Sequential Domain Adaptation by Synthesizing Distributionally Robust Experts

Bahar Taskesen [1] **Man-Chung Yue** [2] **José Blanchet** [3] **Daniel Kuhn** [1] **Viet Anh Nguyen** [3 4]

Abstract

Least squares estimators, when trained on a few target domain samples, may predict poorly. Supervised domain adaptation aims to improve the predictive accuracy by exploiting additional labeled training samples from a source distribution that is close to the target distribution. Given available data, we investigate novel strategies to synthesize a family of least squares estimator experts that are robust with regard to moment conditions. When these moment conditions are specified using Kullback-Leibler or Wasserstein-type divergences, we can find the robust estimators efficiently using convex optimization. We use the Bernstein online aggregation algorithm on the proposed family of robust experts to generate predictions for the sequential stream of target test samples. Numerical experiments on real data show that the robust strategies may outperform non-robust interpolations of the empirical least squares estimators.

1. Introduction

A natural approach to improving predictive performance in data-scarce tasks involves translating informative signals from a data-abundant source domain to the data-scarce target domain. This transfer of knowledge is commonly referred to as domain adaptation or transfer learning, and it is increasingly applied in a wide range of settings, see for example Wilson & Cook (2020); Chu & Wang (2018); Weiss et al. (2016) and Redko et al. (2019).

We consider the supervised domain adaptation setting with scarce labeled target data. The key challenge here is the absence of meaningful data to tune any parameters. However, in many practically relevant applications, new data will arrive sequentially to enrich the information on the target domain. In this case, many online algorithms can be utilized to adaptively learn the best predictor on the target domain, which also guarantee optimal asymptotic regrets (Lattimore & Szepesvári, 2020).

In this paper, we take a pragmatic approach to resolve a specific setup of the domain adaptation problem. We assume access to a scarce labelled target data, and the future target data arrives sequentially. For example, consider understanding the dynamics of ride-sharing platforms requires insights about the demand and supply from both sides of the market. These insights are signalled through the ride fares, which can be explained by characteristics such as the travel distances and the origin-destination pairs of the trips, the time of the day as well as the weather conditions. The capability to correctly predict ride fares directly translates into improved profit forecasts, and thus it vitally supports the growth of new-coming platforms. In a competitive market, a follower (e.g., Lyft) needs to target a slightly different market segment than the leader (e.g., Uber) who had entered earlier. Thus, the demand and supply characteristics for the follower may differ from those of the leader. Nevertheless, as both platforms provide on-demand transportation, it is reasonable to assume that their supply and demand dynamics are similar. The follower, who possesses limited data, can query demand on the leader's platform to collect data in order to leap forward in its predictive precision. Our approach to solve this problem is illustrated in Figure 1 and it consists of two components:

1. **Expert Generation Module:** This module generates a set of competitive experts $\mathcal{E}$ by fine-tuning the explanatory power of the source domain data and harnessing the signal guidance from the scarce target domain data.
2. **Expert Aggregation Module:** Acting on the sequential arrival of the unseen target data, this module aggregates the predictive capability of the generated experts via an online aggregation mechanism. In this work we will use the Bernstein Online Aggregation mechanism.

We will propose two ways to generate the experts. The first

[1]Risk Analytics and Optimization Chair, Ecole Polytechnique Fédérale de Lausanne [2]Department of Applied Mathematics, The Hong Kong Polytechnic University [3]Department of Management Science and Engineering, Stanford University [4]VinAI Research, Vietnam. Correspondence to: Bahar Taskesen <bahar.taskesen@epfl.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

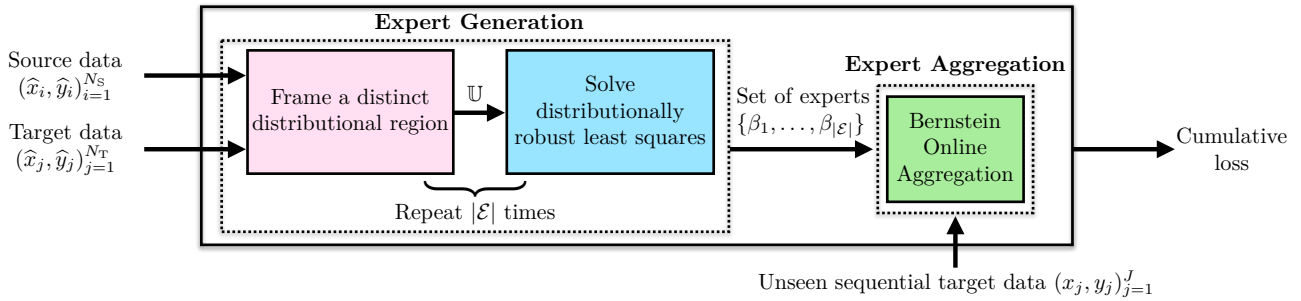


Figure 1. The architecture of our framework for supervised domain adaptation when the unseen target test samples arrive sequentially.

approach generates experts corresponding to optimal decisions along a path, with the intention to interpolate between the source and the target distributions. We will consider two types of trajectories, guided by either the Kullback-Leibler or the Wasserstein divergence. The second approach generates distribution regions around both the source and the target. The intersection of these regions is used to generate distributionally robust experts. The geometrical intuition is to find the "direction" induced by the aforementioned divergences, in which the source data can explain the target data. Once the experts are deployed, the aggregation mechanism is executed without re-adapting the experts.

Our ultimate goal is to ensure a competitive performance in the short term and not in the asymptotic regime when the number of test samples from the target domain tends to infinity. Indeed, as soon as the target sample size is sufficient, training the machine learning model on all available target data becomes more attractive. From a short term horizon benchmark, our approach offers an appealing *warm start* for online training procedure, and it may also lead to a faster convergence rate depending on the underlying algorithm.

Contributions. Our paper explores the expert generation problem in the context of supervised domain adaptation.

- We introduce a novel framework to synthesize a family of robust least squares experts by altering various moment-based distribution sets. These sets gradually interpolate from the source information to the target information, capturing different belief levels on the explanatory power of the source domain onto the target domain.
- We present two intuitive strategies to construct the sets of moment information, namely the "Interpolate, then Robustify" and the "Surround, then Intersect" strategies. Both strategies are simply characterized by two parameters representing the aforementioned explanatory power of belief of the source domain and the level of desired robustness.
- We show that when the moment information is prescribed using a Kullback-Leibler or a Wasserstein-type divergence, the experts are efficiently formed by solving convex optimization problems, that can even be solved by a first-order gradient descent algorithm or off-the-shelf solvers.

This paper is structured as follows. Section 2 delineates the problem setup and describes in details two common strategies to generate experts: the convex combination and the reweighting strategies. Section 3 introduces our framework to generate experts, while Section 4 and 5 dive into details about our "Interpolate, then Robustify" and our "Surround, then Intersect" strategies, respectively. Section 6 demonstrates experimentally that the proposed robust strategies systematically outperform non-robust interpolations of the empirical least squares estimators.

Literature Review. Domain adaptation arises in various applications including natural language processing (Søgaard, 2013; Li, 2012; Jiang & Zhai, 2007; Blitzer et al., 2006), survival analysis (Li et al., 2016) and computer vision (Wang & Deng, 2018; Csurka, 2017). Domain adaptation methods can be classified into three categories. Unsupervised domain adaptation only requires unlabelled target data, but in large amounts (Ghifary et al., 2016; Baktashmotlagh et al., 2013; Ganin & Lempitsky, 2015; Wang et al., 2020; Long et al., 2016; Ben-David et al., 2007; Courty et al., 2017). Semi-supervised domain adaptation requires labelled target data (Yao et al., 2015; Kumar et al., 2010; Sindhwani et al., 2005; Lopez-Paz et al., 2012; Saha et al., 2011; de Mathelin et al., 2020; Sun et al., 2011). Finally, supervised domain adaptation only requires scarce labelled target data (Motiian et al., 2017b;a; Tzeng et al., 2015; Koniusz et al., 2017). If the target data is scarce and label information is available, supervised domain adaptation outperforms unsupervised domain adaptation (Motiian et al., 2017b). The domain adaptation literature further ramifies by imposing different distributional assumptions into covariate shift (Shimodaira, 2000; Sugiyama et al., 2008) or label shift (Lipton et al., 2018; Azizzadenesheli et al., 2019).

The domain adaptation literature for regression problems focuses primarily on instance-based reweighting strategies (Garcke & Vanck, 2014; Sugiyama et al., 2008; Garcke & Vanck, 2014; Huang et al., 2006; Cortes & Mohri, 2014; Chen et al., 2016), which aim to minimize some distance between the source and target distributions. Most of the instance-based methods solve an optimization problem to find the weights of the instances (Garcke & Vanck, 2014; Cortes et al., 2019), which may be computationally expen-

sive when data is abundant. Other approaches rely on deep learning models to minimize the discrepancy between the domain distributions (Zhao et al., 2018; Richard et al., 2020). The literature on regression for domain adaptation also extends towards boosting-based methods (Pardoe & Stone, 2010), and deep learning methods (Salaken et al., 2019).

Our paper also uses ideas and techniques from robust optimization and adversarial training, which have attracted considerable attention in machine learning (Namkoong & Duchi, 2016; Gao et al., 2018; Blanchet et al., 2019; Nguyen et al., 2019a). Robust optimization for least squares problem with uncertain data was studied in Ghaoui & Lebret (1997). Distributionally robust optimization with moment ambiguity sets was proposed in Delage & Ye (2010) and extended in Goh & Sim (2010) and Kuhn et al. (2019). Ambiguity sets prescribed by divergences were previously used to robustify Bayes classification (Nguyen et al., 2019b; 2020).

Our work is also similar to Chen et al. (2016) that consider unsupervised domain adaptation regression, and Wang et al. (2020) that consider robust domain adaption for the classification setting.

Notation. We use I_d to denotes the identity matrix in $\mathbb{R}^d$. The set of p-by-p positive (semi-)definite matrices is denoted by $\mathbb{S}^p_{++}$ ($\mathbb{S}^p_{+}$). All proofs are relegated to the Appendix.

2. Problem Statement and Background

We consider a generic linear regression setting, in which X is a d-dimensional covariate and Y is a univariate response variable. In the context of supervised domain adaptation, we have access to the source domain data $(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$ consisting of N_{S} labelled samples drawn from the source distribution. In addition, we are given a limited number of N_{T} labelled samples $(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$ from the target distribution. Our goal is to predict the responses of the test samples $(x_j, y_j)_{j=1}^{J}$, which are drawn from the target distribution and arrive sequentially. To this end, we will construct several experts.

In the linear regression setting, each expert is characterized by a vector $\beta \in \mathbb{R}^d$. Given a covariate-response pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, we use the square loss function to measure the mismatch between the expert's prediction $\beta^\top x$ and the actual response y. Using the target domain data $(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{T}}}$, one approach is to solve the ridge regression problem

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

for some $\eta \geq 0$ to obtain the empirical target predictor

$$\widehat{\beta}_{\mathrm{T}} = \left(\frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \widehat{x}_j \widehat{x}_j^\top + \eta I_d \right)^{-1} \left(\frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \widehat{x}_j \widehat{y}_j \right).$$

When N_{T} is small, however, the empirical target predictor may perform poorly on the future target data $(x_j, y_j)_{j=1}^{J}$.

If the source domain distribution is sufficiently close to the target domain distribution, it is expedient to exploit the available information in the source domain data to construct better predictors for the target domain data. With this promise, one can synthesize several predictors to form an ensemble of experts, and one can apply an online aggregation scheme to predict on the unseen target data. We now first describe several interpolation schemes to generate experts.

Convex Combination Strategy. Denote by $\widehat{\beta}_{\mathrm{S}}$ the empirical source predictor, which is obtained by solving the ridge regression problem on the source data. The convex combination strategy generates predictors by forming convex combinations between $\widehat{\beta}_{\mathrm{S}}$ and $\widehat{\beta}_{\mathrm{T}}$. More precisely, for any $\lambda \in [0, 1]$ a new predictor is synthesized by setting

$$\widehat{\beta}_\lambda = \lambda \beta_{\mathrm{S}} + (1 - \lambda) \beta_{\mathrm{T}}.$$

The parameter λ represents our *belief* in the explanatory power of the source domain data: if $\lambda = 0$, the source domain has no power to explain the target domain, and we recover $\widehat{\beta}_0 = \beta_{\mathrm{T}}$, the empirical target predictor. If $\lambda = 1$, the source domain has an absolute predictive power on the target domain, and it is beneficial to use $\widehat{\beta}_1 = \widehat{\beta}_{\mathrm{S}}$ because the sample size N_{S} is large. Discretizing λ in the range $[0, 1]$ forms a family of experts $\mathcal{E}$.

