally more tractable. In this case, $f^*(v) = \overline{SQ}^*(v) = \frac{1}{2}v^2$ for $v \geq 0$ and 0 for $v < 0$. Then, according to Equation 32 and 33:

$$D_{\overline{SQ}}(p||q) = \sup_g E_{X\sim q}[g(X)] - E_{X\sim p}[\max(\tfrac{1}{2}g(X)^2, 0)] \tag{42}$$

$$= \sup_{g\geq 0} E_{X\sim q}[g(X)] - E_{X\sim p}[\tfrac{1}{2}g(X)^2] \tag{43}$$

$$g_{\overline{SQ}}(x) = \tfrac{p(x)}{q(x)} \tag{44}$$

Assuming $\mathcal{G}$ to be as in Equation 37, the empirical optimization of Equation 43 can be written as a constrained quadratic program:

$$\hat{D}_{\overline{SQ}}(p||q) = \max_{\alpha\geq 0} -\tfrac{1}{2}\alpha^T\Big(\tfrac{1}{N^p}\sum_i \Phi(x_i^p)\Phi(x_i^p)^T\Big)\alpha + \Big(\tfrac{1}{N^q}\sum_i \Phi(x_i^q)\Big)^T\alpha, \text{ with, based on Equation 44} \tag{45}$$

$$\hat{g}_{\overline{SQ}}(x) = \langle\hat{\alpha}, \Phi(x)\rangle \approx \tfrac{p(x)}{q(x)} \tag{46}$$

3.1.4 Estimation of Dependency Measures based on f-Divergences

Given N samples $\{x_i, y_i\}$ drawn iid from P_{XY}, we can regard $\{(x_i, y_j) : 1 \leq i, j \leq N\}$ to be N^2 samples drawn iid from $P_X P_Y$. To perform empirical estimation of $I_f(X, Y)$, like before we define $\mathcal{G}$, a class of functions whose domain is $\mathcal{X} \times \mathcal{Y}$, as this is the support of P_{XY} and $P_X P_Y$, specifying K basis functions $\{\phi_k(\cdot)\}$ and letting:

$$\mathcal{G} = \{g(\cdot) : (x, y) \to \langle\alpha, \Phi(x, y)\rangle; \alpha \in \mathbb{R}^K\} \tag{47}$$

where $\Phi(x, y) = (\phi_1(x, y), \ldots, \phi_K(x, y))$. Then, empirical estimation of $I_f(X, Y)$ becomes:

$$\hat{I}_f(X, Y) = \sup_{g\in\mathcal{G}} \tfrac{1}{N}\sum_i g(x_i, y_i) - \tfrac{1}{N^2}\sum_{i,j} f^*(g(x_i, y_j)) \tag{48}$$

For example, the empirical estimation of $I_f(X, Y)$ for $f = SQ$ becomes (using the identity of Equation 41):

$$\hat{I}_{SQ}(X, Y) = \hat{I}_{\overline{SQ}}(X, Y) - \tfrac{1}{2}, \text{ where} \tag{49}$$

$$\hat{I}_{\overline{SQ}}(X, Y) = \max_{\alpha\geq 0} -\tfrac{1}{2}\alpha^T\Big(\tfrac{1}{N^2}\sum_{i,j} \Phi(x_i, y_j)\Phi(x_i, y_j)^T\Big)\alpha + \Big(\tfrac{1}{N}\sum_i \Phi(x_i, y_i)\Big)^T\alpha, \text{ with} \tag{50}$$

$$\hat{g}_{I_{\overline{SQ}}}(x, y) = \langle\hat{\alpha}, \Phi(x, y)\rangle \approx \tfrac{P_{XY}(x,y)}{P_X(x)P_Y(x)} \tag{51}$$

Note that I_{SQ} is the *squared mutual information* defined in [5].

