Learning When the Concept Shifts: Confounding, Invariance, and Dimension Reduction

Kulunu Dharmakeerthi[1], YoonHaeng Hur[1], and Tengyuan Liang[*1]

[1]University of Chicago

June 22, 2024

Abstract

Practitioners often deploy a learned prediction model in a new environment where the joint distribution of covariate and response has shifted. In observational data, the distribution shift is often driven by unobserved confounding factors lurking in the environment, with the underlying mechanism unknown. Confounding can obfuscate the definition of the best prediction model (concept shift) and shift covariates to domains yet unseen (covariate shift). Therefore, a model maximizing prediction accuracy in the source environment could suffer a significant accuracy drop in the target environment. This motivates us to study the domain adaptation problem with observational data: given labeled covariate and response pairs from a source environment, and only unlabeled covariates from a target environment, how can one predict the missing target response reliably? We root the adaptation problem in a linear structural causal model to address endogeneity and unobserved confounding. We study the necessity and benefit of leveraging exogenous, invariant covariate representations to cure concept shifts and improve target prediction. This further motivates a new data-driven representation learning method for adaptation that optimizes for a lower-dimensional linear subspace and, subsequently, a prediction model confined to that subspace. The procedure operates on a non-convex objective—that naturally interpolates between predictability and stability/invariance—constrained on the Stiefel manifold, using an analog of projected gradient descent on manifolds. We study the optimization landscape and prove that, when the regularization is sufficient, nearly all local optima align with an invariant linear subspace resilient to both the concept shift and the covariate shift. In terms of predictability, we show a predictive model that uses the learned lower-dimensional subspace can incur a nearly ideal gap between target and source risk. Three real-world data sets are investigated to validate our method and theory against distribution shifts empirically; the tradeoffs between predictability and stability are elucidated therein.

Keywords— Concept shift, distribution shift, unobserved confounding, invariance, structural causal model, representation learning.

1 Introduction

Unobserved confounding can baffle practitioners seeking valid, robust inferences and predictions when data is not collected under carefully designed experiments. Therefore, applying modern statistical learning to observational data is a cautionary tale for practitioners. Due to unobserved confounding, maximizing prediction

accuracy might be shooting at the wrong target. The problem is even more severe when the learned model is deployed to an observational data set collected from a different environment: the unobserved confounding factors can fundamentally alter the data distribution in the new environment and, thus, the concept of the best statistical model.

Gaining actionable knowledge that generalizes well to new data sets is the statistical ideal, yet theory and methodology weaken when the environment underpinning the data set shifts. A litany of works in statistical learning contributes to curing the problem of distribution shifts, notably the domain adaption literature [4, 6, 7, 5, 36, 19, 13, 3] and covariate shift literature [33, 20, 16, 18]. However, studying concept shifts due to unobserved confounding—a central topic in causality for observational data—is a less trespassed path in machine learning.

While conventional wisdom may be pessimistic about the prospect of learning across shifting environments, a growing literature in causal inference has dwelled on the utility of comparing data sets from distinct environments to identify causal relations [26, 28, 15, 3]. Consider covariate and response pair (X, Y) defined on $\mathcal{X} \times \mathcal{Y}$ and their joint probability distributions $\mathcal{P}_\mathcal{E}(X, Y)$ that differ across environments $\mathcal{E}$. Due to the presence of unobserved confounding factors in the environments, both the *covariate distribution* $\mathcal{P}_\mathcal{E}(X)$ and the *conditional concept* $\mathcal{P}_\mathcal{E}(Y|X)$ could be shifted and altered by the environment (see for instance Figure 1). Concept shift is difficult to address due to this lack of invariance. The principle of independent mechanisms in causality [27] alludes to a plausible source of invariance; the mechanism producing the effect given the cause should remain unchanged across environments. This principle can be made practical—much of the invariant risk minimization work [26, 28, 15] aims to identify a causally "stable" representation $X \mapsto \phi(X)$ such that the conditional concept $\mathcal{P}(Y|\phi(X))$ stays invariant across environments if labeled data $\mathcal{P}_\mathcal{E}(X, Y)$ across environments are available.

However, we do not have labeled data across all environments in typical domain adaptation. Consider two environments $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$ corresponding to source and target, respectively. A practitioner has access to labeled covariate-response pairs drawn from the source distribution $\mathcal{P}_\mathcal{S}(X, Y)$, but only unlabeled covariate samples from a target distribution $\mathcal{P}_\mathcal{T}(X)$. Given a representation $Z = \phi(X)$ defined on some space $\mathcal{Z}$ and a model $h : \mathcal{Z} \to \mathcal{Y}$ selected based on source data, an upper bound on target risk was derived in the seminal work [4], which involves a balancing act between the source risk of the model $h \circ \phi : \mathcal{X} \to \mathcal{Y}$, and a distance between probability distributions $\mathcal{P}_\mathcal{S}(\phi(X))$ and $\mathcal{P}_\mathcal{T}(\phi(X))$. Here, a notion of distributional "stability" arises; a representation where the differences between $\mathcal{P}_\mathcal{S}(\phi(X))$ and $\mathcal{P}_\mathcal{T}(\phi(X))$ are minimal is preferred. This notion of invariance is also natural: if the distribution shifts significantly, the model selected based on source data may suffer severely when extrapolating to target data.

In this paper, we introduce a novel approach by postulating a structural causal model to study the domain adaptation problem, addressing both concept and covariate shifts due to unobserved confounding lurking in the environments. By delving into the nature of the distribution shifts, we demonstrate that the notions of causal stability and distributional stability unite in our linear structural model. Guided by this intuition, we present a stability-regularized risk minimization methodology that aims for invariance/stability when adapting to new observational domains. We then introduce a practical first-order manifold optimization method and provide provable guarantees for nearly all local optima, effectively guarding against distribution shifts. We dissect the stability and predictability tradeoff introduced in the domain adaptation literature in a concrete linear setting where low-dimensional projections parametrize representations.

1.1 A Structural Causal Model

Throughout the paper, we postulate a Structural Causal Model (SCM) for the unobserved confounding motivated by the idea of instrumental variables [25, 34], with a crucial difference that these exogenous instruments are latent and unobserved in our model.

Definition 1 (Model). A pair of covariate and response, denoted as $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, are generated from mutually independent exogenous random variables $E \in \mathbb{R}^r, Z \in \mathbb{R}^k$ with $r, k \leq d$ and $U \in \mathbb{R}, W \in \mathbb{R}^d$ via

the following structural causal model,

$$Y = \langle \beta^\star, X \rangle + \langle \gamma, E \rangle + U \,, \tag{1.1}$$

$$X = \Theta Z + \Delta E + W \,, \tag{1.2}$$

where the parameters $\beta^\star \in \mathbb{R}^d$, $\gamma \in \mathbb{R}^r$, $\Theta \in \mathbb{R}^{d \times k}$, and $\Delta \in \mathbb{R}^{d \times r}$ are some unknown coefficients; see Figure 1.