Reweighting Strategy. Reweighting samples is a common strategy in domain adaptation, transfer learning and adversarial training. Garcke & Vanck (2014) synthesize experts, for example, by solving

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{N_{\mathrm{S}}} w_{h,i} (\beta^\top \widehat{x}_i - \widehat{y}_i)^2 + \sum_{j=1}^{N_{\mathrm{T}}} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

for some non-negative weights $w_{h,i}$ determined via a Gaussian kernel with bandwidth $h > 0$ of the form

$$w_{h,i} = \sum_{l=1}^{N_{\mathrm{S}}} \alpha_l \exp\left(- \frac{\|\widehat{x}_i - \widehat{x}_l\|_2^2 + (\widehat{y}_i - \widehat{y}_l)^2}{h^2} \right)$$

for $i = 1, \ldots, N_{\mathrm{S}}$. Here, the parameter vector $\alpha \in \mathbb{R}_+^{N_{\mathrm{S}}}$ solves the exponential cone optimization problem

$$\begin{array}{ll} \max & \displaystyle\sum_{j=1}^{N_{\mathrm{T}}} \log\Big(\sum_{l=1}^{N_{\mathrm{S}}} \alpha_l \exp\Big(- \frac{\|\widehat{x}_j - \widehat{x}_l\|_2^2 + (\widehat{y}_j - \widehat{y}_l)^2}{h^2} \Big) \Big) \\ \text{s.t.} & \displaystyle\sum_{i=1}^{N_{\mathrm{S}}} \sum_{l=1}^{N_{\mathrm{S}}} \alpha_l \exp\left(- \frac{\|\widehat{x}_i - \widehat{x}_l\|_2^2 + (\widehat{y}_i - \widehat{y}_l)^2}{h^2} \right) = N_{\mathrm{S}}. \end{array}$$

The predictor β_h, parametrized by the kernel weight h, that solves the reweighted ridge regression problem has the form

$$\Big(\sum_{j=1}^{N_{\mathrm{T}}} \widehat{x}_j \widehat{x}_j^\top + \sum_{i=1}^{N_{\mathrm{S}}} w_i \widehat{x}_i \widehat{x}_i^\top + \eta I_d \Big)^{-1} \Big(\sum_{j=1}^{N_{\mathrm{T}}} \widehat{x}_j \widehat{y}_j + \sum_{i=1}^{N_{\mathrm{S}}} w_i \widehat{x}_i \widehat{y}_i \Big).$$

Discretizing the bandwidth h forms a family of experts $\mathcal{E}$.

Bernstein Online Aggregation (BOA). We now give a brief overview on the BOA algorithm, which is a recursive expert aggregation procedure for sequential prediction (Cesa-Bianchi & Lugosi, 2006). For a given set of experts $\mathcal{E} = \{\beta_1, \ldots, \beta_{|\mathcal{E}|}\}$ and an incumbent weight $\pi_{k,j-1}$ for expert k at time $j-1$, this algorithm aggregates the individual expert's predictions linearly based on the arrival of the input data (x_j, y_j) as $\sum_{k=1}^{|\mathcal{E}|} \pi_{k,j}\beta_k^\top x_j$. The weights of the experts are updated using the exponential rule

$$\pi_{k,j} = \frac{\exp(-\upsilon(1+\upsilon L_{k,j})L_{k,j})\pi_{k,j-1}}{\sum_{k=1}^{|\mathcal{E}|}\exp(-\upsilon(1+\upsilon L_{kj})L_{k,j})\pi_{k,j-1}},$$

where $\upsilon > 0$ is the learning rate and $L_{k,j} = (\beta_k^\top x_j - y_j)^2 - \sum_{k=1}^{|\mathcal{E}|}(\beta_k^\top x_j - y_j)^2\pi_{k,j-1}$. This algorithm is initialized with weights $\pi_{k,0} \geq 0$ satisfying $\sum_{k=1}^{|\mathcal{E}|}\pi_{k,0} = 1$. The cumulative loss for the stream of test data $(x_j, y_j)_{j=1}^J$ is

$$\sum_{j=1}^{J}\left(\sum_{k=1}^{|\mathcal{E}|}\pi_{k,j}\beta_k^\top x_j - y_j\right)^2. \tag{1}$$

For the square loss, the BOA procedure is optimal for the model selection aggregation problem, that is, the excess risk of its batch version achieves the fast rate of convergence $\log(|\mathcal{E}|)/J$ in deviation; see Wintenberger (2017).

3. Predictor Generation via Distributionally Robust Linear Regression

We now specify our framework to generate the set of competitive experts $\mathcal{E}$ for future prediction. Our construction is based on the premises that the source domain carries the explanatory power on the target domain to a certain extent and that the scarce target data can provide directional guidance to pull information from the source data. Moreover, we also leverage ideas from distributionally robust optimization and adversarial training, which have been shown to significantly improve the out-of-sample predictive performance (Duchi & Namkoong, 2018; Mohajerin Esfahani & Kuhn, 2018; Blanchet et al., 2019; Gao, 2020; Lam, 2019).

With this in mind, our expert generation scheme blends two elements: a distributional probing strategy and a robust estimation procedure. The distributional probing strategy frames the distribution set $\mathbb{B}$, and then each expert is constructed by solving a distributionally robust least squares estimation problem of the form

$$\inf_{\beta\in\mathbb{R}^d}\sup_{\mathbb{Q}\in\mathbb{B}}\mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2], \tag{2}$$

where $\mathbb{Q}$ is a joint distribution over (X, Y). Generating a collection of distribution sets $\mathbb{B}$ in a systematic manner and solving (2) for each such set will form a family of experts $\mathcal{E}$.

In a purely data-driven setting with no additional information, it is attractive to probe into the distributional regions *in between* the empirical source distribution $\widehat{\mathbb{P}}_{\mathrm{S}} = N_{\mathrm{S}}^{-1}\sum_{i=1}^{N_{\mathrm{S}}}\delta_{(\widehat{x}_i,\widehat{y}_i)}$ and the empirical target distribution $\widehat{\mathbb{P}}_{\mathrm{T}} = N_{\mathrm{T}}^{-1}\sum_{j=1}^{N_{\mathrm{T}}}\delta_{(\widehat{x}_j,\widehat{y}_j)}$. Because probability distributions reside in infinite-dimensional spaces, framing $\mathbb{B}$ in between $\widehat{\mathbb{P}}_{\mathrm{S}}$ and $\widehat{\mathbb{P}}_{\mathrm{T}}$ is a non-trivial task. Fortunately, because the expected square loss only depends on the first two moments of the joint distribution of (X, Y), it suffices to prescribe $\mathbb{B}$ using a finite parametrization of distributional moments. To this end, let $p = d+1$ represent the dimension of the joint vector (X, Y). For a given set $\mathbb{U}$ on the space of mean vectors and covariance matrices $\mathbb{R}^p \times \mathbb{S}_+^p$, we consider $\mathbb{B}$ as the lifted distribution set that contains all distributions whose moments belong to $\mathbb{U}$, that is,

$$\mathbb{B} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma),\ (\mu, \Sigma) \in \mathbb{U}\},$$

where $\mathcal{M}(\mathbb{R}^p)$ denotes the set of all distributions on $\mathbb{R}^p$, and the notation $\mathbb{Q} \sim (\mu, \Sigma)$ expresses that $\mathbb{Q}$ has mean μ and covariance matrix Σ. It is convenient to construct the moment information set $\mathbb{U}$ using a divergence on $\mathbb{R}^p \times \mathbb{S}_+^p$.

Definition 3.1 (Divergence)**.** *A divergence ψ on $\mathbb{R}^p \times \mathbb{S}_+^p$ satisfies the following properties:*

- *non-negativity: for any (μ, Σ), $(\widehat{\mu}, \widehat{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}_+^p$, we have $\psi((\mu, \Sigma) \parallel (\widehat{\mu}, \widehat{\Sigma})) \geq 0$,*
- *indiscernability: $\psi((\mu, \Sigma) \parallel (\widehat{\mu}, \widehat{\Sigma})) = 0$ implies $(\mu, \Sigma) = (\widehat{\mu}, \widehat{\Sigma})$.*

In this paper, we will explore two divergences in the space of mean vectors and covariance matrices that are motivated by popular measures of dissimilarity between distributions. The divergence $\mathbb{D}$ is motivated by the Kullback-Leibler (KL) divergence.

Definition 3.2 (Kullback-Leibler-type divergence)**.** *The divergence $\mathbb{D}$ from tuple $(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_{++}^p$ to tuple $(\widehat{\mu}, \widehat{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}_{++}^p$ amounts to*

$$\mathbb{D}\big((\mu, \Sigma) \parallel (\widehat{\mu}, \widehat{\Sigma})\big) \triangleq$$
$$(\widehat{\mu} - \mu)^\top \widehat{\Sigma}^{-1}(\widehat{\mu} - \mu) + \mathrm{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right] - \log\det(\Sigma\widehat{\Sigma}^{-1}) - p.$$

In fact $\mathbb{D}$ is equivalent to the KL divergence between two non-degenerate Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ (up to a factor of 2). As a consequence, $\mathbb{D}$ is non-negative, and it collapses to 0 if and only if $\Sigma = \widehat{\Sigma}$ and $\mu = \widehat{\mu}$. We can also show that $\mathbb{D}$ is affine-invariant. However, we emphasize that $\mathbb{D}$ is not symmetric and $\mathbb{D}\big((\mu, \Sigma) \parallel (\widehat{\mu}, \widehat{\Sigma})\big) \neq \mathbb{D}\big((\widehat{\mu}, \widehat{\Sigma}) \parallel (\mu, \Sigma)\big)$ in general.

We also study the divergence $\mathbb{W}$ which is motivated by the Wasserstein distance.

Definition 3.3 (Wasserstein-type divergence)**.** *The divergence* $\mathbb{W}$ *between two tuples* $(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p$ *and* $(\widehat{\mu}, \widehat{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}_+^p$ *amounts to*

$$\mathbb{W}\big((\mu,\Sigma)\,\|\,(\widehat{\mu},\widehat{\Sigma})\big) \triangleq \|\mu-\widehat{\mu}\|_2^2+\mathrm{Tr}\left[\Sigma+\widehat{\Sigma}-2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right].$$

The divergence $\mathbb{W}$ coincides with the *squared* type-2 Wasserstein distance between two Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ (Givens & Shortt, 1984). One can readily show that $\mathbb{W}$ is non-negative, and it vanishes if and only if $(\mu, \Sigma)=(\widehat{\mu}, \widehat{\Sigma})$. Thus, $\mathbb{W}$ is a symmetric divergence.

In Sections 4 and 5 we examine in detail two strategies to frame $\mathbb{U}$ and its corresponding distribution set $\mathbb{B}$ in a principled manner, and we devise optimization techniques to solve the resulting robust estimation problems.

4. "Interpolate, then Robustify" Strategy

"Interpolate, then Robustify" (IR) is an intuitive strategy to systematically probe into distributional regions between $\widehat{\mathbb{P}}_{\mathrm{S}}$ and $\widehat{\mathbb{P}}_{\mathrm{T}}$. Let $(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$ be the empirical mean vector and covariance matrix of $\widehat{\mathbb{P}}_{\mathrm{S}}$, that is,

$$\widehat{\mu}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}}\sum_{i=1}^{N_{\mathrm{S}}}\begin{pmatrix}\widehat{x}_i\\ \widehat{y}_i\end{pmatrix}, \; \widehat{\Sigma}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}}\sum_{i=1}^{N_{\mathrm{S}}}\begin{pmatrix}\widehat{x}_i\\ \widehat{y}_i\end{pmatrix}\begin{pmatrix}\widehat{x}_i\\ \widehat{y}_i\end{pmatrix}^{\top} - \widehat{\mu}_{\mathrm{S}}\widehat{\mu}_{\mathrm{S}}^{\top},$$

and let $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$ be defined analogously for $\widehat{\mathbb{P}}_{\mathrm{T}}$. The IR strategy applies repeatedly the following two steps to generate distribution sets. First, interpolate between $(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$ and $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$ to obtain a new pair $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ parametrized by $\lambda \in [0, 1]$. Second, construct a moment set $\mathbb{U}_{\lambda,\rho}$ as a ball of radius ρ circumscribing the pair $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$, then lift the moment set $\mathbb{U}_{\lambda,\rho}$ to the corresponding distribution set $\mathbb{B}_{\lambda,\rho}$. More specifically, $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is the ψ-barycenter between $(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$ and $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$, which is obtained by solving

$$\min_{\mu\in\mathbb{R}^p,\Sigma\in\mathbb{S}_+^p} \lambda\psi((\mu,\Sigma)\,\|\,(\widehat{\mu}_{\mathrm{S}},\widehat{\Sigma}_{\mathrm{S}}))+ (1-\lambda)\psi((\mu,\Sigma)\,\|\,(\widehat{\mu}_{\mathrm{T}},\widehat{\Sigma}_{\mathrm{T}})). \tag{3}$$

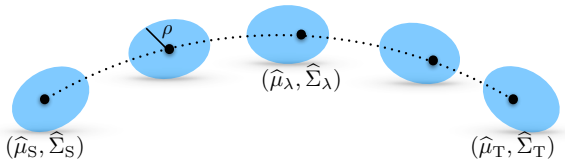


Figure 2. The dashed curve shows the barycenter interpolations parametrized by $\lambda \in [0, 1]$. Ellipses represent $\mathbb{U}_{\lambda,\rho}$ at different λ.

Then, we employ the divergence ψ to construct an uncertainty set $\mathbb{U}_{\lambda,\rho}$ in the mean-covariance matrix space as

$$\mathbb{U}_{\lambda,\rho} \triangleq \left\{(\mu,\Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \psi((\mu,\Sigma)\,\|\,(\widehat{\mu}_\lambda,\widehat{\Sigma}_\lambda)) \le \rho\right\}.$$

The outlined procedure is illustrated in Figure 2. An expert is now obtained by solving the distributionally robust least squares problem (2) with respect to the distribution set

$$\mathbb{B}_{\lambda,\rho} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu,\Sigma), (\mu,\Sigma) \in \mathbb{U}_{\lambda,\rho}\}.$$

Notice that in this strategy the parameter $\lambda \in [0, 1]$ characterizes the explanatory power of the source domain to the target domain: if $\lambda = 0$, then $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$, and if $\lambda = 1$, then $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$. Thus, as λ decreases, $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is moving farther away from the source information $(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$, and $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is pulled towards the target information $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$.

