Appendix
Sequential Domain Adaptation
by Synthesizing Distributionally Robust Experts

A. Appendix

A.1. Proofs 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*

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

admits the dual formulation

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

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.1a) can be re-expressed as a nested program of the form

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

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

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

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

$$\Sigma^\star(\phi) = \phi(\phi\widehat{\Sigma}^{-1} - H)^{-1}, \tag{A.2}$$

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

$$\sup_{\Sigma \in \mathcal{S}_\mu} \ \operatorname{Tr}\big[H\Sigma\big] = \inf_{\substack{\phi \geq 0 \\ \phi\widehat{\Sigma}^{-1} \succ H}} \ \phi\big(\rho - (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu})\big) - \phi \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi). \tag{A.3}$$

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 $\operatorname{Tr}\big[H\widehat{\Sigma}\big]$. 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

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

which implies that the equality holds. Furthermore, when $H = 0$, the left-hand side of (A.3) evaluates to 0, while the infimum problem on the right-hand side of (A.3) also attains the optimal value of 0 asymptotically as ϕ decreases to 0. This implies that (A.3) 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.1a) 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.1a) is equivalent to

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

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

$$(\phi+\nu)[(\phi+\nu)\widehat{\Sigma}^{-1}-H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu}. \tag{A.4}$$

By combining two layers of infimum problem and using a change of variables $\kappa \leftarrow \phi+\nu$, problem (A.1a) can now be written as

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

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

$$g(\kappa)\triangleq\left\{\begin{array}{ll} \inf & -\phi\log\det(I-\widehat{\Sigma}^{\frac{1}{2}}H\widehat{\Sigma}^{\frac{1}{2}}/\phi) \\ \text{s.t.} & \phi\geq 0,\ \phi\widehat{\Sigma}^{-1}\succ H,\ \phi\leq\kappa. \end{array}\right. \tag{A.6}$$

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.6) 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.5) 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*

$$\left\{\begin{array}{ll} \sup\limits_{\mu,\Sigma\succ 0} & w^\top(\Sigma+\mu\mu^\top)w \\ \text{s.t.} & \operatorname{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right]-\log\det(\Sigma\widehat{\Sigma}^{-1})-p+(\mu-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu-\widehat{\mu})\leq\rho, \end{array}\right. \tag{A.7a}$$

which admits the unique optimal solution

$$\Sigma^\star=\kappa^\star(\kappa^\star\widehat{\Sigma}^{-1}-ww^\top)^{-1},\qquad \mu^\star=\Sigma^\star\widehat{\Sigma}^{-1}\widehat{\mu}, \tag{A.7b}$$

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

$$\rho=\frac{(w^\top\widehat{\mu})^2w^\top\widehat{\Sigma}w}{(\kappa-w^\top\widehat{\Sigma}w)^2}+\frac{w^\top\widehat{\Sigma}w}{\kappa-w^\top\widehat{\Sigma}w}+\log\Big(1-\frac{w^\top\widehat{\Sigma}w}{\kappa}\Big). \tag{A.7c}$$

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.7a).

From the duality result in Lemma A.1, problem (A.7a) is equivalent to

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

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

$$f(\beta) = \begin{cases} \inf & \kappa\rho + \frac{\kappa(w^\top\widehat{\mu})^2}{\kappa - w^\top\widehat{\Sigma}w} - \kappa\log\big(1 - w^\top\widehat{\Sigma}w/\kappa\big) \\ \mathrm{s.\,t.} & \kappa > w^\top\widehat{\Sigma}w. \end{cases} \tag{A.8}$$

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.7c) is also the unique minimizer of (A.8). In other words, $f_0(\kappa) = f(\beta)$.