The *unobserved confounding* variable E models the environmental influence, whose distribution shifts across settings. Concretely, in domain adaptation, the distribution of E shifts from the source distribution $\rho_{\mathcal{S}}$ to another the target distribution $\rho_{\mathcal{T}}$, where $\rho_{\mathcal{S}}, \rho_{\mathcal{T}}$ are probability distributions defined on $\mathbb{R}^r$. The *latent invariant* Z models a source of randomness that is not subject to environment shifts: a stable, exogenous structure shared across both source and target environments. It plays a similar role as an instrumental variable, with the notable difference that it is unobserved across environments. The *exogenous noise U, W* are mutually independent and mean-zero, and their distributions do not depend on the environment.
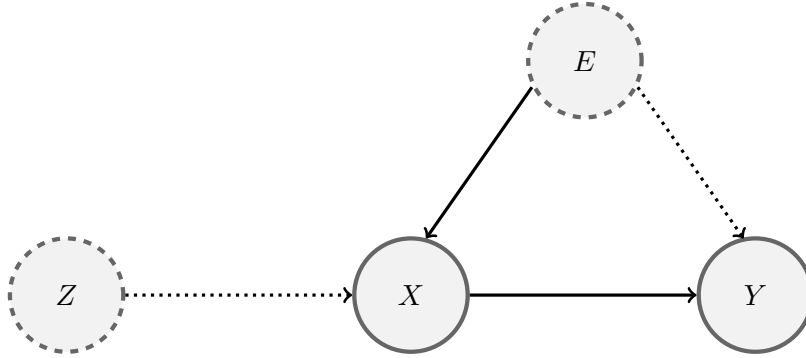

Figure 1: Diagram visualizing the model equations (1.1) and (1.2). Here, the endogenous confounding variable E lurking in the environment and the exogenous invariant variable Z are both latent and unobserved.

Observational Data and the Adaptation Problem Consider the source and target environments $\mathcal{S}$ and $\mathcal{T}$. The joint distributions of (X, Y) under $E \sim \rho_{\mathcal{S}}$ and $E \sim \rho_{\mathcal{T}}$ are denoted as $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X, Y)$, with $\mathcal{P}_{\mathcal{S}}(X)$ and $\mathcal{P}_{\mathcal{T}}(X)$ denoting the covariate distributions. We use $\mathbb{E}_{\mathcal{E}}$ to denote the expectation w.r.t. the joint distribution under the environment $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. For the expectation of random variables depending only on Z, U, W, we drop the subscript $\mathcal{E}$ as it is invariant across environments.

The practitioner observes labeled covariate-response pairs jointly drawn from the source distribution $\mathcal{P}_{\mathcal{S}}(X, Y)$, but only unlabeled covariate data from a target distribution $\mathcal{P}_{\mathcal{T}}(X)$. Given the observational data, the practitioner aims to estimate a linear model based on $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$ that performs well in the target environment.

Confounding and Concept Shift A few remarks follow for our model. First, if the dotted line connections in Figure 1 are removed, we arrive at the prototypical well-specified covariate shift setting. Environment shifts only affect the covariate distribution. In this setting, without confounding, using (weighted) least squares to estimate a model from X to Y consistently is possible. Our model incorporating the dotted lines—capturing unobserved confounding—naturally extends the covariate shift setting. Second, since the unobserved confounding variable E lurking in the environment influences both X and Y, the least squares estimate is biased and unreliable when deploying to new environments. As we shall see in Proposition 1, the distribution shift in E under this structural model induces a *concept shift*: the best linear model depends on the environment. Finally, introducing the invariant, exogenous Z may remind readers of instrumental variables; however, in our domain adaptation problem, Z is unobserved; we merely postulate its existence. We

3

will develop a methodology that leverages invariance across environments to learn a stable, lower-dimensional linear subspace without directly observing Z.

Notations We begin by defining some notations that will streamline the sequel. For a vector $v \in \mathbb{R}^d$, $\|v\|$ will denote the standard Euclidean norm and $\|v\|_M := \|M^{1/2}v\|$ for any symmetric and positive semidefinite matrix $M \in \mathbb{R}^{d \times d}$, with $M^{1/2}$ denoting its matrix root. When dealing with matrix norms, $\|\cdot\|_{\mathrm{op}}$ will refer to the spectral norm and $\|\cdot\|_{\mathrm{F}}$ will denote the Frobenius norm. The Stiefel manifold is defined as $\mathrm{St}(d, \ell) := \{A \in \mathbb{R}^{d \times \ell} : A^\top A = I_\ell\}$. Finally, we denote by $R_\mathcal{E}$ the squared risk depending on environment $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$, $R_\mathcal{E}(\beta) = \mathbb{E}_\mathcal{E}[(Y - \langle X, \beta \rangle)^2]$, where the expectation is w.r.t. the joint distribution $\mathcal{P}_\mathcal{E}(X, Y)$ as in Definition 1.

1.2 Related Literature

A recent line of research has pioneered methods of statistical reasoning for shifting environments, and placed such reasoning in a causal backdrop. Notably, the works identify that the invariance principle of causal relations can be leveraged to generalize across arbitrary interventions—a strong notion of robustness; see [31], [21] and [9]. The ontological relationship between a response and its direct causes is predictive and invariant, and can be shown to be optimal against arbitrarily strong interventions [29, 26]. An influential sequence of papers [26, 28, 15] casts distribution shifts as the action of some intervention. It compares *labeled data sets* across distinct experimental settings to identify these invariant causal conditionals. However, facing observational data, not all direct causes are observed or identifiable, and the interventions might not be diverse or strong enough thus making causal regression overly conservative. Fundamental tradeoffs should be addressed between causality and predictability. The notable work of [30] proposed an actionable methodology to probe the tradeoffs, combining instrumental variable regression (for causality) and ordinary least squares (for predictability), when *access to certain exogenous variables* acting as instruments is available. In contrast, the current paper tackles the domain adaptation problem in observational data when (i) labels are unavailable in the target environment, (ii) unobserved confounding factors corroborate environment shifts, and (iii) there is no access to exogenous, instrumental variables to construct quasi-experiments [8, 2] for causal identification; a setting less studied by the literature. Our setting inherits the quintessential difficulty of domain adaption for observational data, the unobserved confounding, and further delineates its role lurking in the environment. In spirit, our work follows and builds upon the literature on improving out-of-domain generalization in machine learning with causal insights without experimental or quasi-experimental data.