The choice of the divergence ψ influences both the barycenter problem (3) and the formation of the set $\mathbb{U}_{\lambda,\rho}$. Next, we study the special case of the IR strategy with the KL-type divergence and the Wasserstein-type divergence.

4.1. Kullback-Leibler-type Divergence

The KL-type divergence $\mathbb{D}$ in Definition 3.2 is not symmetric. Hence, it is worthwhile to note that the barycenter problem (3) optimizes over (μ, Σ) being placed in the first argument of $\mathbb{D}$, and that the set $\mathbb{U}_{\lambda,\rho}$ is also defined with the pair (μ, Σ) being placed in the first argument. Under the divergence $\mathbb{D}$, the barycenter $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ admits a closed form expression. This fact is well-known in the field of KL fusion of Gaussian distributions (Battistelli et al., 2013).

Proposition 4.1 (KL barycenter)**.** *Suppose that* ψ *is the KL-type divergence. If* $\widehat{\Sigma}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{T}} \succ 0$*, then* $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ *is the minimizer of the barycenter problem* (3) *with*

$$\widehat{\Sigma}_\lambda = (\lambda\widehat{\Sigma}_{\mathrm{S}}^{-1} + (1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1})^{-1} \succ 0,$$
$$\widehat{\mu}_\lambda = \widehat{\Sigma}_\lambda\big(\lambda\widehat{\Sigma}_{\mathrm{S}}^{-1}\widehat{\mu}_{\mathrm{S}} + (1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1}\widehat{\mu}_{\mathrm{T}}\big).$$

For a given $\lambda \in [0, 1]$ and $\rho \ge 0$, the corresponding IR-KL expert is obtained by solving

$$\min_{\beta\in\mathbb{R}^d}\left\{f_{\lambda,\rho}(\beta) \triangleq \sup_{\mathbb{Q}\in\mathbb{B}_{\lambda,\rho}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]\right\}. \tag{4}$$

Problem (4) can be efficiently solved using a gradient-descent algorithm. To do this, the next proposition establishes the relevant properties of $f_{\lambda,\rho}$.

Proposition 4.2 (Properties of $f_{\lambda,\rho}$)**.** *The function* $f_{\lambda,\rho}$ *is convex and continuously differentiable with*

$$\nabla f_{\lambda,\rho}(\beta) = \frac{2\kappa^\star\left(\omega_2\widehat{\Sigma}_\lambda w + (\kappa^\star-\omega_1)(\widehat{\Sigma}_\lambda+\widehat{\mu}_\lambda\widehat{\mu}_\lambda^\top)w\right)_{1:d}}{(\kappa^\star-\omega_1)^2},$$

where $w = [\beta^\top, -1]^\top$, $\omega_1 = w^\top\widehat{\Sigma}_\lambda w$, $\omega_2 = (w^\top\widehat{\mu})^2$ *and* $\kappa^\star \in (\omega_1, \omega_1\big(1+2\rho+\sqrt{1+4\rho\omega_2}\big)/(2\rho)]$ *is the unique solution of the equation*

$$\rho = (\kappa-\omega_1)^{-2}\omega_1\omega_2 + (\kappa-\omega_1)^{-1}\omega_1 + \log(1-\kappa^{-1}\omega_1).$$

Furthermore, $f_{\lambda,\rho}$ is locally smooth at any $\beta \in \mathbb{R}^d$, i.e., *there exist constants $C_\beta, \epsilon_\beta > 0$ such that for any $\beta' \in \mathbb{R}^d$ with $\|\beta' - \beta\|_2 \le \epsilon_\beta$, we have $\|\nabla f_{\lambda,\rho}(\beta') - \nabla f_{\lambda,\rho}(\beta)\|_2 \le C_\beta \|\beta' - \beta\|_2$.*

Thanks to Proposition 4.2, we can apply the adaptive gradient method to solve problem (4) to global optimality, and the algorithm enjoys a sublinear rate $|f_{\lambda,\rho}(\bar{\beta}^k) - f_{\lambda,\rho}(\beta^\star_{\lambda,\rho})| \le O(k^{-1})$, where $\bar{\beta}^k$ is a certain average of the iterates, and $\beta^\star_{\lambda,\rho}$ is an optimal solution of (4). The algorithm and its guarantees are detailed in Malitsky & Mishchenko (2019).

4.2. Wasserstein-type Divergence

Under the divergence $\mathbb{W}$ in Definition 3.3, problem (3) resembles the Wasserstein barycenter in the space of Gaussian distributions. The result from Agueh & Carlier (2011, §6.2) implies that the barycenter $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ admits a closed form expression following the McCann's interpolant (McCann, 1997, Example 1.7).

Proposition 4.3 (Wasserstein interpolation)**.** *Suppose that ψ is the Wasserstein-type divergence. If $\widehat{\Sigma}_{\mathrm{S}} \succ 0$, then $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is the minimizer of problem (3) with*

$$\widehat{\mu}_\lambda = \lambda \widehat{\mu}_{\mathrm{S}} + (1-\lambda)\widehat{\mu}_{\mathrm{T}},$$
$$\widehat{\Sigma}_\lambda = (\lambda I_p + (1-\lambda)L)\widehat{\Sigma}_{\mathrm{S}}(\lambda I_p + (1-\lambda)L),$$

where $L = \widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}(\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}\widehat{\Sigma}_{\mathrm{S}}\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}})^{-\frac{1}{2}}\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}$.

For a given $\lambda \in [0,1]$ and $\rho \ge 0$, we obtain the corresponding IR-Wasserstein expert by solving a conic program using off-the-shelf solvers such as MOSEK ApS (2019).

Proposition 4.4 (IR-Wasserstein expert)**.** *Suppose that ψ is the Wasserstein-type divergence. Problem (2) with $\mathbb{B} \equiv \mathbb{B}_{\lambda,\rho}$ is equivalent to the second order cone program*

$$\min_{\beta \in \mathbb{R}^d} \left\| (\widehat{\Sigma}_\lambda + \widehat{\mu}_\lambda \widehat{\mu}_\lambda^\top)^{\frac{1}{2}} \begin{bmatrix} \beta \\ -1 \end{bmatrix} \right\|_2 + \sqrt{\rho} \left\| \begin{bmatrix} \beta \\ -1 \end{bmatrix} \right\|_2.$$

5. "Surround, then Intersect" Strategy

"Surround, then Intersect" (SI) probes naturally into the distributional space by intersecting two balls centered at the empirical moments. More specifically, this strategy circumscribes $(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$ (respectively, $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$) with a ball of radius ρ_{S} (respectively, ρ_{T}) using the ψ-divergence. Consequentially, the moment information set $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ in the mean vector-covariance matrix space is defined as

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{ \begin{array}{l} (\mu,\Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that:} \\ \psi((\mu,\Sigma) \,\|\, (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \le \rho_{\mathrm{S}} \\ \psi((\mu,\Sigma) \,\|\, (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \le \rho_{\mathrm{T}} \\ \Sigma + \mu\mu^\top \succeq \varepsilon I_p \end{array} \right\},$$

where the small constant $\varepsilon > 0$ improves numerical stability. This construction is graphically illustrated in Figure 3. An expert is now obtained by solving the distributionally robust least squares problem (2) subject to the distributional set

$$\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu,\Sigma),\ (\mu,\Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}.$$

Note that $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is well-defined only if the radii $(\rho_{\mathrm{S}}, \rho_{\mathrm{T}})$ are sufficiently large so that the intersection of the two balls becomes non-empty. A sensible approach to set these parameters is to fix ρ_{S} and to find a sufficiently large ρ_{T} so that $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is non-empty. In this way, the SI strategy characterizes the explanatory power of the source domain to the target domain by the radius ρ_{S}: if $\rho_{\mathrm{S}} = 0$ then $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ becomes a singleton $\{(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})\}$, representing the *belief* that the source domain possess absolute explanatory power onto the target domain. As ρ_{S} increases, $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is gradually pulled towards the empirical target moments $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$. Next, we study the special case of the SI strategy with the KL-type divergence and the Wasserstein-type divergence.

5.1. Kullback-Leibler-type Divergence

Recall that $\mathbb{D}$ is asymmetric and (μ,Σ) is the first argument of $\mathbb{D}$ in the definition of $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$. We first study conditions on ρ_{T} under which the ambiguity set $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is non-empty.

Proposition 5.1 (Minimum radius)**.** *Suppose that ψ is the KL-type divergence. For any $\rho_S > 0$ the sets $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ and $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ are non-empty if $\rho_{\mathrm{T}} \ge \mathbb{D}((\widehat{\mu}_{\gamma^\star}, \widehat{\Sigma}_{\gamma^\star}) \,\|\, (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))$, where $\gamma^\star$ is a maximizer of*

$$\begin{array}{ll} \sup & \mathbb{D}((\widehat{\mu}_\gamma, \widehat{\Sigma}_\gamma) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) + \mathbb{D}((\widehat{\mu}_\gamma, \widehat{\Sigma}_\gamma) \| (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) - \gamma\rho_{\mathrm{S}} \\ \text{s.t.} & \gamma \in \mathbb{R}_+, \widehat{\Sigma}_\gamma = (1+\gamma)(\gamma\widehat{\Sigma}_{\mathrm{S}}^{-1} + \widehat{\Sigma}_{\mathrm{T}}^{-1})^{-1} \in \mathbb{S}_+^p, \\ & \widehat{\mu}_\gamma = \widehat{\Sigma}_\gamma(\gamma\widehat{\Sigma}_{\mathrm{S}}^{-1}\widehat{\mu}_{\mathrm{S}} + \widehat{\Sigma}_{\mathrm{T}}^{-1}\widehat{\mu}_{\mathrm{T}})/(1+\gamma) \in \mathbb{R}^p \end{array}$$

The above optimization problem is effectively one-dimensional and can therefore be solved by bisection on γ. The next theorem asserts that the SI-KL experts are formed by solving a semidefinite program.

Theorem 5.2 (SI-KL Expert)**.** *Suppose that ψ is the KL-type divergence and $\mathbb{B} \equiv \mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is non-empty. Then $\beta^\star = (M^\star_{XX})^{-1}M^\star_{XY}$ solves problem (2), where $(M^\star_{XX}, M^\star_{XY})$ is a solution of the convex semidefinite program*

$$\begin{array}{ll} \sup & \tau \\ \text{s.t.} & M_{XX} \in \mathbb{R}^{d\times d},\ M_{XY} \in \mathbb{R}^{d\times 1},\ M_{YY} \in \mathbb{R} \\ & \tau \in \mathbb{R}_+,\ \mu \in \mathbb{R}^p,\ M \in \mathbb{S}_{++}^p, t \in \mathbb{R}_+ \\ & \widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\widehat{\mu}_k - 2\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\mu + \mathrm{Tr}\left[M\widehat{\Sigma}_k^{-1}\right] - \\ & \log\det(M\widehat{\Sigma}_k^{-1}) - \log(1-t) - p \le \rho_k \ \ \forall k \in \{\mathrm{S},\mathrm{T}\} \\ & \begin{bmatrix} M & \mu \\ \mu^\top & t \end{bmatrix} \succeq 0,\ \begin{bmatrix} M_{XX} & M_{XY} \\ M_{XY}^\top & M_{YY} - \tau \end{bmatrix} \succeq 0 \\ & M = \begin{bmatrix} M_{XX} & M_{XY} \\ M_{XY}^\top & M_{YY} \end{bmatrix} \succeq \varepsilon I_p. \end{array}$$

5.2. Wasserstein-type Divergence

The space $\mathbb{R}^p \times \mathbb{S}_+^p$ can be endowed with a distance inherited from the Wasserstein distance between Gaussian

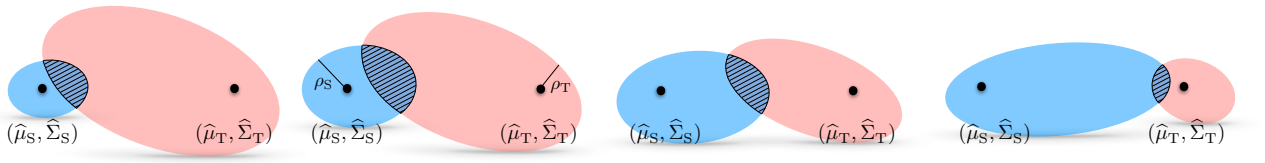


Figure 3. Varying $(\rho_{\mathrm{S}}, \rho_{\mathrm{T}})$ frames different moment sets $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ (hatched regions). The radius ρ_{S} increases from left to right.

distribution. For any $\rho_{\mathrm{S}} > 0$, the minimum radius for ρ_{T} that makes $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ non-empty is known in closed form.

Proposition 5.3 (Minimum radius)**.** *Suppose that ψ is the Wasserstein-type divergence. For any $\rho_{\mathrm{S}} > 0$ the sets $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ and $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ are non-empty if*

$$\rho_{\mathrm{T}} \geq \left(\sqrt{\mathbb{W}((\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}) \parallel (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))} - \sqrt{\rho_{\mathrm{S}}} \right)^2 .$$

The next theorem asserts that the SI-Wasserstein experts are constructed by solving a semidefinite program.