4 Proposed Work

I propose to perform density ratio estimation following the dimension reduction approach of Section 2.1.5, estimating density ratios of the projection of the test and training covariate distributions into some subspace U. Unlike past work, we consider a *supervised* setting, where we assume the end goal is to perform instance re-weighting for domain adaptation of some supervised task. Thus, we will choose U to maximize some general measure of predictive "utility" for the test distribution that can be computed using projected test samples $\{U(x_i^B), y_i\}$. The justification is that if U has high predictive utility for the test distribution, then one can project the features onto U before learning a predictor, following the more general approach of dimensionality reduction to improve robustness. In this situation, instance re-weighting using Equation 1 using the ratio of the *projected* densities is sufficient for obtaining an unbiased estimate of test loss. The benefit of dimension reduction in this pipeline is then two-fold: to improve the robustness of both the learning method and the ratio estimates. In fact, we can even rebrand the primary goal to be supervised dimension reduction under covariate shift.

The problem of course that we cannot directly compute the predictive utility measure of U for the test distribution based on projected test samples $\{U(x_i^B), y_i^B\}$, as we lack test labels. Though, we can obtain an estimate that will require weight ratios that under the dimension reduction approach, require U to begin with. Thus, finding U and estimating the ratios of the densities projected onto U *cannot* be performed sequentially. Instead, U and the projected ratios must be found jointly so that the projected ratios are well estimated, and the predictive test utility of U (relying on the projected ratios) is also high.

Certainly, the predictive utility of a projection is affected by the change in marginal covariate distribution that occurs in covariate shift, so that the projection with the highest predictive utility for the training distribution is not necessarily that for the test distribution. This is why a naive approach of performing supervised dimension reduction for the training data followed by instance reweighting using projected ratios and (possibly cost-sensitive) training to learn a predictor cannot be expected to give low test loss. We illustrate with an example.

4.1 Proposed Work

For this project, I propose to do supervised dimension reduction under covariate shift. That is, given sample features and labels from the training distribution, and sample features (but not labels) from the test distribution, find a lower dimensional subspace U (of dimension d which is fixed beforehand) of the features that has high "predictive" utility in the *test* distribution. The measure of utility of U we consider will be the mutual information between the projected features and labels for the test distribution, which can be calculated using projected test samples $\{U(x_i^B), y_i^B\}$. The problem of course is that we do not have these test labels. Past work [5] solved this problem without covariate shift - given U_{old}, they construct a local estimator f of the utility of U: $f(U; U_{\text{old}}) \approx MI(U(X), Y)$ which is only accurate for U close to U_{old}. They let $U_{\text{old}} \leftarrow \text{argmax}_U f(U; U_{\text{old}})$ and iterate.

Constructing this estimator $f(U; U_{\text{old}})$ required solving an optimization problem whose objective contained an expectation over the training data. (Just estimating the mutual information at U_{old}, which $f(U_{\text{old}}; U_{\text{old}})$ should equal, requires solving Equation 55, which has an expectation. To modify their technique for when test labels are not available, all we need to do is to modify the estimator, estimating expectations over test distribution P^B with importance weighted expectations using samples from P^A. This change is quite minor. Thus, for this project, I also propose to figure out how to make U row-sparse, so the projected features only involve a small subset of the features. I'm not sure how to do this, or whether d should be fixed beforehand. However, the actual problem of supervised dimension reduction under covariate shift has not been addressed before.

4.2 Formulation

We assume there are 2 unknown distributions $P^A_{X,Y}$ and $P^B_{X,Y}$ related via the covariate shift assumption, that

$$P^A_{X,Y}(\cdot) = P^A_X(\cdot) P_{Y|X}(\cdot) \tag{52}$$

$$P^B_{X,Y}(\cdot) = P^B_X(\cdot) P_{Y|X}(\cdot) \tag{53}$$

so that the conditional distribution $P_{Y|X}$ is identical in $P^A_{X,Y}$ and $P^B_{X,Y}$.