In the learning-theoretic literature, out-of-domain generalization bounds show that provided a stable representation is found under which probability distributions on source and target domain are similar, a hypothesis with low source risk will perform well on target data; see the seminal work of [4, 7, 22, 5]. This line of work also pioneered the tradeoffs between representation stability and predictability: stable representations may not harness the maximum predictive power. While methodologies that simultaneously search for predictive and stable representations of data [3, 6, 36, 19] are now ubiquitous, it is unclear under what data generating mechanisms they perform well. Furthermore, whether the practical optimization method obtains provable guarantees for the learned representation remains to be better understood [3]. We anchor the domain adaptation problem in an accessible linear SCM and delineate situations where invariance yields improvement. The methodology in this paper can be interpreted as a bare-bones instantiation of stability regularized risk minimization [13, 36, 32] in the linear SCM. The concrete relationship between stability and predictability will be teased out. We also establish provable characterizations for the learned linear subspace, a concrete step toward representation learning. As we shall see, two notions of stability/invariance unite under the linear SCM: optimizing over a *stable representation* in covariates shifts will align well with searching for the *causal, invariant relationship*, stable across environment shifts.

The problem of covariate shifts, with the conditional concept $Y|X$ unchanged, has spurred considerable interest in statistics and machine learning. The statistical study of covariate shift under parametric models dates back to the pathbreaking work of [33]. [33] established the asymptotic optimality of vanilla Maximum Likelihood Estimation (MLE) in the well-specified setting and that of weighted MLE under misspecification.

Recent non-asymptotic results bolster this result [14], and find that the weighted MLE can be minimax optimal under mis-specification. The above works operate under a mild covariate shift setting—namely, the likelihood ratio between source and target is often required to be bounded and estimable based on data. Several works explore different facets of covariate shifts specific to the linear setting. [17] considers the minimax optimal estimator for linear regression under fixed design, where the learner can access some amount of unlabeled target data. However, in the presence of concept shift near minimax optimal estimation requires target labels. [24] provides lower bounds for out-of-distribution generalization in linear and one-hidden layer neural network models. For hard covariate shifts with an unbounded likelihood ratio, [18] proposed to study adversarial covariate shifts to understand what extrapolation region adversarial covariate shifts will focus on for a given linear model. [18] derived a curious dichotomy: depending on the regression or the classification setting, adversarial covariate shift can either be a blessing or a curse to the subsequent learning. In comparison, the current paper anchors the problem in a linear SCM where confounding generates both concept and covariate shifts—diverging from the well-specified settings explored in the literature. Our method does not require the bounded likelihood ratio assumption nor the need to estimate such a ratio. Specific to the linear SCM, our paper addresses when and how access to unlabeled target samples can boost target domain performance in the presence of concept shifts.

1.3 Contributions

The paper studies domain adaptation in the presence of unobserved confounding, a likely obstacle for the practitioner attempting to extrapolate with only observational data. We extend the standard covariate shift setting to incorporate confounding factors lurking in the environment that simultaneously influence covariate X and response Y, thereby inducing concept shift across environments. Concretely, we postulate a linear structural causal model where confounding highlights the shortcomings of vanilla risk minimization and we study the necessity and benefit of leveraging a lower-dimensional, invariant subspace to assist with prediction. The invariant subspace in this model will unify concept stability and distributional stability, two distinct concepts in the literature a priori.

Guided by the insights about the invariance subspace, we proceed to solve the domain adaptation task. Given distributional access to the unlabeled target, $\mathcal{P}_{\mathcal{T}}(X)$, and labeled data from the source, $\mathcal{P}_{\mathcal{S}}(X,Y)$, we seek a methodology gauging towards the inaccessible target risk by learning a subspace that balances predictability and stability/invariance. Concretely, we seek an estimator decoupled as a composition of a linear subspace representation, $V \in \mathbb{R}^{d \times \ell}$, and a low-dimensional ridge regressor, $\alpha \in \mathbb{R}^{\ell}$, jointly optimized according to the objective:

$$\beta^{\eta,\upsilon} = V\alpha_{\mathcal{S}}^{V}, \ \{V, \ \alpha_{\mathcal{S}}^{V}\} := \arg\min \tfrac{1}{2} \left\{ \mathbb{E}_{\mathcal{S}}[(Y - \langle X, V\alpha \rangle)^2] + \upsilon\|\alpha\|^2 + \tfrac{\eta}{2}\|V^{\top}(\mathbb{E}_{\mathcal{T}}[XX^{\top}] - \mathbb{E}_{\mathcal{S}}[XX^{\top}])V\|_{\mathrm{F}}^2 \right\} . \tag{1.3}$$

We derive objective (1.3) as an upper bound of the inaccessible target risk $\mathbb{E}_{\mathcal{T}}[(Y - \langle X, \beta \rangle)^2]$. The term $\|V^{\top}(\mathbb{E}_{\mathcal{T}}[XX^{\top}] - \mathbb{E}_{\mathcal{S}}[XX^{\top}])V\|_{\mathrm{F}}^2$ is a natural notion of invariance for domain adaptation under the linear SCM, jointly determined by loss function and estimator class. Shown to be an upper bound on the target risk, (1.3) acts as an actionable proxy for optimization based on the information available. A practical first-order manifold optimization method is devised to navigate the non-convex landscape. By optimizing over the choice of regularization parameters η, υ in (1.3), the target risk can effectively be optimized, an empirical fact demonstrated using real-world data examples.

We then provide a theoretical characterization of the landscape of the non-convex manifold optimization. When the regularization parameter η is sufficiently large, we show that almost all local optima align well with an exogenous, invariant linear subspace and are resilient to distribution shifts driven by endogenous confounding factors. Indeed, denoting $\mathrm{Endo} \subset \mathbb{R}^d$ as the endogenous subspace corresponding to Δ in (1.2), we find the first-order stationary point $V \in \mathrm{St}(d, \ell)$ of (1.3) satisfies:

$$\max_{v \in V, w \in \mathrm{Endo}} \big(\cos \angle(v, w) \big)^6 \leq O\left(\frac{1}{\upsilon\eta^2}\right) .$$

5

This alignment result enables deriving a stability bound between the target and source risks that directly addresses the domain adaptation problem. For any first-order stationary point of (1.3),

$$R_{\mathcal{T}}(\beta^{\eta,\upsilon}) - R_{\mathcal{S}}(\beta^{\eta,\upsilon}) \leq \underbrace{\inf_{\beta \perp \text{Endo}} \{R_{\mathcal{T}}(\beta) - R_{\mathcal{S}}(\beta)\}}_{\text{oracle term}} + O\left(\frac{1}{\upsilon^{4/3}\eta^{2/3}}\right).$$

Namely, a predictive model using the learned lower-dimensional subspace could incur a nearly ideal gap between target and source risk, quantified by the oracle term. The theoretical results on invariance alignment and risk stability corroborate empirical observations in real-world data sets. Despite the difficult nature of non-convex manifold optimization, the practical first-order method often identifies good local optima for domain adaptation.