Theorem 5.4 (SI-Wasserstein expert)**.** *Suppose that ψ is the Wasserstein-type divergence and $\mathbb{B} \equiv \mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is non-empty. Then $\beta^\star = (M_{XX}^\star)^{-1} M_{XY}^\star$ solves problem (2), where $(M_{XX}^\star, M_{XY}^\star)$ is a solution of the linear semidefinite program*

$$\begin{array}{ll} \sup & \tau \\ \text{s.t.} & M_{XX} \in \mathbb{R}^{d\times d}, M_{XY} \in \mathbb{R}^{d\times 1}, M_{YY} \in \mathbb{R} \\ & \tau \in \mathbb{R}_+, \mu \in \mathbb{R}^p, M, H \in \mathbb{S}_+^p, C_{\mathrm{S}}, C_{\mathrm{T}} \in \mathbb{R}^{p\times p} \\ & \left.\begin{array}{l} \|\widehat{\mu}_k\|_2^2 - 2\widehat{\mu}_k^\top \mu + \mathrm{Tr}\left[M + \widehat{\Sigma}_k - 2C_k\right] \leq \rho_k \\ \begin{bmatrix} H & C_k \\ C_k^\top & \widehat{\Sigma}_k \end{bmatrix} \succeq 0 \end{array}\right\} k \in \{\mathrm{S}, \mathrm{T}\} \\ & \begin{bmatrix} M - H & \mu \\ \mu^\top & \end{bmatrix} \succeq 0 \\ & \begin{bmatrix} M_{XX} & M_{XY} \\ M_{XY}^\top & M_{YY} - \tau \end{bmatrix} \succeq 0, \; M = \begin{bmatrix} M_{XX} & M_{XY} \\ M_{XY}^\top & M_{YY} \end{bmatrix} \succeq \varepsilon I_p. \end{array}$$

6. Numerical Experiments

The second-order cone and semidefinite programs are modelled in MATLAB via YALMIP (Löfberg, 2004) and solved with MOSEK ApS (2019). All experiments are run on an Intel i7-8700 CPU (3.2 GHz) computer with 16GB RAM. The corresponding codes are available at `https://github.com/RAO-EPFL/DR-DA.git`.

We now aim to assess the performance of experts and demonstrate the effects of robustness. In all experiments we generate the set $\mathcal{E} = \{\beta_1, \ldots, \beta_{|\mathcal{E}|}\}$ of experts with $|\mathcal{E}| = 10$.

We consider four family of robust experts generated by:

- IR-KL: with $\rho = \mathbb{D}((\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}) \,\|\, (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}))/(3|\mathcal{E}|)$ and λ is spaced from 1 to 0 in exponentially increasing steps.[1]
- IR-WASS: with $\rho = \mathbb{W}((\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}))/(3|\mathcal{E}|)$ and λ is spaced from 1 to 0 in exponentially increasing steps.
- SI-KL: with ρ_{S} spaced from 10^{-3} to $\mathbb{D}((\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}) \parallel (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}))-1$ in exponentially increasing steps. For a given ρ_{S}, ρ_{T} is set to the sum of the minimum target radius satisfying the condition of Proposition 5.1 and $\rho_{\mathrm{S}}/2$.[2]
- SI-WASS: with ρ_S spaced from 10^{-4} to $\mathbb{W}((\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}) \parallel (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}))$ in increasing exponential steps. For a given ρ_{S}, ρ_{T} is set to the sum of the minimum radius that satisfies the condition in Proposition 5.3 and $\rho_{\mathrm{S}}/2$.

We benchmark against the Convex Combination (CC) and Reweighting (RW) experts in Section 2 generated by

- CC-L: with λ equally spaced in $[0, 1]$, thus provides uniformly spaced distributional regions in between domains.
- CC-TL: with λ equally spaced in $[0, 0.5]$, thus distributional regions are formed around the target domain.
- CC-SL: with λ equally spaced in $[0.5, 1]$, thus distributional regions are formed around the source domain.
- CC-TE: with λ spaced from 0 to 1 in exponentially increasing steps, thus the constructed distributional regions are concentrated towards the target domain.
- CC-SE: with λ spaced from 1 to 0 in exponentially increasing steps, thus the constructed distributional regions are concentrated towards the source domain.
- RWS: with h equally spaced in $[0.5, 10]$.

We consider a family of sequential empirical ridge regression estimators generated by training for each J over

- LSE-T, the union of the target dataset $(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$, and the sequentially arriving target test data $(x_j, y_j)_{j=1}^{J-1}$,
- LSE-T&S, the union of the source data $(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$, the target data $(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$ and the sequentially arriving target test data $(x_j, y_j)_{j=1}^{J-1}$.

Note that both LSE-T and LSE-T&S predictors dynamically incorporate the new data to adapt the prediction. Thereby, they have an unfair advantage in the long run over the other experts that are trained only once at the beginning with N_{T} samples from the test domain.

[1] We say that λ is spaced from a to b in K exponentially increasing steps if $\lambda_1 = a$ and $\lambda_{k+1} = \lambda_k - (a - b)\exp(k)/\sum_{i=1}^{K-1}\exp(i)$ for all $k \in \{2, \ldots, K-1\}$.

[2] If $d \geq 15$, then the minimum value of ρ_{S} is set to 5 to improve numerical stability.

Data Set	Time	IR-KL	IR-WASS	SI-KL	SI-WASS	CC-L	CC-TL	CC-SL	CC-TE	CC-SE	RWS	LSE-T	LSE-T&S
Uber&Lyft	5	17.65	**1.00**	199.28	1.01	34.04	98.43	12.03	155.71	1.74	1.45	119.65	11.08
	10	13.67	**1.00**	111.52	1.01	30.85	99.22	11.40	161.72	1.58	1.34	137.15	6.32
	50	13.39	**1.00**	60.29	1.01	25.87	85.06	9.72	147.45	1.42	1.16	57.85	2.12
	100	15.24	**1.00**	59.06	1.01	26.01	85.77	9.91	148.49	1.41	1.12	31.25	1.57
US Births (2018)	5	79.83	1.02	44.71	**1.00**	64.99	257.60	25.13	432.09	2.07	4.50	727.88	39.17
	10	115.47	1.02	39.35	**1.00**	45.59	195.14	18.33	339.11	1.60	3.29	524.39	19.28
	50	107.40	1.01	40.04	**1.00**	42.74	192.46	13.12	361.51	1.31	2.00	191.27	5.20
	100	117.03	1.01	53.13	**1.00**	45.35	208.65	12.94	397.33	1.22	1.75	104.75	3.19
Life Expectancy	5	33.18	**1.00**	6.24	1.03	17.24	77.06	7.38	125.71	1.46	1.15	255.08	20.72
	10	25.59	**1.00**	5.45	1.02	12.49	60.19	5.50	104.00	1.40	1.15	167.15	10.73
	50	19.81	**1.00**	8.70	1.01	7.57	44.00	3.10	84.98	1.38	1.10	39.83	3.15
	100	19.02	**1.00**	8.25	1.005	6.82	41.40	2.68	83.60	1.38	1.08	20.42	2.10
House Prices in KC	5	1.58	**1.00**	1.21	1.01	3.98	8.87	2.12	13.31	1.29	1.23	11.75	3.70
	10	1.52	**1.00**	1.20	1.01	3.58	7.77	2.02	11.70	1.27	1.23	6.93	2.25
	50	1.34	**1.00**	1.31	1.01	2.79	6.52	1.86	10.37	1.27	1.20	3.91	1.30
	100	1.34	**1.00**	1.30	1.01	2.65	6.54	1.91	10.74	1.27	1.18	2.72	1.12
California Housing	5	63.33	1.05	3.31	**1.00**	27.63	102.82	9.60	181.52	1.35	1.17	96.43	54.34
	10	68.08	1.04	2.42	**1.00**	20.57	91.86	6.23	169.87	1.19	1.17	45.64	24.76
	50	70.08	1.01	1.97	**1.00**	11.79	81.72	2.49	170.18	1.05	1.13	10.17	5.63
	100	72.80	1.003	1.90	**1.00**	9.71	79.19	1.83	173.96	1.04	1.14	5.81	3.39

Table 1. Normalized cumulative loss values averaged over 100 independent runs.

The main reason behind using exponential step sizes originates from the asymmetric nature of $\mathbb{D}$. For simplicity, we also use it for experts with $\mathbb{W}$. To ensure fairness in the competition between experts, we vary the parameters of the non-robust experts also in exponential steps.

We compare the performance of our model against the above non-robust benchmarks on 5 Kaggle datasets:[3]

- **Uber&Lyft** contains $d=38$ features of Uber and Lyft cab rides in Boston including the distances, date and time of the hailing, a weather summary for that day. The prediction target is the price of the ride. We divide the dataset based on the company, Uber (source) and Lyft (target).
- **US Births (2018)** has $d = 36$ predictive features of child births in the United States in the year of 2018 including the gender of the infant, mother's weight gain, and mother's per-pregnancy body mass index. The task is to predict the weight of the infants. We divide the dataset based on gender: male (source) and female (target).
- **Life Expectancy** contains $d = 19$ predictive features, and the target variable is the life expectancy at birth. The dataset is divided into two subgroups: developing (source) and developed (target) countries.
- **House Prices in King Country** contains $d=14$ predictive variables, the target variable is the transaction price of the houses. We split the dataset into two domains: houses built in $[1950, 2000)$ (source) and $[2000, 2010]$ (target).
- **California Housing Prices** has $d = 9$ predictive features, the target variable is the price of houses. We divide this dataset into houses with less than an hour drive to the ocean shore (source) and houses in inland (target).

We use all samples from the source domain for training, and we form the target training set by drawing $N_{\mathrm{T}}=d$ samples from the target dataset. Later, we randomly sample $J = 1000$ data points from the remaining target samples to form the sequentially arriving target test samples. Note that the performance of the experts is sensitive to the data, and thus we replicate this procedure 100 times. We set the regularization parameter of the ridge regression problem to $\eta = 10^{-6}$ and the learning rate of the BOA algorithm to $\upsilon = 0.5$. We measure the performance of the experts by the cumulative loss (1) calculated for every J.

Table 1 shows the average cumulative loss of each aggregated expert obtained by the BOA algorithm for all datasets and for $J = \{5, 10, 50, 100\}$ across 100 independent runs. In each row, the minimum loss is normalized to 1, and the remaining entries are presented by the multiplicative factor of the minimum value. This result suggests that the IR-WASS and SI-WASS experts perform favorably over the competitors in that their cumulative loss at each time step is substantially lower than that of most other competitors.

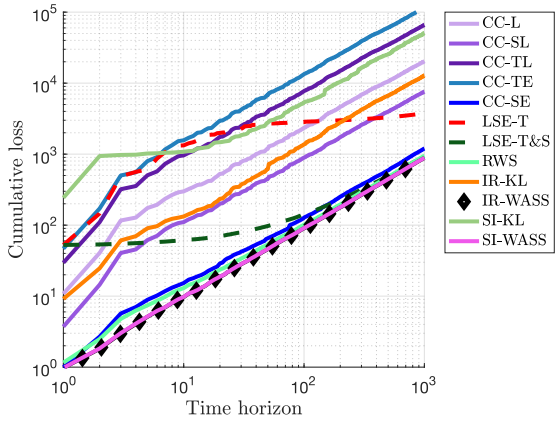


Figure 4. Cumulative loss averaged over 100 runs, Uber&Lyft.

Figure 4 demonstrates how the average cumulative loss

[3]Descriptions and download links are provided in the appendix.

in (1) grows over time for the Uber&Lyft dataset. Figure 4 shows that the loss of LSE-T&S is initially constant at a high level, which highlights the discrepancy between the two domain distributions. The growth rate of LSE-T decays faster than that of other experts, and the time when LSE-T saturates indicates when the combined target domain data alone is sufficient to construct a single, competitive predictor without using any source domain data.

Concluding Remarks. The theoretical and experimental results in this paper suggest that IR-WASS and SI-WASS are attractive schemes to generate a family of robust least squares experts. Moreover, the IR-WASS and SI-WASS experts are extremely easy to compute because it requires solving only a second-order cone or a linear semidefinite program. We observe that KL-type divergence schemes are less numerically stable due to the computation of the log-determinant and the inverse of a nearly singular covariance matrix $\widehat{\Sigma}_{\mathrm{T}}$. Setting the parameters for KL-type divergence schemes is also harder due to the asymmetry of the divergence $\mathbb{D}$. While this paper focuses solely on *interpolating* schemes, it would also be interesting to explore *extrapolating* schemes in future research.

Acknowledgments

Material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF grants 1915967, 1820942, 1838676, and also from the China Merchant Bank. Man-Chung Yue gratefully acknowledges the support by HKRGC under the Early Career Scheme Funding 25302420.

References

Agueh, M. and Carlier, G. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019.

Baktashmotlagh, M., Harandi, M. T., Lovell, B. C., and Salzmann, M. Unsupervised domain adaptation by domain invariant projection. In *IEEE International Conference on Computer Vision*, pp. 769–776, 2013.

