A Proofs

Proof of Theorem 1. With strong duality holds,

$$\begin{aligned}
&\min_{\hat{f}(y|\mathbf{x})\in\Delta}\ \max_{f(y|\mathbf{x})\in\Delta\cap\Xi}\ \text{rel-loss}_{f_{\text{trg}}(\mathbf{x})}\left(f(Y|X),\hat{f}(Y|X),f_0(Y|X)\right)\\
&=\max_{f(y|\mathbf{x})\in\Delta\cap\Xi}\ \min_{\hat{f}(y|\mathbf{x})\in\Delta}\ \text{rel-loss}_{f_{\text{trg}}(\mathbf{x})}\left(f(Y|X),\hat{f}(Y|X),f_0(Y|X)\right)\\
&\left(\text{Given } f(y|\mathbf{x}),\ \text{the minimum of relative logloss is achived when } \hat{f}(y|\mathbf{x})=f(y|\mathbf{x})\right)\\
&=\max_{f(y|\mathbf{x})\in\Delta\cap\Xi}\ \text{rel-loss}_{f_{\text{trg}}(\mathbf{x})}\left(f(Y|X),f(Y|X),f_0(Y|X)\right)\\
&=\min_{f(y|\mathbf{x})\in\Delta\cap\Xi} D_{f_{\text{trg}}(\mathbf{x}),f(y|\mathbf{x})}\left(f(Y|X)||f_o(Y|X)\right)\\
&=\min_{\hat{f}(y|\mathbf{x})\in\Delta\cap\Xi} D_{f_{\text{trg}}(\mathbf{x}),\hat{f}(y|\mathbf{x})}\left(\hat{f}(Y|X)||f_o(Y|X)\right)
\end{aligned}$$

□

Proof of Theorem 2. Let

$$\hat{f}_\theta(y|\mathbf{x})=\frac{f_o(y|\mathbf{x})e^{-\frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}\theta^T\Phi(\mathbf{x},y)}}{Z(\mathbf{x},\theta)},$$

where $Z(\mathbf{x},\theta)$ is the normalization term which guarantees $\hat{f}_\theta(y|\mathbf{x})$ is a valid conditional probability, and θ is chosen so that $\hat{f}_\theta(y|\mathbf{x})$ satisfies statistic constraint. Formally, $\hat{f}_\theta(y|\mathbf{x})\in\Delta\cap\Xi$. Then $\hat{f}_\theta(y|\mathbf{x})$ uniquely minimizes the Kullback-Leibler divergence over all conditional probabilities $g(y|\mathbf{x})\in\Delta\cap\Xi$.

To simplify the proof let,

$$\phi(\mathbf{x},y)=-\frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}\theta^T\Phi(\mathbf{x},y)-\log Z(\mathbf{x},\theta),$$

and then,

$$\begin{aligned}
&D(g(Y|X)||f_o(Y|X))\\
&=\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\log\frac{g(y|\mathbf{x})}{f_o(y|\mathbf{x})}\\
&=\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\log\frac{g(y|\mathbf{x})}{\hat{f}(y|\mathbf{x})}\hat{f}(y|\mathbf{x})-\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\log f_o(y|\mathbf{x})\\
&=D(g(Y|X)||\hat{f}_\theta(Y|X))+\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\log\hat{f}(y|\mathbf{x})-\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\log f_o(y|\mathbf{x})\\
&\geq\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\log\hat{f}(y|\mathbf{x})-\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\log f_o(y|\mathbf{x})\\
&=\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})(\log f_o(y|\mathbf{x})+\phi(\mathbf{x},y))-\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\log f_o(y|\mathbf{x})\\
&=\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X} g(y|\mathbf{x})\phi(\mathbf{x},y)\ \left(\text{Both } g(y|\mathbf{x}) \text{ and } \hat{f}(y|\mathbf{x})\in\Delta\cap\Xi\right)\\
&=\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X}\hat{f}(y|\mathbf{x})\phi(\mathbf{x},y)\\
&=\int_X f_{\text{trg}}(\mathbf{x})\int_{Y|X}\hat{f}(y|\mathbf{x})\log\frac{\hat{f}(y|\mathbf{x})}{f_o(y|\mathbf{x})}\\
&=D(\hat{f}_\theta(Y|X)||f_o(Y|X))
\end{aligned}$$