We now proceed to show that $(\mu^\star, \Sigma^\star)$ defined as in (A.7b) is feasible and optimal. First, we prove feasibility of $(\mu^\star, \Sigma^\star)$. By direct computation,

$$(\mu^\star - \widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu^\star - \widehat{\mu}) = \widehat{\mu}^\top(\widehat{\Sigma}^{-1}\Sigma^\star - I)\widehat{\Sigma}^{-1}(\Sigma^\star\widehat{\Sigma}^{-1} - I)\widehat{\mu} = \frac{(\widehat{\mu}^\top w)^2 w^\top\widehat{\Sigma}w}{(\kappa^\star - w^\top\widehat{\Sigma}w)^2}. \tag{A.9a}$$

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

$$\mathrm{Tr}\left[\Sigma^\star\widehat{\Sigma}^{-1}\right] - \log\det(\Sigma^\star\widehat{\Sigma}^{-1}) - p = (\kappa^\star - w^\top\widehat{\Sigma}w)^{-1}w^\top\widehat{\Sigma}w + \log\big(1 - \frac{w^\top\widehat{\Sigma}w}{\kappa^\star}\big). \tag{A.9b}$$

Combining (A.9a) and (A.9b), 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.7c). 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

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

where the antepenultimate equality follows from the fact that $\kappa^\star$ solves (A.7c), and the last equality holds because $\kappa^\star$ is the minimizer of (A.8). Therefore, $(\mu^\star, \Sigma^\star)$ is optimal to problem (A.7a). 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.2) and (A.4), respectively.

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

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

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.8) for the function $f(\beta)$:

$$f(\beta) = \left\{ \begin{array}{ll} \inf & \kappa\rho + \frac{\kappa (w^\top \widehat{\mu})^2}{\kappa - w^\top \widehat{\Sigma} w} - \kappa \log\left(1 - w^\top \widehat{\Sigma} w / \kappa\right) \\ \text{s. t.} & \kappa > w^\top \widehat{\Sigma} w. \end{array} \right. \tag{A.10}$$

Problem (A.10) has only one constraint. Therefore, LICQ (hence MFCQ) always holds, which implies that the Lagrange multiplier ζ_β of problem (A.10) is unique for any β. Also, it is easy to see that the constraint of problem (A.10) 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.10) is precisely the $\kappa^\star$ defined by equation (A.7c) (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),

$$\nabla f(\beta) = \nabla_\beta L_\beta(\kappa_\beta, \zeta_\beta) = \nabla_\beta L_\beta(\kappa_\beta, 0) \quad \forall \beta \in \mathbb{R}^d. \tag{A.11}$$

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.11), 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

$$\begin{aligned}
&\sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] = \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(w^\top \xi)^2] \\
&= \left\{ \begin{array}{ll} \inf & \kappa\big(\rho - \|\widehat{\mu}\|_2^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\big) + z + \mathrm{Tr}\left[Z\right] \\ \mathrm{s.\,t.} & \kappa \in \mathbb{R}_+, \; z \in \mathbb{R}_+, \; Z \in \mathbb{S}_+^p \\ & \begin{bmatrix} \kappa I - ww^\top & \kappa \widehat{\Sigma}^{\frac{1}{2}} \\ \kappa \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0, \; \begin{bmatrix} \kappa I - ww^\top & \kappa \widehat{\mu} \\ \kappa \widehat{\mu}^\top & z \end{bmatrix} \succeq 0 \end{array} \right. \\
&= \left\{ \begin{array}{ll} \inf & \kappa\big(\rho - \|\widehat{\mu}\|_2^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\big) + \kappa^2 \widehat{\mu}^\top (\kappa I - ww^\top)^{-1} \widehat{\mu} + \kappa^2 \,\mathrm{Tr}\left[\widehat{\Sigma}(\kappa I - ww^\top)^{-1}\right] \\ \mathrm{s.\,t.} & \kappa \geq \|w\|_2^2, \end{array} \right.
\end{aligned} \tag{A.12}$$

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

$$(\kappa I - ww^\top)^{-1} = \kappa^{-1} I + \kappa^{-2}\big(1 - \|w\|_2^2/\kappa\big)^{-1} ww^\top. \tag{A.13}$$