Battistelli, G., Chisci, L., Fantacci, C., Farina, A., and Graziano, A. Consensus CPHD filter for distributed multitarget tracking. *IEEE Journal of Selected Topics in Signal Processing*, 7(3):508–520, 2013.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19: 137, 2007.

Bernstein, D. S. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

Bertsekas, D. *Convex Optimization Theory*. Athena Scientific, 2009.

Blanchet, J., Kang, Y., and Murthy, K. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Blitzer, J., McDonald, R., and Pereira, F. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, pp. 120–128, 2006.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Chen, X., Monfort, M., Liu, A., and Ziebart, B. D. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pp. 1270–1279, 2016.

Chu, C. and Wang, R. A survey of domain adaptation for neural machine translation. In *International Conference on Computational Linguistics*, pp. 1304–1319. Association for Computational Linguistics, 2018.

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

Cortes, C., Mohri, M., and Medina, A. M. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (9):1853–1865, 2017.

Csurka, G. *A Comprehensive Survey on Domain Adaptation for Visual Applications*, pp. 1–35. Springer International Publishing, 2017.

de Mathelin, A., Richard, G., Mougeot, M., and Vayatis, N. Adversarial weighting for domain adaptation in regression. *arXiv preprint arXiv:2006.08251*, 2020.

Delage, E. and Ye, Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pp. 1180–1189, 2015.

Gao, R. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

Gao, R., Xie, L., Xie, Y., and Xu, H. Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pp. 7913–7923, 2018.

Garcke, J. and Vanck, T. Importance weighted inductive transfer learning for regression. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 466–481, 2014.

Ghaoui, L. E. and Lebret, H. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D., and Li, W. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pp. 597–613, 2016.

Givens, C. and Shortt, R. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

Goh, J. and Sim, M. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, 19:601–608, 2006.

Jiang, J. and Zhai, C. Instance weighting for domain adaptation in NLP. In *Association of Computational Linguistics*, pp. 264–271, 2007.

Koniusz, P., Tas, Y., and Porikli, F. Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4478–4487, 2017.

Kuhn, D., Mohajerin Esfahani, P., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pp. 130–166. 2019.

Kumar, A., Saha, A., and Daume, H. Co-regularization based semi-supervised domain adaptation. *Advances in Neural Information Processing Systems*, pp. 478–486, 2010.

Lam, H. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Li, Q. Literature survey: Domain adaptation algorithms for natural language processing. *Department of Computer Science The Graduate Center, The City University of New York*, pp. 8–10, 2012.

Li, Y., Wang, L., Wang, J., Ye, J., and Reddy, C. K. Transfer learning for survival analysis via efficient L2,1-norm regularized Cox regression. In *IEEE International Conference on Data Mining*, pp. 231–240, 2016.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pp. 3122–3130, 2018.

Löfberg, J. YALMIP: A toolbox for modeling and optimization in MATLAB. In *IEEE International Conference on Robotics and Automation*, pp. 284–289, 2004.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Unsupervised domain adaptation with residual transfer networks. In *International Conference on Neural Information Processing Systems*, pp. 136–144, 2016.

Lopez-Paz, D., Hernández-Lobato, J. M., and Schölkopf, B. Semi-supervised domain adaptation with non-parametric copulas. In *International Conference on Neural Information Processing Systems*, pp. 665–673, 2012.

Malitsky, Y. and Mishchenko, K. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019.

McCann, R. J. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

Mohajerin Esfahani, P. and Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

MOSEK ApS. *The MOSEK optimization toolbox. Version 9.2.*, 2019.

Motiian, S., Jones, Q., Iranmanesh, S., and Doretto, G. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6670–6680, 2017a.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. Unified deep supervised domain adaptation and generalization. In *IEEE International Conference on Computer Vision*, pp. 5715–5725, 2017b.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2208–2216, 2016.

Nguyen, V. A., Shafieezadeh-Abadeh, S., Yue, M.-C., Kuhn, D., and Wiesemann, W. Calculating optimistic likelihoods using (geodesically) convex optimization. In *Advances in Neural Information Processing Systems*, 2019a.

Nguyen, V. A., Shafieezadeh-Abadeh, S., Yue, M.-C., Kuhn, D., and Wiesemann, W. Optimistic distributionally robust optimization for nonparametric likelihood approximation. In *Advances in Neural Information Processing Systems 32*, 2019b.

Nguyen, V. A., Si, N., and Blanchet, J. Robust Bayesian classification using an optimistic score ratio. In *International Conference on Machine Learning*, 2020.

Pardoe, D. and Stone, P. Boosting for regression transfer. In *International Conference on Machine Learning*, 2010.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. *Advances in Domain Adaptation Theory*. Elsevier, 2019.

Richard, G., de Mathelin, A., Hébrail, G., Mougeot, M., and Vayatis, N. Unsupervised multi-source domain adaptation for regression. 2020.

Saha, A., Rai, P., Daumé, H., Venkatasubramanian, S., and DuVall, S. L. Active supervised domain adaptation. In *Machine Learning and Knowledge Discovery in Databases*, pp. 97–112, 2011.

Salaken, S. M., Khosravi, A., Nguyen, T., and Nahavandi, S. Seeded transfer learning for regression problems with deep learning. *Expert Systems with Applications*, 115: 565 – 577, 2019.

Shafieezadeh-Abadeh, S., Nguyen, V. A., Kuhn, D., and Mohajerin Esfahani, P. Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8474–8483, 2018.

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

Sindhwani, V., Niyogi, P., and Belkin, M. A co-regularization approach to semi-supervised learning with multiple views. In *ICML workshop on learning with multiple views*, pp. 74–79, 2005.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Søgaard, A. Semi-supervised learning and domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies*, 6(2):1–103, 2013.

Still, G. *Lectures on Parametric Optimization: An Introduction*. 2018.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Sun, Q., Chattopadhyay, R., Panchanathan, S., and Ye, J. A two-stage weighting framework for multi-source domain adaptation. In *Advances in Neural Information Processing Systems*, volume 24, pp. 505–513, 2011.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In *IEEE International Conference on Computer Vision*, pp. 4068–4076, 2015.

Villani, C. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

Wang, H., Liu, A., Yu, Z., Yue, Y., and Anandkumar, A. Distributionally robust learning for unsupervised domain adaptation. *arXiv preprint arXiv:2010.05784*, 2020.

Wang, M. and Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135 – 153, 2018.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big Data*, 3(1):1–40, 2016.

Wilson, G. and Cook, D. J. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5):1–46, 2020.

Wintenberger, O. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

Yao, T., Pan, Y., Ngo, C.-W., Li, H., and Mei, T. Semi-supervised domain adaptation with subspace learning for visual recognition. In *IEEE conference on Computer Vision and Pattern Recognition*, pp. 2142–2150, 2015.

Zhao, H., Zhang, S., Wu, G., Moura, J. M. F., Costeira, J. P., and Gordon, G. J. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

A. Appendix

A.1. Proof of Section 4

Proof of Proposition 4.1. Note that optimization problem (3) constitutes an unbounded convex optimization problem when ψ is the Kullback-Leibler-type divergence of Definition 3.1. Let $g(\mu,\Sigma) \triangleq \lambda \mathbb{D}((\mu,\Sigma) \parallel (\widehat{\mu}_{\mathrm{S}},\widehat{\Sigma}_{\mathrm{S}})) + (1-\lambda)\mathbb{D}((\mu,\Sigma) \parallel (\widehat{\mu}_{\mathrm{T}},\widehat{\Sigma}_{\mathrm{T}}))$, then, the first order optimality condition reads

$$\begin{aligned}\nabla_\mu g(\mu,\Sigma) &= 2\lambda\widehat{\Sigma}_{\mathrm{S}}^{-1}(\mu-\widehat{\mu}_{\mathrm{S}}) + 2(1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1}(\mu-\widehat{\mu}_{\mathrm{T}}) = 0,\\ \nabla_\Sigma g(\mu,\Sigma) &= \lambda\widehat{\Sigma}_{\mathrm{S}}^{-1} - \lambda\Sigma^{-1} + (1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1} - (1-\lambda)\Sigma^{-1} = 0.\end{aligned}$$

One can then show $(\widehat{\mu}_\lambda,\widehat{\Sigma}_\lambda)$ provided in statement of Proposition 4.1 solves the system of equalities above. □

Below we prove Proposition 4.2. In the proof of Proposition 4.2 and its auxiliary lemmas, Lemma A.1 and Lemma A.2, we omit the subscripts λ and ρ to avoid clutter.

Lemma A.1 (Dual problem)**.** *Fix $(\widehat{\mu},\widehat{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}_{++}^p$ and $\rho \geq 0$. For any symmetric matrix $H \in \mathbb{S}^p$, the optimization problem*

$$\left\{\begin{array}{ll}\sup\limits_{\mu,\Sigma} & \mathrm{Tr}\left[H(\Sigma+\mu\mu^\top)\right]\\ \mathrm{s.t.} & \mathrm{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right] - \log\det(\Sigma\widehat{\Sigma}^{-1}) - p + (\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu}) \leq \rho,\\ & \Sigma \succ 0\end{array}\right. \tag{A.5a}$$

admits the dual formulation

$$\left\{\begin{array}{ll}\inf & \kappa(\rho - \widehat{\mu}^\top\widehat{\Sigma}^{-1}\widehat{\mu}) + \kappa^2\widehat{\mu}^\top\widehat{\Sigma}^{-1}[\kappa\widehat{\Sigma}^{-1}-H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu} - \kappa\log\det(I - \widehat{\Sigma}^{\frac{1}{2}}H\widehat{\Sigma}^{\frac{1}{2}}/\kappa)\\ \mathrm{s.t.} & \kappa \geq 0,\ \kappa\widehat{\Sigma}^{-1} \succ H.\end{array}\right. \tag{A.5b}$$

Proof of Lemma A.1. For any $\mu \in \mathbb{R}^p$ such that $(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu}) \leq \rho$, denote the set $\mathcal{S}_\mu$ as

$$\mathcal{S}_\mu \triangleq \left\{\Sigma \in \mathbb{S}_{++}^p : \mathrm{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right] - \log\det\Sigma \leq \rho_\mu\right\},$$

where $\rho_\mu \in \mathbb{R}$ is defined as $\rho_\mu \triangleq \rho + p - \log\det\widehat{\Sigma} - (\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu})$. Using these auxiliary notations, problem (A.5a) can be re-expressed as a nested program of the form

$$\begin{array}{ll}\sup\limits_{\mu} & \mu^\top H\mu + \sup\limits_{\Sigma\in\mathcal{S}_\mu} \mathrm{Tr}\left[H\Sigma\right]\\ \mathrm{s.t.} & (\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu}) \leq \rho,\end{array}$$

where we emphasize that the constraint on μ is redundant, but it is added to ensure the feasibility of the inner supremum over Σ for every feasible value of μ of the outer problem. We now proceed to reformulate the supremum subproblem over Σ.

Assume momentarily that $H \neq 0$ and that μ satisfies $(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu}) < \rho$. In this case, one can verify that $\widehat{\Sigma}$ is a Slater point of the convex set $\mathcal{S}_\mu$. Using a duality argument, we find

$$\begin{aligned}\sup_{\Sigma\in\mathcal{S}_\mu} \mathrm{Tr}\left[H\Sigma\right] &= \sup_{\Sigma\succ 0}\inf_{\phi\geq 0} \mathrm{Tr}\left[H\Sigma\right] + \phi\big(\rho_\mu - \mathrm{Tr}\left[\widehat{\Sigma}^{-1}\Sigma\right] + \log\det\Sigma\big)\\ &= \inf_{\phi\geq 0}\left\{\phi\rho_\mu + \sup_{\Sigma\succ 0}\left\{\mathrm{Tr}\left[(H-\phi\widehat{\Sigma}^{-1})\Sigma\right] + \phi\log\det\Sigma\right\}\right\},\end{aligned}$$

where the last equality follows from strong duality (Bertsekas, 2009, Proposition 5.3.1). If $H - \phi\widehat{\Sigma}^{-1} \not\prec 0$, then the inner supremum problem becomes unbounded. To see this, let $\sigma \in \mathbb{R}_+$ be the maximum eigenvalue of $H - \phi\widehat{\Sigma}^{-1}$ with the corresponding eigenvector v, then the sequence $(\Sigma_k)_{k\in\mathbb{N}}$ with $\Sigma_k = I + kvv^\top$ attains the asymptotic maximum objective value of $+\infty$. If $H - \phi\widehat{\Sigma}^{-1} \prec 0$ then the inner supremum problem admits the unique optimal solution

$$\Sigma^\star(\phi) = \phi(\phi\widehat{\Sigma}^{-1} - H)^{-1}, \tag{A.6}$$

which is obtained by solving the first-order optimality condition. By placing this optimal solution into the objective function and arranging terms, we have

$$\sup_{\Sigma\in\mathcal{S}_\mu} \operatorname{Tr}\left[H\Sigma\right] = \inf_{\substack{\phi\geq 0\\ \phi\widehat{\Sigma}^{-1}\succ H}} \phi\big(\rho-(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu})\big) - \phi\log\det(I-\widehat{\Sigma}^{\frac{1}{2}}H\widehat{\Sigma}^{\frac{1}{2}}/\phi). \tag{A.7}$$

We now argue that the above equality also holds when μ is chosen such that $(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu})=\rho$. In this case, $\mathcal{S}_\mu$ collapses into a singleton $\{\widehat{\Sigma}\}$, and the left-hand side supremum problem attains the value $\operatorname{Tr}\left[H\widehat{\Sigma}\right]$. The right-hand side infimum problem becomes

$$\inf_{\substack{\phi\geq 0\\ \phi\widehat{\Sigma}^{-1}\succ H}} -\phi\log\det(I-\widehat{\Sigma}^{\frac{1}{2}}H\widehat{\Sigma}^{\frac{1}{2}}/\phi).$$

One can show using the l'Hopital rule that

$$\lim_{\phi\uparrow+\infty} -\phi\log\det(I-\widehat{\Sigma}^{\frac{1}{2}}H\widehat{\Sigma}^{\frac{1}{2}}/\phi) = \operatorname{Tr}\left[H\widehat{\Sigma}\right],$$

which implies that the equality holds. Furthermore, when $H=0$, the left-hand side of (A.7) evaluates to 0, while the infimum problem on the right-hand side of (A.7) also attains the optimal value of 0 asymptotically as ϕ decreases to 0. This implies that (A.7) holds for all $H\in\mathbb{S}^p$ and for any μ satisfying $(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu})\leq\rho$.