The inequality holds because the Kullback-Leibler divergence is always non-negative with zero if and only if $g(y|\mathbf{x})=\hat{f}_\theta(y|\mathbf{x})$ almost everywhere, thus proving uniqueness.

The non-negative constraint of (9) is superfluous since the log function in the objective function requires non-negative real numbers. Hence, the Lagrangian of the optimization problem is:

$$\begin{aligned}
\mathcal{L}(\theta) &= D_{f_{\text{trg}}(\mathbf{x}),\hat{f}(y|\mathbf{x})}\left(\hat{f}(Y|X)||f_o(Y|X)\right) + \theta^T(\mathbb{E}_{f_{\text{src}}(\mathbf{x})f(y|\mathbf{x})}[\Phi(X,Y)] - c) \\
&= \int_X f_{\text{trg}}(\mathbf{x}) \int_{Y|X} \hat{f}(y|\mathbf{x}) \log \frac{\hat{f}(y|\mathbf{x})}{f_o(y|\mathbf{x})} + \theta^T(\mathbb{E}_{f_{\text{src}}(\mathbf{x})f(y|\mathbf{x})}[\Phi(X,Y)] - c) \\
&= \int_X f_{\text{trg}}(\mathbf{x}) \int_{Y|X} \hat{f}(y|\mathbf{x})(-\frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}\theta^T\Phi(\mathbf{x},y)) - \int_X f_{\text{trg}}(\mathbf{x}) \int_{Y|X} \hat{f}(y|\mathbf{x}) \log Z(\mathbf{x},\theta) + \\
&\qquad \theta^T(\int_X f_{\text{src}}(\mathbf{x}) \int_{Y|X} \hat{f}(y|\mathbf{x})\Phi(\mathbf{x},y) - c) \\
&= -\int_X f_{\text{trg}}(\mathbf{x}) \log Z(\mathbf{x},\theta) - \theta^T c
\end{aligned}$$

Then:

$$\begin{aligned}
\arg\max_\theta \mathcal{L}(\theta) &= \arg\max_\theta \left(-\int_X f_{\text{trg}}(\mathbf{x}) \log Z(\mathbf{x},\theta) - \theta^T c\right) \\
&= \arg\max_\theta \left(-\int_X f_{\text{trg}}(\mathbf{x}) \log Z(\mathbf{x},\theta) - \theta^T \int_X f_{\text{trg}}(\mathbf{x}) \int_{Y|X} f(y|\mathbf{x})\Phi(\mathbf{x},y)\right) \\
&= \arg\max_\theta \left(-\int_X f_{\text{trg}}(\mathbf{x}) \log Z(\mathbf{x},\theta) - \theta^T \int_X f_{\text{trg}}(\mathbf{x}) \frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})} \int_{Y|X} f(y|\mathbf{x})\Phi(\mathbf{x},y)\right) \\
&= \arg\max_\theta \left(\int_X f_{\text{trg}}(\mathbf{x}) \int_{Y|X} f(y|\mathbf{x}) \left(\log \frac{1}{Z(\mathbf{x},\theta)} - \frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}\theta^T\Phi(\mathbf{x},y)\right)\right) \\
&= \arg\max_\theta \mathbb{E}_{f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})}\left[\log \frac{\hat{f}_\theta(Y|X)}{f_o(Y|X)}\right] \\
&= \arg\max_\theta \mathbb{E}_{f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})}\left[\log \hat{f}_\theta(Y|X)\right]
\end{aligned}$$