2 Undesired Properties of Source Risk Minimization

2.1 Concept Shift, Confounding, and Subspace Invariance

We first state the assumptions used in this section.

Assumption 1. In Definition 1, the exogeneous variables Z, U, W are mean-zero, namely, $\mathbb{E}[Z] = 0, \mathbb{E}[U] = 0, \mathbb{E}[W] = 0$, and satisfy $\mathbb{E}[WW^\top] = \tau^2 I_d$ with $\tau > 0$.

Assumption 2 (Orthogonality). In Definition 1, $k + r \leq d$, and the matrix satisfies $[\Theta, \Delta] \in \text{St}(d, k+r)$, the Stiefel manifold.

For each environment $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$, classical risk minimization takes on the following form:

$$\arg\min_{\beta \in \mathbb{R}^d} R_{\mathcal{E}}(\beta) = \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}_{\mathcal{E}}[(Y - \langle X, \beta \rangle)^2]. \tag{2.1}$$

Recall the SCM in Definition 1, when the endogeneity parameter $\gamma \neq 0$, the unobserved variable E lurking in the environment plays a confounding role in the relationship between X and Y. A natural question is: will there be a linear model simultaneously optimal for source and target risk minimization? The answer is no, thus implying that the best model will be a moving target. We shall see that a shift in the distribution of the confounding variable $E \sim \rho_{\mathcal{S}}$ to $E \sim \rho_{\mathcal{T}}$ can lead to a stark difference in the best linear model across distributions. The following proposition formally defines the above phenomenon as a *concept shift*.

Proposition 1 (Concept Shift). *Under Assumptions 1 and 2, the risk minimization in (2.1) admits a unique minimizer, which we denote as $\beta_{\mathcal{E}}$ for $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. If $\mathbb{E}_{\mathcal{S}}[EE^\top] \neq \mathbb{E}_{\mathcal{T}}[EE^\top]$, then*

$$\exists \gamma \in \mathbb{R}^r, \text{ such that } \beta_{\mathcal{S}} \neq \beta_{\mathcal{T}},$$

namely, the best linear predictors (concept) shift across the two environments for some endogeneity parameter γ.

Remark. It is immediate to show that under Assumption 1, the best linear model

$$\beta_{\mathcal{E}} = \beta^\star + (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma$$

is well-defined. The above result shows that the best linear model shifts as long as the environment changes. The curious reader may wonder whether the Bayes optimal model $\mathbb{E}_{\mathcal{E}}[Y|X]$ also changes. If we assume in addition that (E, Z, W) is drawn from a multivariate Gaussian and E is mean-zero, we have

$$\mathbb{E}_{\mathcal{E}}[Y|X] = \langle \beta^\star + (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma, X \rangle,$$

which matches the best linear model. Therefore, the Bayes optimal concept is also environment dependent. The above simple derivation also shows that the re-weighting method [33] according to the likelihood ratio of covariate distributions will not fix the concept shift issue. Due to endogeneity, a new methodology is needed.

In plain language, source risk minimization fails to recover a model that stays invariant across environments. The concept shift induced by the movement of the second moments of E reflects a bias that cannot be reconciled through traditional least squares. The root cause of the above concept shift is endogeneity. Inspired by the instrumental variable literature, we instead resort to exogenous linear subspaces of covariates invariant to environment shifts. The following proposition shows two desiderata—dimension reduction and invariance—can be achieved simultaneously.

Proposition 2 (Subspace Invariance). *Under Assumptions 1 and 2, let $\beta_{\mathcal{E}}^{\Theta}$ be the best linear predictor restricted to the linear subspace $\Theta \in \mathrm{St}(d, k)$ for each environment $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$,*

$$\beta_{\mathcal{E}}^{\Theta} := \Theta \alpha_{\mathcal{E}}^{\Theta}, \ \alpha_{\mathcal{E}}^{\Theta} = \underset{\alpha \in \mathbb{R}^k}{\arg \min} \, \mathbb{E}_{\mathcal{E}}[(Y - \langle X, \Theta \alpha \rangle)^2] \, .$$

Then,

$$\forall \gamma \in \mathbb{R}^r, \ we \ have \ \beta_{\mathcal{S}}^{\Theta} = \beta_{\mathcal{T}}^{\Theta} \, ,$$

namely, $\beta_{\mathcal{S}}^{\Theta} = \beta_{\mathcal{T}}^{\Theta}$ regardless of the endogenity parameter γ.

Remark. The above result shows that deconfounding is possible with an invariant subspace. In our model, the effect of the confounding variable, E, is restricted to the linear subspace Δ. Therefore, the covariate distribution projected to a lower-dimensional linear subspace, Θ, remains stable despite the environment shifts $\mathcal{P}_{\mathcal{S}}(X, Y) \to \mathcal{P}_{\mathcal{T}}(X, Y)$. Proposition 2 should be read in contrast to Proposition 1; risk minimization constrained to this "invariant" linear subspace resists arbitrary confounding effects parametrized by γ. And so, restricting to the invariant, exogenous space Θ yields stability in learned concepts, resilient to environment shifts.

In view of the decomposition,

$$Y = \langle \Theta^{\top} \beta^{\star}, Z \rangle + \langle \Delta^{\top} \beta^{\star} + \gamma, E \rangle + \text{noise} \, ,$$

the subspace model can be shown as $\beta_{\mathcal{E}}^{\Theta} \equiv \Theta \Theta^{\top} \beta^{\star}, \forall \mathcal{E}$. The above equation separates variation in Y into invariant, exogenous component $\langle \Theta^{\top} \beta^{\star}, Z \rangle$ and environment dependent, endogeneous component $\langle \Delta^{\top} \beta^{\star} + \gamma, E \rangle$. In particular, when $(I - \Theta \Theta^{\top}) \beta^{\star} = 0$, learning in the exogenous space consistently recovers the true signal; $\beta_{\mathcal{E}}^{\Theta} = \beta^{\star}$.