The above line of argument shows that problem (A.5a) can now be expressed as the following maximin problem

$$\sup_{\mu:(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu})\leq\rho} \inf_{\substack{\phi\geq 0\\ \phi\widehat{\Sigma}^{-1}\succ H}} \mu^\top H\mu + \phi\big(\rho-(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu})\big) - \phi\log\det(I-\widehat{\Sigma}^{\frac{1}{2}}H\widehat{\Sigma}^{\frac{1}{2}}/\phi).$$

For any $\phi\geq 0$ such that $\phi\widehat{\Sigma}^{-1}\succ H$, the objective function is concave in μ. For any μ, the objective function is convex in ϕ. Furthermore, the feasible set of μ is convex and compact, and the feasible set of ϕ is convex. As a consequence, we can apply Sion's minimax theorem (Sion, 1958) to interchange the supremum and the infimum operators, and problem (A.5a) is equivalent to

$$\inf_{\substack{\phi\geq 0\\ \phi\widehat{\Sigma}^{-1}\succ H}} \left\{ \begin{array}{l} \phi\rho - \phi\log\det(I-\widehat{\Sigma}^{\frac{1}{2}}H\widehat{\Sigma}^{\frac{1}{2}}/\phi) \\ \qquad + \sup\limits_{\mu:(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu})\leq\rho} \mu^\top H\mu - \phi(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu}) \end{array} \right\}.$$

For any ϕ which is feasible for the outer problem, the inner supremum problem is a convex quadratic optimization problem because $\phi\widehat{\Sigma}^{-1}\succ H$. Using a strong duality argument, the value of the inner supremum equals to the value of

$$\begin{aligned} &\inf_{\nu\geq 0} \left\{\nu\rho-(\nu+\phi)\widehat{\mu}^\top\widehat{\Sigma}^{-1}\widehat{\mu} + \sup_{\mu}\ \mu^\top(H-(\phi+\nu)\widehat{\Sigma}^{-1})\mu + 2(\nu+\phi)(\widehat{\Sigma}^{-1}\widehat{\mu})^\top\mu\right\} \\ &= \inf_{\nu\geq 0}\ \nu\rho-(\nu+\phi)\widehat{\mu}^\top\widehat{\Sigma}^{-1}\widehat{\mu} + (\nu+\phi)^2(\widehat{\Sigma}^{-1}\widehat{\mu})^\top[(\phi+\nu)\widehat{\Sigma}^{-1}-H]^{-1}(\widehat{\Sigma}^{-1}\widehat{\mu}), \end{aligned}$$

where the equality follows from the fact that the unique optimal solution in the variable μ is given by

$$(\phi+\nu)[(\phi+\nu)\widehat{\Sigma}^{-1}-H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu}. \tag{A.8}$$

By combining two layers of infimum problem and using a change of variables $\kappa\leftarrow\phi+\nu$, problem (A.5a) can now be written as

$$\left\{ \begin{array}{ll} \inf & \kappa(\rho-\widehat{\mu}^\top\widehat{\Sigma}^{-1}\widehat{\mu}) + \kappa^2\widehat{\mu}^\top\widehat{\Sigma}^{-1}[\kappa\widehat{\Sigma}^{-1}-H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu} - \phi\log\det(I-\widehat{\Sigma}^{\frac{1}{2}}H\widehat{\Sigma}^{\frac{1}{2}}/\phi) \\ \text{s.t.} & \phi\geq 0,\ \phi\widehat{\Sigma}^{-1}\succ H,\ \kappa-\phi\geq 0. \end{array} \right. \tag{A.9}$$

We now proceed to eliminate the multiplier ϕ from the above problem. To this end, rewrite the above optimization problem as

$$\begin{array}{ll} \inf & \kappa(\rho-\widehat{\mu}^\top\widehat{\Sigma}^{-1}\widehat{\mu}) + \kappa^2\widehat{\mu}^\top\widehat{\Sigma}^{-1}[\kappa\widehat{\Sigma}^{-1}-H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu} + g(\kappa) \\ \text{s.t.} & \kappa\geq 0,\ \kappa\widehat{\Sigma}^{-1}\succ H, \end{array}$$

where $g(\kappa)$ is defined for every feasible value of κ as

$$g(\kappa) \triangleq \left\{ \begin{array}{ll} \inf & -\phi \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi) \\ \text{s.t.} & \phi \geq 0,\ \phi\widehat{\Sigma}^{-1} \succ H,\ \phi \leq \kappa. \end{array} \right. \tag{A.10}$$

Let $g_0(\phi)$ denote the objective function of the above optimization, which is independent of κ. Let $\sigma_1, \ldots, \sigma_p$ be the eigenvalues of $\widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}$, we can write the function g directly using the eigenvalues $\sigma_1, \ldots, \sigma_p$ as

$$g_0(\phi) = -\phi \sum_{i=1}^{p} \log(1 - \sigma_i/\phi).$$

It is easy to verify by basic algebra manipulation that the gradient of g_0 satisfies

$$\nabla g_0(\phi) = \sum_{i=1}^{p} \left[\log\left(\frac{\phi}{\phi - \sigma_i}\right) - \frac{\phi}{\phi - \sigma_i} \right] + p \leq 0,$$

which implies that the value of ϕ that solves (A.10) is κ, and thus $g(\kappa) = -\kappa \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\kappa)$. Substituting ϕ by κ in problem (A.9) leads to the desired claim. □

Lemma A.2 (Optimal solution attaining $f(\beta)$)**.** *For any $(\widehat{\mu}, \widehat{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}_{++}^p$, $\rho \in \mathbb{R}_{++}$ and $w \in \mathbb{R}^p$, $f(\beta)$ equals to the optimal value of the optimization problem*

$$\left\{ \begin{array}{ll} \sup\limits_{\mu, \Sigma \succ 0} & w^\top (\Sigma + \mu\mu^\top) w \\ \text{s.t.} & \mathrm{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right] - \log\det(\Sigma\widehat{\Sigma}^{-1}) - p + (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) \leq \rho, \end{array} \right. \tag{A.11a}$$

which admits the unique optimal solution

$$\Sigma^\star = \kappa^\star(\kappa^\star\widehat{\Sigma}^{-1} - ww^\top)^{-1}, \qquad \mu^\star = \Sigma^\star\widehat{\Sigma}^{-1}\widehat{\mu}, \tag{A.11b}$$

with $\kappa^\star > w^\top\widehat{\Sigma}w$ being the unique solution of the nonlinear equation

$$\rho = \frac{(w^\top\widehat{\mu})^2 w^\top\widehat{\Sigma}w}{(\kappa - w^\top\widehat{\Sigma}w)^2} + \frac{w^\top\widehat{\Sigma}w}{\kappa - w^\top\widehat{\Sigma}w} + \log\Big(1 - \frac{w^\top\widehat{\Sigma}w}{\kappa}\Big). \tag{A.11c}$$

Moreover, we have $\kappa^\star \leq w^\top\widehat{\Sigma}w\big(1 + 2\rho + \sqrt{1 + 4\rho(w^\top\widehat{\mu})^2}\big)/(2\rho)$.

Proof of Lemma A.2. First, note that

$$f(\beta) = \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}\left[(\beta^\top X - Y)^2\right] = \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}\left[w^\top \xi\xi^\top w\right] = \sup_{(\mu,\Sigma) \in \mathbb{U}} w^\top\left(\Sigma + \mu\mu^\top\right) w,$$

which, by the definition of $\mathbb{U}$ and definition (3.2), equals to the optimal value of problem (A.11a).

From the duality result in Lemma A.1, problem (A.11a) is equivalent to

$$\begin{array}{ll} \inf & \kappa(\rho - \widehat{\mu}^\top\widehat{\Sigma}^{-1}\widehat{\mu}) + (\kappa\widehat{\Sigma}^{-1}\widehat{\mu})^\top[\kappa\widehat{\Sigma}^{-1} - ww^\top]^{-1}(\kappa\widehat{\Sigma}^{-1}\widehat{\mu}) - \kappa\log\det(I - \widehat{\Sigma}^{\frac{1}{2}} ww^\top \widehat{\Sigma}^{\frac{1}{2}}/\kappa) \\ \text{s.t.} & \kappa \geq 0,\ \kappa\widehat{\Sigma}^{-1} \succ ww^\top. \end{array}$$

Applying Bernstein (2009, Fact 2.16.3), we have the equalities

$$\begin{aligned} \det(I - \widehat{\Sigma}^{\frac{1}{2}} ww^\top \widehat{\Sigma}^{\frac{1}{2}}/\kappa) &= 1 - w^\top\widehat{\Sigma}w/\kappa \\ (\kappa\widehat{\Sigma}^{-1} - ww^\top)^{-1} &= \kappa^{-1}\widehat{\Sigma} + \kappa^{-2}\big(1 - w^\top\widehat{\Sigma}w/\kappa\big)^{-1}\widehat{\Sigma}ww^\top\widehat{\Sigma}, \end{aligned}$$

and thus by some algebraic manipulations we can rewrite

$$f(\beta) = \left\{ \begin{array}{ll} \inf & \kappa\rho + \frac{\kappa(w^\top\widehat{\mu})^2}{\kappa - w^\top\widehat{\Sigma}w} - \kappa\log\left(1 - w^\top\widehat{\Sigma}w/\kappa\right) \\ \text{s.t.} & \kappa > w^\top\widehat{\Sigma}w. \end{array} \right. \tag{A.12}$$

Let f_0 be the objective function of the above optimization problem. The gradient of f_0 satisfies

$$\nabla f_0(\kappa) = \rho - \frac{(w^\top \widehat{\mu})^2 w^\top \widehat{\Sigma} w}{(\kappa - w^\top \widehat{\Sigma} w)^2} - \frac{w^\top \widehat{\Sigma} w}{\kappa - w^\top \widehat{\Sigma} w} - \log\Big(1 - \frac{w^\top \widehat{\Sigma} w}{\kappa}\Big).$$

By the above expression of $\nabla f_0(\kappa)$ and the strict convexity of $f_0(\kappa)$, the value $\kappa^\star$ that solves (A.11c) is also the unique minimizer of (A.12). In other words, $f_0(\kappa) = f(\beta)$.

We now proceed to show that $(\mu^\star, \Sigma^\star)$ defined as in (A.11b) is feasible and optimal. First, we prove feasibility of $(\mu^\star, \Sigma^\star)$. By direct computation,

$$(\mu^\star - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (\mu^\star - \widehat{\mu}) = \widehat{\mu}^\top (\widehat{\Sigma}^{-1}\Sigma^\star - I)\widehat{\Sigma}^{-1}(\Sigma^\star \widehat{\Sigma}^{-1} - I)\widehat{\mu} = \frac{(\widehat{\mu}^\top w)^2 w^\top \widehat{\Sigma} w}{(\kappa^\star - w^\top \widehat{\Sigma} w)^2}. \tag{A.13a}$$

Moreover, because $\Sigma^\star \widehat{\Sigma}^{-1} = I + (\kappa^\star - w^\top \widehat{\Sigma} w)^{-1} \widehat{\Sigma} w w^\top$, we have

$$\mathrm{Tr}\left[\Sigma^\star \widehat{\Sigma}^{-1}\right] - \log\det(\Sigma^\star \widehat{\Sigma}^{-1}) - p = (\kappa^\star - w^\top \widehat{\Sigma} w)^{-1} w^\top \widehat{\Sigma} w + \log\big(1 - \frac{w^\top \widehat{\Sigma} w}{\kappa^\star}\big). \tag{A.13b}$$

Combining (A.13a) and (A.13b), we have

$$\mathrm{Tr}\left[\Sigma^\star \widehat{\Sigma}^{-1}\right] - \log\det(\Sigma^\star \widehat{\Sigma}^{-1}) - p + (\mu^\star - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu^\star - \widehat{\mu}) = \rho,$$

where the first equality follows from the definition of $\mathbb{D}$, and the second equality follows from the fact that $\kappa^\star$ solves (A.11c). This shows the feasibility of $(\mu^\star, \Sigma^\star)$.