Combining (A.12) and (A.13), we get

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

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. Proofs of Section 5

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

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

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

$$\begin{aligned}
&\mathbb{D}\big((\mu, M - \mu\mu^\top) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)\big) \\
=&(\mu - \widehat{\mu}_k)^\top \widehat{\Sigma}_k^{-1} (\mu - \widehat{\mu}_k) + \mathrm{Tr}\left[(M - \mu\mu^\top)\widehat{\Sigma}^{-1}\right] - \log\det((M - \mu\mu^\top)\widehat{\Sigma}_k^{-1}) - p \\
=&\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1} \widehat{\mu}_k - 2\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1} \mu + \mathrm{Tr}\left[M\widehat{\Sigma}_k^{-1}\right] - \log\det(M\widehat{\Sigma}_k^{-1}) - \log(1 - \mu^\top M^{-1}\mu) - p,
\end{aligned} \tag{A.14}$$

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

$$\mathbb{V}_k = \left\{ (\mu, M) : \begin{array}{l} (\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++},\ M - \mu\mu^\top \in \mathbb{S}^p_{++},\ \exists t \in \mathbb{R}_+ : \\ \widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1} \widehat{\mu}_k - 2\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1} \mu + \mathrm{Tr}\left[M\widehat{\Sigma}_k^{-1}\right] - \log\det(M\widehat{\Sigma}_k^{-1}) - \log(1 - t) - p \leq \rho \\ \begin{bmatrix} M & \mu \\ \mu^\top & t \end{bmatrix} \succeq 0 \end{array} \right\}, \tag{A.15}$$

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

Next, we show that $\mathbb{V}_k$ is compact. Denote by $\mathbb{U}_k = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}^p_+ : \mathbb{D}((\mu, \Sigma) \,\|\, (\widehat{\mu}_k, \widehat{\Sigma}_k)) \leq \rho_k\}$. 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

$$\begin{aligned}
&\mathbb{V} = \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}\} \\
=&\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M) \in \mathbb{V}_{\mathrm{S}}\} \cap \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M) \in \mathbb{V}_{\mathrm{T}}\} \cap \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : M \succeq \varepsilon I\}.
\end{aligned} \tag{A.16}$$

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.16). 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.16) 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

$$\begin{aligned}
&\min_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] \\
&= \min_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top \mathbb{E}_{\mathbb{Q}}[\xi \xi^\top] \begin{bmatrix} \beta \\ -1 \end{bmatrix} \\
&= \min_{\beta \in \mathbb{R}^d} \sup_{(\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} \\
&= \min_{\beta \in \mathbb{R}^d} \sup_{(\mu, M) \in \mathbb{V}} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} \\
&= \sup_{(\mu, M) \in \mathbb{V}} \min_{\beta \in \mathbb{R}^d} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} && \text{(A.17a)} \\
&= \sup_{(\mu, M) \in \mathbb{V}} M_{YY} - M_{XY}^\top M_{XX}^{-1} M_{XY}, && \text{(A.17b)}
\end{aligned}$$

where (A.17a) 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.17b) 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.17b), the nonlinear semidefinite program in the theorem statement follows from representations (A.15) and (A.16). This completes the proof. □

Proof of Proposition 5.3. It is well-known that the space of probability measures equipped with the type-2 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 $\mathbb{U}_{\mathrm{T}}$ have a non-empty intersection. □

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

$$\begin{aligned}
&\min_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] && \text{(A.18a)} \\
&= \min_{\beta \in \mathbb{R}^d} \sup_{(\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} \\
&= \sup_{(\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} \min_{\beta \in \mathbb{R}^d} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} && \text{(A.18b)} \\
&= \sup_{(\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}} M_{YY} - M_{XY}^\top M_{XX}^{-1} M_{XY}, && \text{(A.18c)}
\end{aligned}$$

where (A.18b) 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.18c) 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**[1] has $N_{\mathrm{S}} = 5000$ instances in the source domain and 5000 available samples in the target domain.
- **US Births (2018)**[2] has $N_{\mathrm{S}} = 5172$ samples in the source domain and 4828 available samples in the target domain.
- **Life Expectancy**[3] has $N_{\mathrm{S}} = 1407$ instances in the source domain and 242 available samples in the target domain.
- **House Prices in King County**[4] has $N_{\mathrm{S}} = 543$ instances in the source domain and 334 available samples in the target domain.
- **California Housing Prices**[5] has $N_{\mathrm{S}} = 9034$ instances in the source domain, and 6496 available instances in the target domain.