Proposition 2 only demonstrates the existence of an invariant linear subspace Θ. However, since the exogenous variable Z is unobserved, a data-driven approach to learning such a lower-dimensional linear subspace is still largely unclear. Motivated by invariance and stability, in Section 3, we will develop a methodology that inherits a manifold optimization problem to learn a lower-dimensional linear subspace. Later in Section 4, we will derive a theoretical characterization of the subspace alignment between the invariant Θ and a first-order stationary point of the manifold optimization. A target risk bound exploiting this alignment will also be established therein.

2.2 Target Risk Improvement via Invariance

For domain adaptation, the lingering question is whether leveraging the invariant subspace Θ can improve the target risk compared to $\beta_{\mathcal{S}}$, the vanilla source risk minimizer. This section provides a sufficient and necessary condition for the above question. The following proposition delineates when one can simultaneously achieve three desiderata: target risk improvement, invariance, and dimension reduction.

Proposition 3 (Target Risk Improvement). *Let Assumptions 1 and 2 hold with $k + r = d$. Denote $\Lambda_{\mathcal{E}} := \mathbb{E}_{\mathcal{E}}[EE^{\top}]$ for $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. The linear subspace predictor indexed by Θ based on source risk, that is, $\beta_{\mathcal{S}}^{\Theta}$, has a smaller target risk than the usual source risk minimizer $\beta_{\mathcal{S}}$*

$$R_{\mathcal{T}}(\beta_{\mathcal{S}}^{\Theta}) < R_{\mathcal{T}}(\beta_{\mathcal{S}}) \, ,$$

if and only if

$$\|\Delta^{\top} \beta^{\star} + (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1} \Lambda_{\mathcal{T}} \gamma\|_{\tau^2 I_r + \Lambda_{\mathcal{T}}}^2 < \|(\tau^2 I_r + \Lambda_{\mathcal{S}})^{-1} \Lambda_{\mathcal{S}} \gamma - (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1} \Lambda_{\mathcal{T}} \gamma\|_{\tau^2 I_r + \Lambda_{\mathcal{T}}}^2 \, .$$
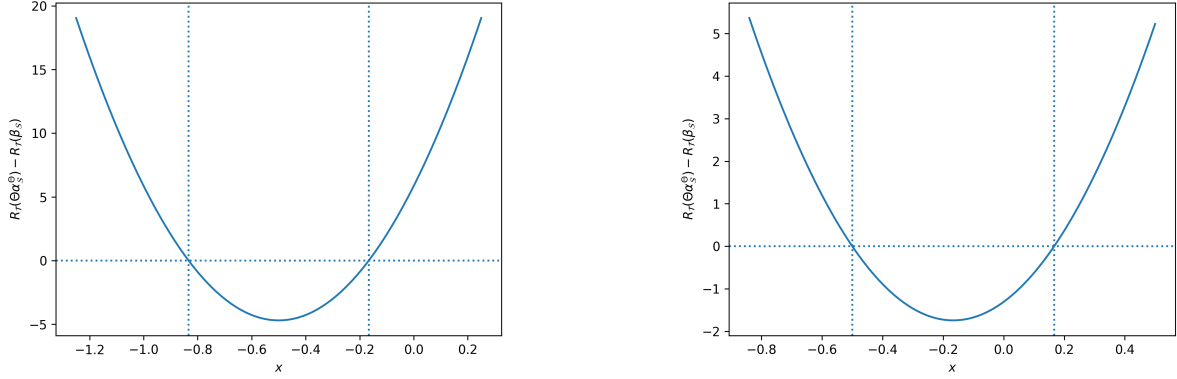
Figure 2: Left: Target Rich Regime with $\sigma_t^2 = 10$, $\sigma_s^2 = 2$, $\tau^2 = 10$; Right: Source Rich Regime with $\sigma_t^2 = 2$, $\sigma_s^2 = 10$, $\tau^2 = 10$. Here x-axis illustrates the scalar parameter $x \in \mathbb{R}$, and y-axis shows the risk improvement $R_\mathcal{T}(\beta_\mathcal{S}^\Theta) - R_\mathcal{T}(\beta_\mathcal{S})$.

The right-hand side can be interpreted as the amount of concept shifts, and the left-hand side can be interpreted as the sub-optimality of target risk due to the (invariant) subspace model. A special case helps to unpack the result. Consider $(I - \Theta\Theta^\top)\beta^\star = 0$, then the above expression reads

$$\|\Lambda_\mathcal{T}\gamma\|_{(\tau^2 I_r + \Lambda_\mathcal{T})^{-1}} < \|\tau^2 \underbrace{(\Lambda_\mathcal{T} - \Lambda_\mathcal{S})}_{\text{environment shift}} (\tau^2 I_r + \Lambda_\mathcal{S})^{-1}\gamma\|_{(\tau^2 I_r + \Lambda_\mathcal{T})^{-1}} \, .$$

Conceptually, when the environment shift—parametrized by the second moment of the confounding variable E—is large enough, the subspace model strictly improves upon the vanilla source risk minimizer. When such a condition holds, target risk improvement, invariance, and dimension reduction can be achieved simultaneously.

Before concluding this section, we use the following simple example to elucidate the condition in Proposition 3.

Example. Consider the simple setting where $\Lambda_\mathcal{S} = \sigma_s^2 \cdot I_r$, $\Lambda_\mathcal{T} = \sigma_t^2 \cdot I_r$, and $\Delta^\top \beta^\star = x \cdot \gamma$ with a scalar $x \in \mathbb{R}$. The equivalent condition to $R_\mathcal{T}(\beta_\mathcal{S}^\Theta) < R_\mathcal{T}(\beta_\mathcal{S})$ in Proposition 3 boils down to

$$\left(x + \frac{\sigma_t^2}{\sigma_t^2 + \tau^2}\right)^2 < \left(\frac{\sigma_s^2}{\sigma_s^2 + \tau^2} - \frac{\sigma_t^2}{\sigma_t^2 + \tau^2}\right)^2 \, .$$

Fixing other parameters σ_t, σ_s, τ, the condition can be interpreted as inequalities for the one-dimensional parameter x:

- *Target Rich Regime* with $\sigma_t > \sigma_s$, $\frac{-\sigma_s^2}{\sigma_s^2 + \tau^2} + \frac{-2\tau^2(\sigma_t^2 - \sigma_s^2)}{(\sigma_t^2 + \tau^2)(\sigma_s^2 + \tau^2)} < x < \frac{-\sigma_s^2}{\sigma_s^2 + \tau^2}$;

- *Source Rich Regime* with $\sigma_s > \sigma_t$, $\frac{-\sigma_s^2}{\sigma_s^2 + \tau^2} < x < \frac{-\sigma_s^2}{\sigma_s^2 + \tau^2} + \frac{-2\tau^2(\sigma_t^2 - \sigma_s^2)}{(\sigma_t^2 + \tau^2)(\sigma_s^2 + \tau^2)}$.