Next, we prove the optimality of $(\mu^\star, \Sigma^\star)$. Through a tedious computation, one can show that

$$\begin{aligned}
&w^\top(\Sigma^\star + (\mu^\star)(\mu^\star)^\top) w = w^\top(\Sigma^\star + \Sigma^\star \widehat{\Sigma}^{-1} \widehat{\mu}\widehat{\mu}^\top \widehat{\Sigma}^{-1} \Sigma^\star) w \\
=& w^\top \widehat{\Sigma} w \Big(1 + \frac{w^\top \widehat{\Sigma} w}{\kappa^\star - w^\top \widehat{\Sigma} w}\Big) + (\widehat{\mu}^\top w)^2 \Big(1 + \frac{2 w^\top \widehat{\Sigma} w}{\kappa^\star - w^\top \widehat{\Sigma} w}\Big) + \frac{(w^\top \widehat{\mu})^2 (w^\top \widehat{\Sigma} w)^2}{(\kappa^\star - w^\top \widehat{\Sigma} w)^2} \\
=& \frac{\kappa^\star w^\top \widehat{\Sigma} w}{\kappa^\star - w^\top \widehat{\Sigma} w} + \frac{(\kappa^\star)^2 (\widehat{\mu}^\top w)^2}{(\kappa^\star - w^\top \widehat{\Sigma} w)^2} \\
=& \frac{\kappa^\star w^\top \widehat{\Sigma} w}{\kappa^\star - w^\top \widehat{\Sigma} w} + \frac{\kappa^\star (\widehat{\mu}^\top w)^2 w^\top \widehat{\Sigma} w}{(\kappa^\star - w^\top \widehat{\Sigma} w)^2} + \frac{\kappa^\star (\widehat{\mu}^\top w)^2}{\kappa^\star - w^\top \widehat{\Sigma} w} \\
=& \kappa^\star \rho - \kappa^\star \log\big(1 - \frac{w^\top \widehat{\Sigma} w}{\kappa^\star}\big) + \frac{\kappa^\star (\widehat{\mu}^\top w)^2}{\kappa^\star - w^\top \widehat{\Sigma} w} = f_0(\kappa^\star) = f(\beta),
\end{aligned}$$

where the antepenultimate equality follows from the fact that $\kappa^\star$ solves (A.11c), and the last equality holds because $\kappa^\star$ is the minimizer of (A.12). Therefore, $(\mu^\star, \Sigma^\star)$ is optimal to problem (A.11a). The uniqueness of $(\mu^\star, \Sigma^\star)$ now follows from the unique solution of Σ and μ with respect to the dual variables from (A.6) and (A.8), respectively.

It now remains to show the upper bound on $\kappa^\star$. Towards that end, we note that for any $\kappa > w^\top \widehat{\Sigma} w$,

$$0 = \rho - \frac{(w^\top \widehat{\mu})^2 w^\top \widehat{\Sigma} w}{(\kappa^\star - w^\top \widehat{\Sigma} w)^2} - \frac{w^\top \widehat{\Sigma} w}{\kappa^\star - w^\top \widehat{\Sigma} w} - \log\Big(1 - \frac{w^\top \widehat{\Sigma} w}{\kappa^\star}\Big) > \rho - \frac{(w^\top \widehat{\mu})^2 w^\top \widehat{\Sigma} w}{(\kappa^\star - w^\top \widehat{\Sigma} w)^2} - \frac{w^\top \widehat{\Sigma} w}{\kappa^\star - w^\top \widehat{\Sigma} w}.$$

Solving the above quadratic inequality in the variable $\kappa^\star - w^\top \widehat{\Sigma} w$ yields the desired bound. This completes the proof. □

We are now ready to prove Proposition 4.2.

Proof of Proposition 4.2. The convexity of f follows immediately by noting that it is the pointwise supremum of the family of convex functions $\mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$ parametrized by $\mathbb{Q}$.

To prove the continuously differentiability and the formula for the gradient, recall the expression (A.12) for the function $f(\beta)$:

$$f(\beta) = \begin{cases} \inf & \kappa\rho + \frac{\kappa(w^\top \widehat{\mu})^2}{\kappa - w^\top \widehat{\Sigma} w} - \kappa \log\left(1 - w^\top \widehat{\Sigma} w/\kappa\right) \\ \text{s.t.} & \kappa > w^\top \widehat{\Sigma} w. \end{cases} \tag{A.14}$$

Problem (A.14) has only one constraint. Therefore, LICQ (hence MFCQ) always holds, which implies that the Lagrange multiplier ζ_β of problem (A.14) is unique for any β. Also, it is easy to see that the constraint of problem (A.14) is never binding. So, $\zeta_\beta = 0$ for any β. The Lagrangian function $L_\beta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is given by

$$L_\beta(\kappa, \zeta) = \rho\kappa + \frac{\omega_2 \kappa}{\kappa - \omega_1} - \kappa \log\left(1 - \frac{\omega_1}{\kappa}\right) + \zeta(\omega_1 - \kappa),$$

where $\omega_1 = w^\top \widehat{\Sigma} w$ and $\omega_2 = (w^\top \widehat{\mu})^2$. The first derivative with respect to κ is

$$\frac{\mathrm{d}L_\beta}{\mathrm{d}\kappa}(\kappa, \zeta) = \rho - \frac{\omega_1 \omega_2}{(\kappa - \omega_1)^2} - \log\left(1 - \frac{\omega_1}{\kappa}\right) - \frac{\omega_1}{\kappa - \omega_1} - \zeta.$$

The second derivative with respect to κ is

$$\frac{\mathrm{d}^2 L_\beta}{\mathrm{d}\kappa^2}(\kappa, \zeta) = \frac{\omega_1}{(\kappa - \omega_1)^3}\left(2\omega_2 + \frac{\omega_1}{\kappa}(\kappa - \omega_1)\right).$$

From the proof of Lemma A.2, we have that the minimizer κ_β of problem (A.14) is precisely the $\kappa^\star$ defined by equation (A.11c) (below we write κ_β instead of $\kappa^\star$ to emphasize and keep track of the dependence on β). Therefore, for any β, the minimizer κ_β exists and is unique. So, there exists some constant $\eta_\beta > 0$ such that

$$\frac{\mathrm{d}^2 L_\beta}{\mathrm{d}\kappa^2}(\kappa_\beta, \zeta_\beta) \geq \eta_\beta > 0.$$

Therefore, for any β, the strong second order condition at κ_β holds (see Still (2018, Definition 6.2)). By Still (2018, Theorem 6.7),

$$\nabla f(\beta) = \nabla_\beta L_\beta(\kappa_\beta, \zeta_\beta) = \nabla_\beta L_\beta(\kappa_\beta, 0) \quad \forall \beta \in \mathbb{R}^d. \tag{A.15}$$

Then we compute

$$\begin{aligned} \nabla_w L_\beta(\kappa, \zeta) &= \nabla_w \left[\frac{\kappa(w^\top \widehat{\mu})^2}{\kappa - w^\top \widehat{\Sigma} w} - \kappa \log\left(1 - \frac{w^\top \widehat{\Sigma} w}{\kappa}\right) + \zeta(w^\top \widehat{\Sigma} w - \kappa)\right] \\ &= \frac{2\kappa\omega_2}{(\kappa - \omega_1)^2}\widehat{\Sigma} w + \frac{2\kappa}{(\kappa - \omega_1)}\widehat{\mu}\widehat{\mu}^\top w + \frac{2\kappa}{(\kappa - \omega_1)}\widehat{\Sigma} w + 2\zeta\widehat{\Sigma} w. \end{aligned}$$

Hence,

$$\nabla_\beta L_\beta(\kappa, \zeta) = \frac{dw}{d\beta}^\top \cdot \nabla_w L_\beta(\kappa, \zeta) = [I_d \ \mathbf{0}_d] \cdot \nabla_w L_\beta(\kappa, \zeta),$$

which, when combined with (A.15), yields the desired gradient formula

$$\nabla f(\beta) = \frac{2\kappa_\beta \left(\omega_2 \widehat{\Sigma} w + (\kappa_\beta - \omega_1)(\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top) w\right)_{1:d}}{(\kappa_\beta - \omega_1)^2}.$$

By Still (2018, Theorem 6.5), the function $\beta \mapsto \kappa_\beta$ is locally Lipschitz continuous, *i.e.*, for any $\beta \in \mathbb{R}^d$, there exists $c_\beta, \epsilon_\beta > 0$ such that if $\|\beta' - \beta\|_2 \leq \epsilon_\beta$, then

$$|\kappa_{\beta'} - \kappa_\beta| \leq c_\beta \|\beta' - \beta\|_2.$$

Note that ω_1 and ω_2 are both locally Lipschitz continuous in β. Also, it is easy to see that $\kappa_\beta > \omega_1$ for any β. Thus, $\nabla f(\beta)$ is locally Lipschitz continuous in β. □

Proof of 4.3. Noting that problem (3) is the barycenter problem between two Gaussian distributions with respect to the Wasserstein distance, the proof then directly follows from Agueh & Carlier (2011, §6.2) and McCann (1997, Example 1.7). □

Proof of Proposition 4.4. Again we omit the subscripts λ and ρ. Reminding that $\xi = (X, Y)$, we find

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] = \sup_{\mathbb{Q}\in\mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(w^\top \xi)^2] \\
&= \left\{ \begin{array}{ll} \inf & \kappa\big(\rho - \|\widehat{\mu}\|_2^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + z + \operatorname{Tr}\big[Z\big] \\ \text{s.t.} & \kappa \in \mathbb{R}_+,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}_+^p \\ & \begin{bmatrix} \kappa I - ww^\top & \kappa\widehat{\Sigma}^{\frac{1}{2}} \\ \kappa\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \kappa I - ww^\top & \kappa\widehat{\mu} \\ \kappa\widehat{\mu}^\top & z \end{bmatrix} \succeq 0 \end{array} \right. \\
&= \left\{ \begin{array}{ll} \inf & \kappa\big(\rho - \|\widehat{\mu}\|_2^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + \kappa^2\widehat{\mu}^\top(\kappa I - ww^\top)^{-1}\widehat{\mu} + \kappa^2 \operatorname{Tr}\big[\widehat{\Sigma}(\kappa I - ww^\top)^{-1}\big] \\ \text{s.t.} & \kappa \geq \|w\|_2^2, \end{array} \right.
\end{aligned} \tag{A.16}$$

where the second equality follows from Kuhn et al. (2019, Lemma 2). By applying Bernstein (2009, Fact 2.16.3), we find

$$(\kappa I - ww^\top)^{-1} = \kappa^{-1} I + \kappa^{-2}\big(1 - \|w\|_2^2/\kappa\big)^{-1} ww^\top. \tag{A.17}$$

Combining (A.16) and (A.17), we get

$$\sup_{\mathbb{Q}\in\mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] = \left\{ \begin{array}{ll} \inf & \kappa\rho + \kappa w^\top(\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top)w/(\kappa - \|w\|_2^2) \\ \text{s.t.} & \kappa \geq \|w\|_2^2. \end{array} \right.$$

One can verify through the first-order optimality condition that the optimal solution $\kappa^\star$ is

$$\kappa^\star = \|w\|_2 \left(\|w\|_2 + \sqrt{\frac{w^\top(\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top)w}{\rho}} \right),$$

and by replacing this value $\kappa^\star$ into the objective function, we find

$$\sup_{\mathbb{Q}\in\mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] = \big(\sqrt{w^\top(\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top)w} + \sqrt{\rho}\|w\|_2\big)^2,$$

which then completes the proof. □

A.2. Proof of Section 5

Lemma A.3 (Compactness). *For $k \in \{\mathrm{S}, \mathrm{T}\}$, the set*

$$\mathbb{V}_k = \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : M - \mu\mu^\top \in \mathbb{S}^p_{++}, \mathbb{D}((\mu, M - \mu\mu^\top) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)) \le \rho_k\}$$

is convex and compact. Furthermore, the set

$$\mathbb{V} \triangleq \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_\mathrm{S}, \rho_\mathrm{T}}\}$$

is also convex and compact.

Proof of Lemma A.3. For any $(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++}$ such that $M - \mu\mu^\top \in \mathbb{S}^p_{++}$, we find

$$\begin{aligned}
&\mathbb{D}\big((\mu, M - \mu\mu^\top) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)\big) \\
=&(\mu - \widehat{\mu}_k)^\top \widehat{\Sigma}_k^{-1}(\mu - \widehat{\mu}_k) + \mathrm{Tr}\left[(M - \mu\mu^\top)\widehat{\Sigma}^{-1}\right] - \log\det((M - \mu\mu^\top)\widehat{\Sigma}_k^{-1}) - p \\
=&\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\widehat{\mu}_k - 2\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\mu + \mathrm{Tr}\left[M\widehat{\Sigma}_k^{-1}\right] - \log\det(M\widehat{\Sigma}_k^{-1}) - \log(1 - \mu^\top M^{-1}\mu) - p,
\end{aligned} \tag{A.18}$$

where in the last expression, we have used the determinant formula (Bernstein, 2009, Fact 2.16.3) to rewrite

$$\det(M - \mu\mu^\top) = (1 - \mu^\top M^{-1}\mu)\det M.$$

Because $M - \mu\mu^\top \in \mathbb{S}^p_{++}$, one can show that $1 - \mu^\top M^{-1}\mu > 0$ by invoking the Schur complement, and as such, the logarithm term in the last expression is well-defined. Moreover, we can write

$$\mathbb{V}_k = \left\{ (\mu, M) : \begin{array}{l} (\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++},\ M - \mu\mu^\top \in \mathbb{S}^p_{++},\ \exists t \in \mathbb{R}_+ : \\ \widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\widehat{\mu}_k - 2\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\mu + \mathrm{Tr}\left[M\widehat{\Sigma}_k^{-1}\right] - \log\det(M\widehat{\Sigma}_k^{-1}) - \log(1 - t) - p \le \rho \\ \begin{bmatrix} M & \mu \\ \mu^\top & t \end{bmatrix} \succeq 0 \end{array} \right\}, \tag{A.19}$$

which is a convex set. Notice that by Schur complement, the semidefinite constraint is equivalent to $t \ge \mu^\top M^{-1}\mu$.