□

Proof of Theorem 3. Taking the partial derivative of $\mathcal{L}(\theta)$ with respect to θ (by Leibniz's rule for differentiation under the integral sign):

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta)}{\partial \theta} &= -\int_X f_{\text{trg}}(\mathbf{x}) \frac{1}{Z(\mathbf{x},\theta)} \frac{\partial Z(\mathbf{x},\theta)}{\partial \theta} - c \\
&= -\int_X f_{\text{trg}}(\mathbf{x}) \int_{Y|X} \hat{f}_\theta(-\frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}\Phi(\mathbf{x},y)) - c \\
&= \mathbb{E}_{f_{\text{src}}(\mathbf{x})\hat{f}_\theta(y|\mathbf{x})}[\Phi(X,Y)] - c
\end{aligned} \tag{24}$$

□

Proof of Corollary 1. Since

$$f_o(y|\mathbf{x}) \propto e^{-\frac{1}{2}y^T\Sigma_o^{-1}y + y^T\Sigma_o^{-1}\mu_o(\mathbf{x})},$$

Then

$$\begin{aligned}
\hat{f}_\theta(y|\mathbf{x}) &\propto f_o(y|\mathbf{x}) e^{-\frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}\theta^T \begin{bmatrix} y \\ \mathbf{x} \\ 1 \end{bmatrix}^T \mathbf{M} \begin{bmatrix} y \\ \mathbf{x} \\ 1 \end{bmatrix}} \\
&\propto e^{-\frac{1}{2}y^T(2\frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}M_{(y,y)} + \Sigma_o^{-1})y + y^T(-2\frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}M_{(y,\mathbf{x}1)}\begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} + \Sigma_o^{-1}\mu_o(\mathbf{x}))}
\end{aligned}$$

□

Proof of Theorem 4. When the new constraint is applied, the constrained optimization problem (9) becomes:

$$\min_{\hat{f}(Y|X)} D_{f_{\text{trg}}(\mathbf{x}),\hat{f}(y|\mathbf{x})}\left(\hat{f}(Y|X)||f_0(Y|X)\right)$$
$$\text{such that}: \mathbb{E}_{f_{\text{trg}}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(X,Y)] = \tilde{\mathbf{c}}'$$

Following the same procedure of solving the optimization problem in Theorem 2, the solution takes the form $\hat{f}_\theta(y|\mathbf{x}) = f_0(y|\mathbf{x})\frac{e^{-\theta^T\Phi(\mathbf{x},y)}}{Z(\mathbf{x})}$, with $Z(\mathbf{x}) = \int_{y\in\mathcal{Y}} f_0(y|\mathbf{x})e^{-\theta^T\Phi(\mathbf{x},y)}$, which is not the same form as (10). Similar to the proof of Theorem 2, the Lagrangian after plugging in $\hat{f}_\theta(y|\mathbf{x})$ becomes:

$$\begin{aligned}\mathcal{L}(\theta) =& D_{f_{\text{trg}}(\mathbf{x}),\hat{f}(y|\mathbf{x})}\left(\hat{f}(Y|X)||f_0(Y|X)\right) + \theta^T\left(\mathbb{E}_{f_{\text{trg}}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(X,Y)] - \tilde{\mathbf{c}}'\right)\\ =& -\mathbb{E}_{f_{\text{trg}}(\mathbf{x})}[\log Z(X,\theta)] - \theta^T\tilde{\mathbf{c}}'.\end{aligned}$$

Take the gradient with respect to θ, the gradient is

$$\mathbb{E}_{f_{\text{trg}}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(X,Y)] - \tilde{\mathbf{c}}' = \mathbb{E}_{f_{\text{src}}(\mathbf{x})\hat{f}(y|\mathbf{x})}\left[\frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}\Phi(X,Y)\right] - \tilde{\mathbf{c}}',$$

which becomes

$$\mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\hat{f}(y|\mathbf{x})}\left[\frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}\Phi(X,Y)\right] - \tilde{\mathbf{c}}' - \lambda\theta,$$

when constraint slack and dual regularization is applied to allow for the noise from finite sample approximation. We first prove the Lagrangian maximization problem above is equivalent to the reweighted conditional log likelihood maximization problem.