Figure 2 visualizes the quantitative condition: the magnitude of confounding and the amount of environment shift altogether determine when invariance renders risk improvement.

Leveraging subspace invariance allows one to surpass the limitations of source risk minimization in situations where the concept and the covariate shift. In the next section, we will propose a new domain adaptation method that learns a linear subspace resilient to environment shifts, making the insights obtained in this section actionable. The proposed methodology—an instantiation of stability/invariance regularized risk minimization—optimizes over a representation parametrized by a lower-dimensional subspace projection. The method trades off predictability and stability, using lower-dimensional representation as a vehicle.

3 Methodology: Domain Adaptation via Manifold Optimization

The previous section briefly touches upon the fundamental tradeoff between robustness/invariance and predictive performance. In general, an invariant subspace can improve upon the source risk minimizer but may not be optimal for the target risk in domain adaptation. This section contributes to delineating the abovementioned tradeoff in the SCM defined in Definition 1. Concretely, we design a representation learning method for domain adaptation, parametrized by subspace projections, and navigate the tradeoff between representation invariance and risk minimization. The proposed manifold optimization is non-convex. However, we demonstrate that a standard first-order method works both empirically in Section 3.3 with real data sets, and later provably in Section 4, where we develop theory matching the empirics.

We motivate our main methodology in two ways: (i) as a surrogate for an upper bound on the target risk directly, and (ii) as a tradeoff to balance robustness/invariance and predictive performance.

3.1 A Surrogate for the Target Risk

In typical domain adaptation, labels from the target environment $Y \sim \mathcal{P}_{\mathcal{T}}(Y)$ are unavailable, and thus $R_{\mathcal{T}}(\beta)$ cannot be directly minimized. To circumvent this issue, we design a simple surrogate objective that does not require labels from the target environment, requiring only information about $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$. In the sequel we will denote $\Sigma_{\mathcal{E}} := \mathbb{E}_{\mathcal{E}}[XX^{\top}]$ for $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. We start with a simple algebraic relation between the source and the target.

Proposition 4. *Under Assumption 1 and $\Delta^{\top}\Delta = I_r$, we have*

$$R_{\mathcal{T}}(\beta) = R_{\mathcal{S}}(\beta) + \langle \beta - (\beta^{\star} + \Delta\gamma), (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})(\beta - (\beta^{\star} + \Delta\gamma)) \rangle, \quad \forall \beta \in \mathbb{R}^d . \tag{3.1}$$

The geometry of the target risk landscape involves a balance of two terms with clear distance interpretations. The first term, $R_{\mathcal{S}}(\beta)$, is an actionable objective in domain adaptation based on source risk. The second term encodes a distance between β and the endogenous component, $\beta^{\star} + \Delta\gamma$, under a geometry determined by the covariance shift matrix $\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}$. The second term is not actionable based on data because $\beta^{\star} + \Delta\gamma$ is unknown. In the following proposition, we define an actionable objective, which serves as a surrogate upper bound for the target risk, for all β and $\beta^{\star} + \Delta\gamma$.

To make the presentation simple, we impose the following additional assumption.

Assumption 3 (Target Richer than Source). Let $\Lambda_{\mathcal{E}} := \mathbb{E}_{\mathcal{E}}[EE^{\top}]$ for $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. Assume that $\Lambda_{\mathcal{T}} \succ \Lambda_{\mathcal{S}}$.

Note that the assumption is for simplicity of presentation: we can state more general results by separating the positive and negative parts of $\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}}$. More importantly, we use Assumption 3 to focus on the interesting regime of distribution shift where the target distribution strictly dominates the source. This is when out-of-domain extrapolation can happen. Informally speaking, the Loewner order, $\Lambda_{\mathcal{T}} \succ \Lambda_{\mathcal{S}}$, indicates more variability present in the target domain.

With Assumption 3 in hand, an upper bound on target risk can be derived that suggests a practical proxy for $R_{\mathcal{T}}(\beta)$. To elucidate the low-dimensional subspace dependence, we explicitly construct the estimator β to be a composite map of a lower-dimensional representation, parametrized by $V \in \mathbb{R}^{d \times \ell}$, and a linear model restricted to the representation, parametrized by $\alpha \in \mathbb{R}^{\ell}$.

Proposition 5 (Surrogate for Target). *Assume Assumption 1 with $\Delta^{\top}\Delta = I_r$, and Assumption 3 with $D := \Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}} \succ 0$. Then $\forall \xi, \zeta > 0$*

$$R_{\mathcal{S}}(V\alpha) \leq R_{\mathcal{T}}(V\alpha) \leq R_{\mathcal{S}}(V\alpha) + \tfrac{(1+\xi)\zeta}{2}\|\alpha\|^4 + \tfrac{(1+\xi)}{2\zeta}\|V^{\top}DV\|_{\mathrm{F}}^2 + (1 + \tfrac{1}{\xi})\langle \beta^{\star} + \Delta\gamma, D(\beta^{\star} + \Delta\gamma) \rangle .$$

Proposition 5 highlights that the gap between target and source risks of a composite estimator $\beta = V\alpha$ can be controlled by two complexities: $\|\alpha\|$ and $\|V^{\top}(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$. The term $\|V^{\top}(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$ quantifies the alignment of the subspace V with the directions of covariance shift; in essence, a term describing stability

for extrapolation. The term may also be interpreted as a quadratic maximum mean discrepancy (MMD) between covariate distributions in the linear subspace V. Indeed, under the map $\psi_V(X) = V^\top X X^\top V$,

$$\left\| V^\top (\mathbb{E}_\mathcal{T}[XX^\top] - \mathbb{E}_\mathcal{S}[XX^\top])V \right\|_\mathrm{F} = \sup_{M : \|M\|_\mathrm{F} \leq 1} \mathbb{E}_\mathcal{T}[\langle M, \psi_V(X)\rangle] - \mathbb{E}_\mathcal{S}[\langle M, \psi_V(X)\rangle] . \tag{3.2}$$

Finally, note that the right-hand side in Proposition 5 is a regularized source risk minimization, regularized by ridge and stability penalties. More importantly, it is an actionable surrogate objective for domain adaptation, solely depending on $\mathcal{P}_\mathcal{S}(X, Y)$ and $\mathcal{P}_\mathcal{T}(X)$.