We are given N^A samples $\{x_i^A, y_i^A\}$ drawn iid from $P^A_{X,Y}$ and N^B samples $\{x_i^B\}$ drawn iid from P^B_X. The goal is to perform sufficient dimension reduction under covariate shift - to find:

$$\hat{U} = \text{argmax}_{U(\cdot)} I_{SQ}(U(X^B), Y^B) = \text{argmax}_{U(\cdot)} I_{\overline{SQ}}(U(X^B), Y^B) - \tfrac{1}{2} \tag{54}$$

where

$$I_{\overline{SQ}}(U(X^B), Y^B) = \sup_{g \geq 0} E_{U(X),Y \sim P^B_{U(X),Y}}[g(U(X), Y)] - E_{U(X),Y \sim P^B_{U(X)} P^B_Y}[\tfrac{1}{2} g(U(X), Y)^2] \tag{55}$$

$$= \sup_{g \geq 0} -\tfrac{1}{2} E_{U(X),Y \sim P^A_{U(X)} P^A_Y}[w_{U(X)}(U(X)) w_Y(Y) g(U(X), Y)^2] + E_{U(X),Y \sim P^A_{U(X),Y)}}[w_{U(X),Y}(U(X), Y) g(U(X), Y)] \tag{56}$$

where

$$w_{U(X),Y}(x, y) := \frac{P^B_{U(X),Y}(x,y)}{P^A_{U(X),Y}(x,y)} = \frac{P^B_{U(X)}(x) P_{Y|U(X)}(y)}{P^A_{U(X)}(x) P_{Y|U(X)}(y)} = \frac{P^B_{U(X)}(x)}{P^A_{U(X)}(x)} := w_{U(X)}(x) \tag{57}$$

$$w_Y(y) := \frac{P^B_Y(y)}{P^A_Y(y)} \tag{58}$$

Note that all expectations are now over $P^A_{U(X),Y}, P^A_{U(X)}, P^A_Y$, which are the distributions we actually have empirical distributions for. Making use of Equation 57 and assuming the function class $\mathcal{G}$ of Equation 47, we can write down the empirical estimation version of Equation 56:

$$\hat{I}_{\overline{SQ}}(U(X^B), Y^B) = \max_{\alpha \geq 0} \Big(-\tfrac{1}{2} \alpha^T \Big(\sum_i w_{U(X)}(U(x_i^A)) w_Y(y_i^A) \Phi(U(x_i^A), y_i^A) \Phi(U(x_i^A), y_i^A)^T \Big) \alpha \tag{59}$$

$$+ \Big(\sum_i w_{U(X)}(U(x_i^A)) \Phi(U(x_i^A), y_i^A) \Big)^T \alpha \Big) \tag{60}$$

This expression requires the density ratio functions $w_{U(X)}(\cdot)$ and $w_Y(\cdot)$. Given U, $w_{U(X)}(\cdot)$ can be estimated using samples $\{U(x_i^A)\}, \{U(x_i^B)\}$ using Equation 45, where $p \leftarrow P^B_{U(X)}, q \leftarrow P^A_{U(X)}$. Estimating $w_Y(\cdot)$ requires a modification of the same technique, as we have samples $\{y_i^A\}$, but not $\{y_i^B\}$. Note that the maximizing $g_{\overline{SQ}}(\cdot)$ of Equation 43 with $p \leftarrow P^B_Y, q \leftarrow P^A_Y$ will be such that $g_{\overline{SQ}}(y) = \frac{P^B_Y(y)}{P^A_Y(y)} = w_Y(y)$, and so we need to empirically estimate

$$D_{\overline{SQ}}(P^B_Y, P^B_X) = \sup_{g \geq 0} E_{Y \sim P^A_Y}[g(Y)] - E_{Y \sim P^B_Y}[\tfrac{1}{2} g(Y)^2] \tag{61}$$

With samples $\{y_i^A\}$, a simple average suffices to estimate $E_{Y \sim P^A_Y}[g(Y)] \approx \frac{1}{N^A} \sum_i g(y_i^A)$, where I now use $\approx$ to mean "is an unbiased estimate of". For the other term:

$$\begin{aligned}
E_{Y \sim P^B_Y}[\tfrac{1}{2} g(Y)^2] &= E_{X \sim P^B_{U(X)}}[E_{Y \sim P^B_{Y|U(X)}}[\tfrac{1}{2} g(Y)^2]] && (62)\\
&= E_{X \sim P^B_{U(X)}}[E_{Y \sim P^A_{Y|U(X)}}[\tfrac{1}{2} g(Y)^2]] && (63)\\
&= E_{X \sim P^A_{U(X)}}[w_{U(X)}(X) E_{Y \sim P^A_{Y|U(X)}} \tfrac{1}{2} g(Y)^2]] && (64)\\
&= E_{X,Y \sim P^A_{U(X),Y}}[w_{U(X)}(X) \tfrac{1}{2} g(Y)^2] && (65)\\
&\approx \tfrac{1}{N^A} \sum_i w_{U(X)}(x_i^A) \tfrac{1}{2} g(y_i^A)^2 && (66)
\end{aligned}$$