Next, we show that $\mathbb{V}_k$ is compact. Denote by $\mathbb{U}_k = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}^p_+ : \mathbb{D}((\mu, \Sigma) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)) \le \rho_k\}$. Then, it is easy to see that $\mathbb{V}_k$ is the image of $\mathbb{U}_k$ under the continuous mapping $(\mu, \Sigma) \mapsto (\mu, \Sigma + \mu\mu^\top)$. Therefore, it suffices to prove the compactness of $\mathbb{U}_k$. Towards that end, we note that

$$\mathbb{D}\big((\mu, \Sigma) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)\big) = (\widehat{\mu}_k - \mu)^\top \widehat{\Sigma}_k^{-1}(\widehat{\mu}_k - \mu) + \mathrm{Tr}\left[\Sigma\widehat{\Sigma}_k^{-1}\right] - \log\det(\Sigma\widehat{\Sigma}_k^{-1}) - p$$

is a continuous and coercive function in (μ, Σ). Thus, as a level set of $\mathbb{D}\big((\mu, \Sigma) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)\big)$, $\mathbb{U}_k$ is closed and bounded, and hence compact.

To prove the last claim, by the definitions of $\mathbb{V}$ and $\mathbb{U}_{\rho_\mathrm{S}, \rho_\mathrm{T}}$ we write

$$\begin{aligned}
&\mathbb{V} = \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_\mathrm{S}, \rho_\mathrm{T}}\} \\
=&\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M) \in \mathbb{V}_\mathrm{S}\} \cap \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M) \in \mathbb{V}_\mathrm{T}\} \cap \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : M \succeq \varepsilon I\}.
\end{aligned} \tag{A.20}$$

The convexity of $\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_\mathrm{S}, \rho_\mathrm{T}}\}$ then follows from the convexity of the three sets in (A.20). Furthermore, from the first part of the proof, we know that both $\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M) \in \mathbb{V}_\mathrm{S}\}$ and $\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M) \in \mathbb{V}_\mathrm{T}\}$ are compact sets, so is their intersection. Also, the last set $\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : M \succeq \varepsilon I\}$ in (A.20) is closed. Since any closed subset of a compact set is again compact, we conclude that $\mathbb{V}$ is compact. This completes the proof.

□

Proof of Theorem 5.2. As $\xi = (X, Y)$, we can rewrite

$$\min_{\beta\in\mathbb{R}^d} \sup_{\mathbb{Q}\in\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] \tag{A.21a}$$

$$= \min_{\beta\in\mathbb{R}^d} \sup_{\mathbb{Q}\in\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}} \begin{bmatrix}\beta\\-1\end{bmatrix}^\top \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top] \begin{bmatrix}\beta\\-1\end{bmatrix} \tag{A.21b}$$

$$= \min_{\beta\in\mathbb{R}^d} \sup_{(\mu, M-\mu\mu^\top)\in\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}} \begin{bmatrix}\beta\\-1\end{bmatrix}^\top M \begin{bmatrix}\beta\\-1\end{bmatrix}$$

$$= \min_{\beta\in\mathbb{R}^d} \sup_{(\mu, M)\in\mathbb{V}} \begin{bmatrix}\beta\\-1\end{bmatrix}^\top M \begin{bmatrix}\beta\\-1\end{bmatrix}$$

$$= \sup_{(\mu, M)\in\mathbb{V}} \min_{\beta\in\mathbb{R}^d} \begin{bmatrix}\beta\\-1\end{bmatrix}^\top M \begin{bmatrix}\beta\\-1\end{bmatrix} \tag{A.21c}$$

$$= \sup_{(\mu, M)\in\mathbb{V}} M_{YY} - M_{XY}^\top M_{XX}^{-1} M_{XY} \tag{A.21d}$$

where (A.21c) follows from the Sion's minimax theorem, which holds because the objective function is convex in β, concave in M, and Lemma A.3. Equation (A.21d) exploits the unique optimal solution in β as $\beta^\star = M_{XX}^{-1} M_{XY}$, in which the matrix inverse is well defined because $M \succ 0$ for any feasible M.

Finally, after an application of the Schur complement reformulation to (A.21d), the nonlinear semidefinite program in the theorem statement follows from representations (A.19) and (A.20). This completes the proof. □

Proof of Proposition 5.3. It is well-known that the space of probability measures equipped with the Wasserstein distance W_2 is a geodesic metric space (see Villani (2008, Section 7) for example), meaning that for any two probability distributions $\mathcal{N}_0$ and $\mathcal{N}_1$, there exists a constant-speed geodesic curve $[0, 1] \ni a \mapsto \mathcal{N}_a$ satisfying

$$W_2(\mathcal{N}_a, \mathcal{N}_{a'}) = |a - a'| W_2(\mathcal{N}_0, \mathcal{N}_1) \quad \forall a, a' \in [0, 1].$$

The claim follows trivially if $W_2(\mathcal{N}_{\mathrm{S}}, \mathcal{N}_{\mathrm{T}}) \leq \sqrt{\rho_{\mathrm{S}}}$. Therefore, we assume $W_2(\mathcal{N}_{\mathrm{S}}, \mathcal{N}_{\mathrm{T}}) > \sqrt{\rho_{\mathrm{S}}}$.

Consider the the geodesic $\mathcal{N}_t$ from $\mathcal{N}_0 = \mathcal{N}_{\mathrm{S}}$ to $\mathcal{N}_1 = \mathcal{N}_{\mathrm{T}}$. Also, denote by $\mathbb{U}_k = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \mathbb{D}((\mu, \Sigma) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)) \leq \rho_k\}$ for $k \in \{\mathrm{S}, \mathrm{T}\}$. Then, $\mathbb{U}_{\mathrm{S}}$ and $\mathbb{U}_{\mathrm{T}}$ has empty intersection if and only if

$$W_2(\mathcal{N}_a, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{S}}} \Longrightarrow W_2(\mathcal{N}_a, \mathcal{N}_{\mathrm{T}}) > \sqrt{\rho_{\mathrm{T}}} \quad \forall a \in [0, 1],$$

which is in turn equivalent to

$$a W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{S}}} \Longrightarrow (1 - a) W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{T}}} \quad \forall a \in [0, 1].$$

Picking $a = \frac{\sqrt{\rho_{\mathrm{S}}}}{W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}})} \in (0, 1)$, then we have

$$\left(1 - \frac{\sqrt{\rho_{\mathrm{S}}}}{W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}})}\right) W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{T}}}.$$

The above inequality can be rewritten as

$$W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{S}}} + \sqrt{\rho_{\mathrm{T}}},$$

which contradicts with our supposition

$$\rho_{\mathrm{T}} \geq \left(\sqrt{\mathbb{W}((\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}) \parallel (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))} - \sqrt{\rho_{\mathrm{S}}}\right)^2.$$

Thus, $\mathbb{U}_{\mathrm{S}}$ and $\mathbb{U}_{\mathrm{T}}$ has non-empty intersection. □

Proof of Theorem 5.4. As $\xi = (X, Y)$, we can rewrite

$$\min_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] \tag{A.22a}$$

$$= \min_{\beta \in \mathbb{R}^d} \sup_{(\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix}$$

$$= \sup_{(\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} \min_{\beta \in \mathbb{R}^d} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} \tag{A.22b}$$

$$= \sup_{(\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} M_{YY} - M_{XY}^\top M_{XX}^{-1} M_{XY} \tag{A.22c}$$

where (A.22b) follows from the Sion's minimax theorem, which holds because the objective function is convex in β, concave in M, and the set $\mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}$ is compact (Shafieezadeh-Abadeh et al., 2018, Lemma A.6). Equation (A.22c) exploits the unique optimal solution in β as $\beta^\star = M_{XX}^{-1} M_{XY}$, in which the matrix inverse is well defined because $M - \mu\mu^\top \succeq \varepsilon I$ for any feasible M. □

B. Additional Numerical Results

In the following the details of the datasets used in Section 6 are presented.

- **Uber&Lyft**[4] has $N_{\mathrm{S}} = 5000$ instances in the source domain and 5000 available samples in the target domain.
- **US Births (2018)**[5] has $N_{\mathrm{S}} = 5172$ samples in the source domain and 4828 available samples in the target domain.
- **Life Expectancy**[6] has $N_{\mathrm{S}} = 1407$ instances in the source domain and 242 available samples in the target domain.
- **House Prices in King County**[7] has $N_{\mathrm{S}} = 543$ instances in the source domain and 334 available samples in the target domain.
- **California Housing Prices**[8] has $N_{\mathrm{S}} = 9034$ instances in the source domain, and 6496 available instances in the target domain.

Figure A.5 demonstrates how the average cumulative loss in (1) grows over time for the US Births (2018), Life Expectancy, House Prices in KC and California Housing datasets. The results suggest that the IR-WASS and SI-WASS experts perform favorably over the competitors in that their cumulative loss at each time step is lower than that of most other competitors.

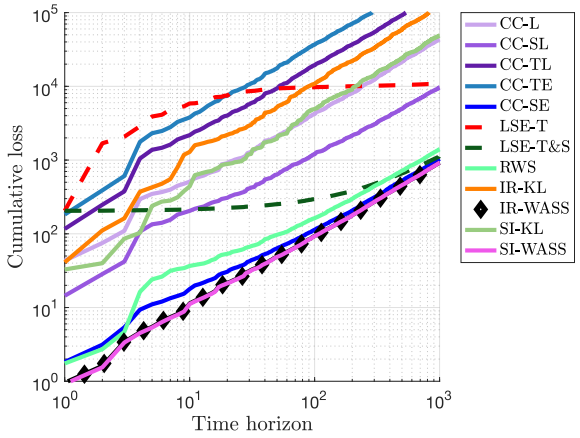
[4] Available publicly at https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma

[5] Available publicly at https://www.kaggle.com/des137/us-births-2018

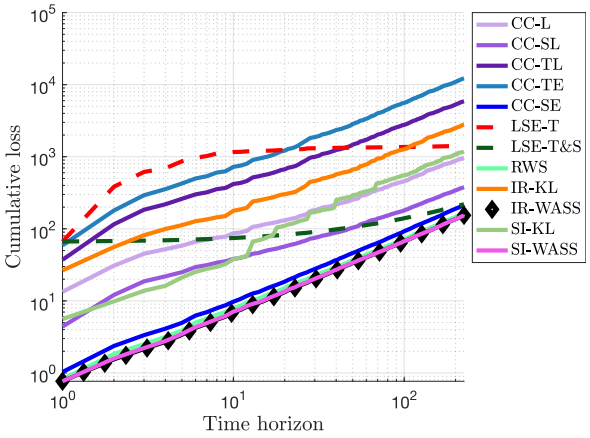
[6] Available publicly at https://www.kaggle.com/kumarajarshi/life-expectancy-who

[7] Available publicly at https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

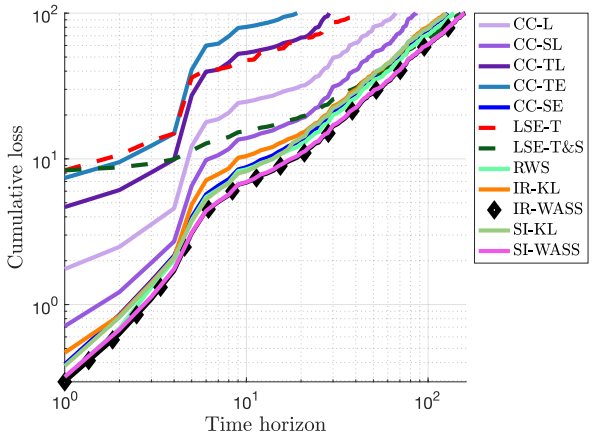
[8] The modified version that we use is available publicly at https://www.kaggle.com/camnugent/california-housing-prices and the original dataset is available publicly at https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html



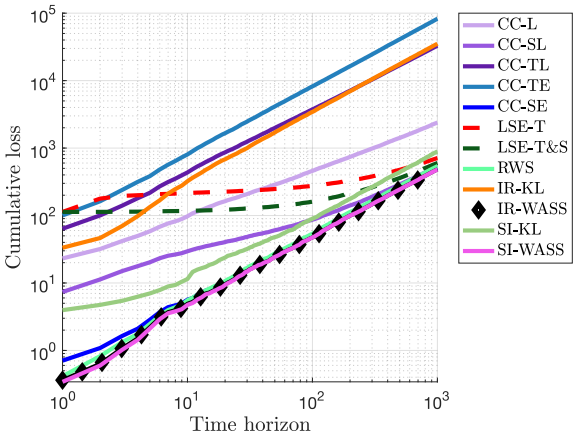
(a) US Births (2018)



(b) Life Expectancy



(c) House Prices in KC



(d) California Housing

Figure A.5. Cumulative loss averaged over 100 runs on logarithmic scale