The reweighted conditional log likelihood maximization is:

$$\begin{aligned}ll(\theta) &= \mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\tilde{f}(y|\mathbf{x})}\left[\frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}\log\frac{\hat{f}_\theta(Y|X)}{f_0(Y|X)}\right]\\ &= -\mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})}\left[\frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}\log Z(X,\theta)\right] - \theta^T\tilde{\mathbf{c}}'.\end{aligned}$$

If the same regularization is applied, the gradient with respect to θ is

$$\mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\hat{f}(y|\mathbf{x})}\left[\frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}\Phi(X,Y)\right] - \tilde{\mathbf{c}}' - \lambda\theta,$$

which is the same with the gradient of the Lagrangian maximization problem. Therefore, robust bias-aware regression is equivalent with reweighted conditional log likelihood maximization problem.

Furthermore, we will prove that when the feature function takes the quadratic form as in Corollary 1, reweighted conditional log likelihood maximization problem is equivalent with the importance weighted least square regression.

Since $f_0(y|\mathbf{x})$ is a Gaussian distribution $N(\mu_0,\Sigma_0)$ and feature function takes quadratic form, the resulting distribution $\hat{f}_{\mathbf{M}}(y|\mathbf{x})$ is also a Gaussian, where

$$\begin{aligned}\mu &= (2\mathbf{M}_{(y,y)} + \sigma_0^{-2})^{-1}(-2\mathbf{M}_{(y,\mathbf{x}1)}\begin{bmatrix}\mathbf{x}\\1\end{bmatrix} + \sigma_0^{-2}\mu_0)\\ \sigma^2 &= (2\mathbf{M}_{(y,y)} + \sigma_0^{-2})^{-1}\end{aligned}$$

$\hat{f}_{\mathbf{M}}(y|\mathbf{x})$ can take form $N(a+\mathbf{b}^T\mathbf{x},\sigma^2)$, the reweighted conditional log likelihood maximization problem is represented as:

$$\mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\tilde{f}(y|\mathbf{x})}\left[\frac{1}{2}\frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}\log(2\pi\sigma^2) + \frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}\frac{(Y-a-\mathbf{b}^TX)^2}{2\sigma^2}\right] + \mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\tilde{f}(y|\mathbf{x})}[\frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}\log(f_0(Y|X))]$$

If we consider σ^2 as a constant, minimizing the above function is equivalent with minimizing the reweighted squared loss $\mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\tilde{f}(y|\mathbf{x})}[\frac{f_{\text{trg}}(X)}{f_{\text{src}}(X)}(Y-a-\mathbf{b}^TX)^2]$. □

B Data Description

We apply our approach to four datasets with a natural bias between the source and the target distributions and five datasets with a synthetically created bias as described in Table 1 (Parkinsons dataset has both natural and synthetically created bias settings). All of the datasets are selected from the UCI repository [3].

- **Airfoil** is a National Aeronautics and Space Administration (NASA) dataset, obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel.
- **Concrete** reports the compressive strength of concrete.
- **Housing** contains the housing values in suburbs of Boston.
- **Music** was collected to predict the origin of the music, represented by latitude in our experiment setting.
- **Crime** combines the communities and crime rate information of the cities in different states.
- **Parkinsons** is composed of a range of biomedical voice measurements from 42 people for remote symptom progression monitoring.
- **WineQuality** includes two datasets, related to red and white vinho verde wine samples.
- **IndoorLocation** is a multi-building multi-floor indoor localization database to test indoor positioning system that rely on WLAN/WiFi fingerprint.