Figure A.1 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.

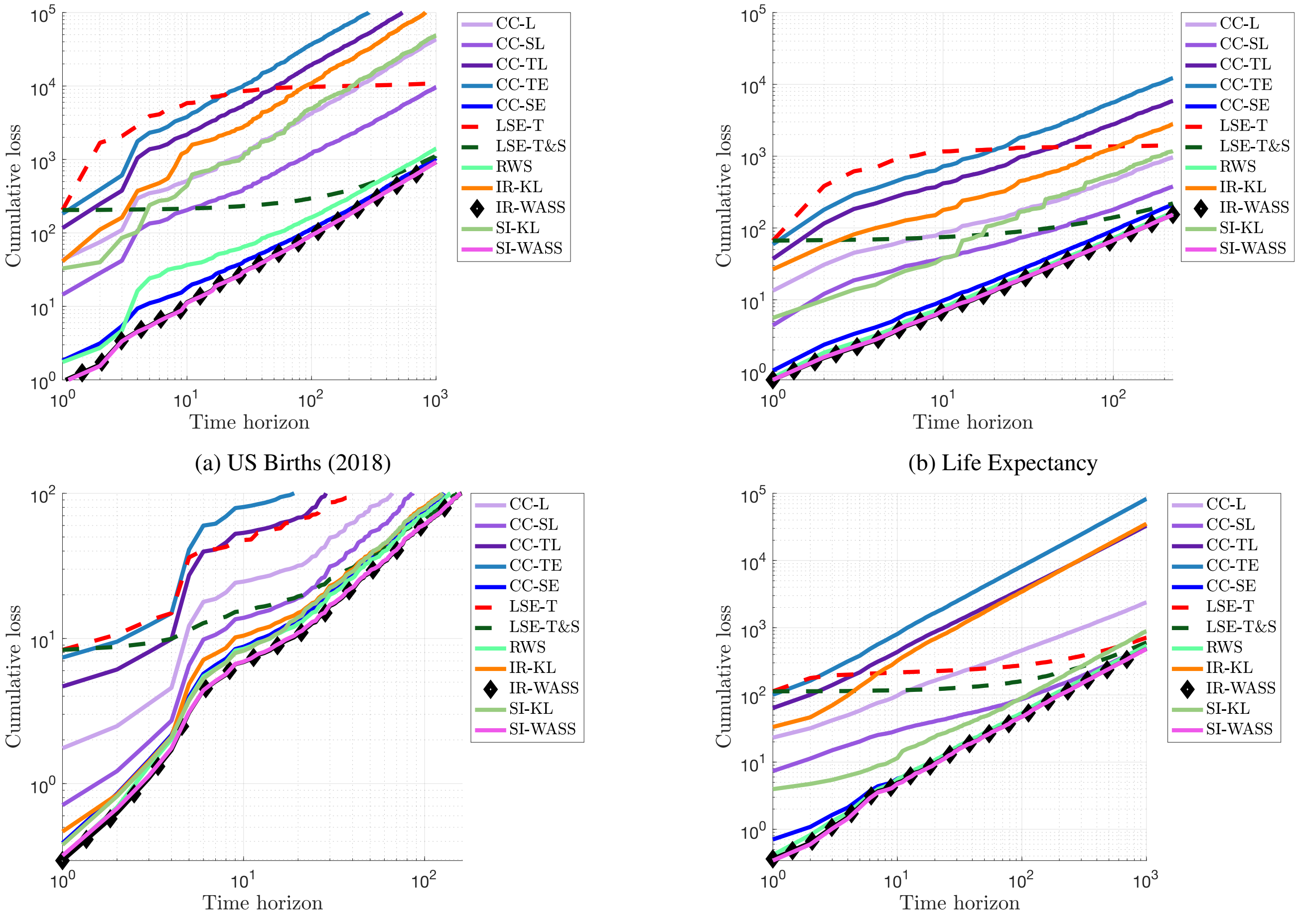
[1] Available publicly at https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma

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

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

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

[5] 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.1. Cumulative loss averaged over 100 runs on logarithmic scale

References

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

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

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

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.

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

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.

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

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

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