3.2 An Optimization Procedure on the Stiefel Manifold

Motivated by the upper bound on the target risk shown in Proposition 5, we define the following penalized objective to search for a linear subspace V, and a model restricted to that subspace parametrized by α,

$$F_{v,\eta}(V, \alpha) := \tfrac{1}{2} \left\{ R_\mathcal{S}(V\alpha) + v\|\alpha\|^2 + \tfrac{\eta}{2} \|V^\top (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})V\|_\mathrm{F}^2 \right\} . \tag{3.3}$$

By optimizing over a linear subspace $V \in \mathbb{R}^{d \times \ell}$ that balances predictive power and stability, the above objective provides an explicit tradeoff. Notice that the two-level optimization problem is convex in α for fixed V, where the inner optimum $\min_{\alpha \in \mathbb{R}^\ell} F_{v,\eta}(V, \alpha)$ is attained at

$$\alpha_V := (V^\top \mathbb{E}_\mathcal{S}[XX^\top]V + vI_\ell)^{-1} V^\top \mathbb{E}_\mathcal{S}[XY] . \tag{3.4}$$

We search for a representation confined to the Stiefel manifold $V \in \mathrm{St}(d, \ell)$

$$\mathrm{St}(d, \ell) := \{V \in \mathbb{R}^{d \times \ell} \ : \ V^\top V = I_\ell\} . \tag{3.5}$$

Putting things together, we arrive at the following optimization on the Stiefel manifold,

$$\min_{V \in \mathrm{St}(d,\ell)} \Phi_{v,\eta}(V) := \min_{V \in \mathrm{St}(d,\ell)} F_{v,\eta}(V, \alpha_V) \tag{3.6}$$

where α_V is defined in (3.4). The problem $\Phi_{v,\eta}(V)$ is non-convex in V; however, we will derive provable results for any first-order stationarity of the landscape. Informally speaking, as we shall see in Section 4, under certain conditions, almost all local optima demonstrate good alignment with the invariant subspace.

Let us first introduce a skeletal implementation of a first-order manifold optimization method to find local optima of $\Phi_{v,\eta}(V)$. Then, we move to fill in the essential background on the geometry of the Stiefel manifold to rationalize the algorithm.

Algorithm 1 Optimization on Stiefel Manifold

Require: Initial point: $V_0 \in \mathrm{St}(d, \ell)$.
    for $k = 0, 1, 2, \dots$ **do**
        Gradient: $G_k \leftarrow (I - V_k V_k^\top)((\Sigma_\mathcal{S} V_k \alpha_{V_k} - \mathbb{E}_S[XY])\alpha_{V_k}^\top + \eta D V_k V_k^\top D V_k)$ as per (3.10).
        Retraction: $V_{k+1}(t) \leftarrow (V_k + t \cdot G_k)(I_\ell + t^2 \cdot G_k^\top G_k)^{-1/2}$ as per (3.11), $t \in \mathbb{R}_+$.
        Line-Search: choose step-size $t = t_k$ via line-search, and set $V_{k+1} \leftarrow V_{k+1}(t_k)$.
    end for

Geometry on the Stiefel Manifold   A few essential geometric items must be defined to understand the first-order method optimizing (3.6) detailed in Algorithm 1. We treat $\mathrm{St}(d, \ell)$ as an embedded submanifold in the vector space $\mathbb{R}^{d \times \ell}$ and introduce appropriate notions of tangent space, projection, and gradient. We provide a cursory overview of the Stiefel geometry for a self-contained treatment. A familiar reader may skip this part.

First, the tangent space at a point on a sub-manifold is intuitively the plane tangent to the sub-manifold at that point. The normal space is the corresponding orthogonal complement. The tangent space and normal space to $\mathrm{St}(d,\ell)$ at V are respectively:

$$T_V\mathrm{St}(d,\ell) = \left\{ Z \in \mathbb{R}^{d\times\ell} : V^\top Z + Z^\top V = 0 \right\}, \quad (T_V\mathrm{St}(d,\ell))^\perp = \left\{ VS : S = S^\top \in \mathbb{R}^{\ell\times\ell} \right\} . \tag{3.7}$$

We view the Stiefel manifold as a sub-manifold of $\mathbb{R}^{d\times\ell}$, with its tangent spaces inheriting the Euclidean metric (Frobenius norm), $\|\cdot\|_\mathrm{F}$, and inner product, $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$. With the metric defined, projections of $\xi \in \mathbb{R}^{d\times\ell}$ on to the tangent and the normal spaces at V are given by:

$$P_V(\xi) = \left(I_d - VV^\top\right)\xi + V\,\mathrm{skew}\left(V^\top\xi\right), \quad P_V^\perp(\xi) = V\,\mathrm{sym}\left(V^\top\xi\right) , \tag{3.8}$$

where $\mathrm{skew}(A) = (A - A^\top)/2$ and $\mathrm{sym}(A) = (A + A^\top)/2$.

Finally, if F is a smooth function defined on $\mathbb{R}^{d\times\ell}$, and $\bar{F}$ is its restriction to the Riemannian sub-manifold $\mathrm{St}(d,\ell)$, the gradient of $\bar{F}$ is equal to the projection of the gradient of F onto $T_V\mathrm{St}(d,\ell)$:

$$\mathrm{grad}\ \bar{F}(V) = P_V(\mathrm{grad}\ F(V)). \tag{3.9}$$

A rigorous treatment of the above can be found in [1, Chapters 3 and 4].

Gradient Formula and Retraction With the geometry understood, the Riemannian gradient of the objective (3.3) can be readily calculated.

Proposition 6 (Riemannian Gradient and Retraction). *Consider* $\bar{\Phi}_{v,\eta} : \mathrm{St}(d,\ell) \to \mathbb{R}$, *the restriction of* $\Phi_{v,\eta}$ *to the Riemannian sub-manifold* $\mathrm{St}(d,\ell)$. *The Riemannian gradient of the objective* (3.6) *is*

$$\mathrm{grad}\ \bar{\Phi}_{v,\eta}(V) = (I_d - VV^\top)\left((\Sigma_\mathcal{S}V\alpha_V - \mathbb{E}_S[XY])\alpha_V^\top + \eta DVV^\top DV\right) \in T_V\mathrm{St}(d,\ell) . \tag{3.10}$$

Moreover, the gradient formula for the objective (3.3) *is the same for either Stiefel or Grassmann manifold.*

For any $\xi \in T_V\mathrm{St}(d,\ell)$, the retraction map based on matrix polar factorization is defined as

$$\mathrm{R}_V(\xi) = (V + \xi)(I_\ell + \xi^\top\xi)^{-1/2} . \tag{3.11}$$

Proposition 6 allows us to define first-order descent methods to optimize (3.6). For example, a simple line-search method on the Stiefel manifold follows the update formula:

$$V_{k+1} = \mathrm{R}_{V_k}(t_k \cdot G_k), \quad G_k = \mathrm{grad}\ \bar{\Phi}_{v,\eta}(V_k)$$

where R_{V_k} is the retraction map from (3.11) and t_k is a step-size. Now, all ingredients of Algorithm 1 readily unfold.