Thus, the estimation of $D_{\overline{SQ}}(P^B_Y, P^A_Y)$ that gives us an estimate of $w_Y(\cdot)$ is:

$$\hat{D}_{\overline{SQ}} = \max_{\alpha \geq 0} -\tfrac{1}{2} \alpha^T \Big(\tfrac{1}{N^A} \sum_i w_{U(X)}(x_i^A) \Phi(y_i^A) \Phi(y_i^A)^T \Big) \alpha + \Big(\tfrac{1}{N^A} \sum_i \Phi(y_i^A) \Big)^T \alpha \text{ with} \tag{67}$$

$$\hat{w}_Y(y) = \hat{g}_{\overline{SQ}}(y) = \hat{\alpha}^T \Phi(y) \tag{68}$$

In summary, the grand optimization objective of Equation 54 is over U, but the objective function itself is another optimization problem that depends on U. This suggests initializing U and repeating the following procedure until convergence:

- Update $w_{U(X)}(\cdot)$ based on U (Equation 43)
- Update $w_Y(\cdot)$ based on $w_{U(X)}(\cdot)$ (Equation 54)
- Update U, viewing $\hat{I}_{\overline{SQ}}(U(X^B), Y^B)$ a function of U, with $\alpha, W_{U(X)}, W_Y$ fixed.

4.3 Generic Formulation Based on Predictive Loss

Note: I will use U to denote the subspace spanned by rows of $U \in \mathbb{R}^{D \times P}$, and also use $U(\cdot)$ as a function from $\mathbb{R}^P \to \mathbb{R}^D$ that projects data to the subspace.

Input:

- subspace dimension D
- function class $\mathcal{F}$ consisting of functions $f : \mathbb{R}^D \to \mathbb{R}$
- training covariates and labels $\{x_i^A, y_i^A\}$ drawn from $P_{X,Y}^A$
- test covariates $\{x_i^B\}$ drawn from P_X^B, with $x_i^A, x_i^B \in \mathbb{R}^P$
- prediction loss function $L(\cdot, \cdot)$
- black box weight ratio function parameterized by U that uses training/test covariate samples $\{x_i^A\}, \{x_i^B\}$: $w(\cdot; U) := \mathbb{R}^P \to \mathbb{R} : x \mapsto \frac{P_{U(X)}^B(U(x))}{P_{U(X)}^A(U(x))}$

Output: subspace $U = \text{argmax}_U \min_{f \in \mathcal{F}} \sum_i w(x_i^A) L(f(x_i^A), y_i^A)$. U is the most useful subspace for prediction, the subspace within which if all covariates were projected, achievable loss is the lowest.

Iterative algorithm:

1. $w_{\text{current}}(\cdot) \leftarrow w(\cdot; U_{\text{current}})$
2. $U_{\text{current}} \leftarrow \text{argmax}_U \min_{f \in \mathcal{F}} \sum_i w_{\text{current}}(x_i^A) L(f(U(x_i^A)), y_i^A)$

Problem: suppose $\mathcal{F}$ is the set of linear functions. Then there are many U in the argmax - those that the optimal regression coefficient $B^* \in \mathbb{R}^P$ (in the original feature space) lies within. In other words, the argmax is any U contained in a $P - 1$ dimensional subspace.

References

[1] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM, 2007.

[2] 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*, pages 601–608, 2006.

[3] Takafumi Kanamori. Density ratio estimation: A comprehensive review. , 1703:10–31, 2010.

[4] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization.

[5] Taiji Suzuki and Masashi Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural computation*, 25(3):725–758, 2013.