3.3 Real-World Data Examples

Having established a framework for first-order optimization of (3.6), we apply the method to several real world data sets exhibiting distribution shifts. The following examples naturally admit distinct environments with markedly different covariate distributions. In addition, the optimal least-squares regressor in a source domain is notably different from that of the target domain, firmly landing us in the concept shift setting of Proposition 1.

- **Forest Fires Data** [10]: The goal is to predict the burned area of forest fires in the northeast region of Portugal using various meteorological features. We choose seven covariates ($d = 7$): temperature, relative humidity, wind speed, precipitation, and three fire weather indices. We emulate seasonal shifts by considering two environments separated in time; the data corresponding to June, July, and August constitute the target, while we take the remaining data as the source.

- **Bike Sharing Data** [12]: Obtained from the bikeshare system called Capital Bikeshare serving Washington, D.C., USA., the goal of this data set is to predict hourly counts of bike rentals using features representing weather information. We postulate two different environments depending on the seasons, spring and fall for source and target, respectively. Here, we take a subset of this data by focusing on workdays and take five features ($d = 5$): hour (0 to 23), temperature, feeling temperature, humidity, and wind speed.

- **Wine Quality Data** [11]: The goal in this exercise is to predict the quality of wine using eleven physicochemical features ($d = 11$): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. We take the data corresponding to white and red wine as the source and target, respectively.

For each dataset, we obtain the linear subspace predictor $V\alpha$ by solving (3.6) for different values of regularization parameters υ and η. We implement the optimization using the PYMANOPT package [35]. Figures 3, 4, and 5 illustrate the performance of the obtained predictor for the forest fires, bike sharing, and wine quality data, respectively. Each figure consists of three subfigures: (a) the target risk $R_{\mathcal{T}}(V\alpha)$, the primary objective that is inaccessible, (b) the source risk $R_{\mathcal{S}}(V\alpha)$, and (c) the difference $R_{\mathcal{T}}(V\alpha) - R_{\mathcal{S}}(V\alpha)$, a quantifier for the stability/invariance. For (a), the solid red horizontal line shows $R_{\mathcal{T}}(\beta_{\mathcal{S}})$, the target risk of the vanilla source risk minimizer $\beta_{\mathcal{S}}$, and the solid black line shows $R_{\mathcal{T}}(\beta_{\mathcal{T}})$, the target risk of the target risk minimizer $\beta_{\mathcal{T}}$ if we were given the labeled access to the target distribution $\mathcal{P}_{\mathcal{T}}(X, Y)$—a best possible oracle benchmark infeasible in practice.
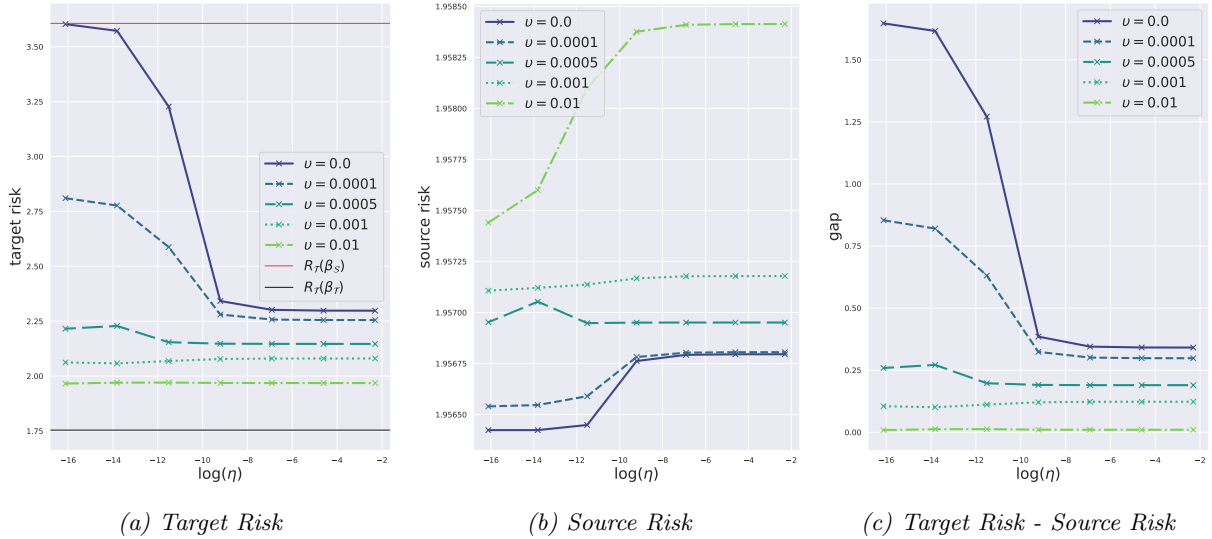

(a) Target Risk   (b) Source Risk   (c) Target Risk - Source Risk

Figure 3: Forest Fires Data: Performance of the linear subspace predictor $V\alpha$ obtained by solving (3.6) for different values of υ and η, where $d = 7$ and $\ell = 6$. (a) plots the risk on target dataset $R_{\mathcal{T}}(V\alpha)$, where the solid red horizontal line shows $R_{\mathcal{T}}(\beta_{\mathcal{S}})$ and the solid black line shows $R_{\mathcal{T}}(\beta_{\mathcal{T}})$. (b) shows the risk of on source dataset $R_{\mathcal{S}}(V\alpha)$. (c) plots the difference $R_{\mathcal{T}}(V\alpha) - R_{\mathcal{S}}(V\alpha)$.

From Figures 3(a), 4(a), and 5(a), we can see that for each υ, there is a range of η such that the resulting target risk, $R_{\mathcal{T}}(V\alpha)$, is smaller than the target risk of the source risk minimizer $R_{\mathcal{T}}(\beta_{\mathcal{S}})$ (visually below the red solid line). This indicates that the proposed procedure can improve the target risk by balancing the stability and the predictive power of the estimator for certain combinations of the parameters υ and η. Notably, for the forest fires data, we can see improvement over the source risk minimizer $\beta_{\mathcal{S}}$ for any pair (υ, η). The PYMANOPT implementation converged to the desired first-order local optima for the forest fires and bike sharing data sets, as we numerically checked the first-order conditions for the solution. The optimization for the wine quality data did not seem to converge to stationarity before reaching the maximum number of
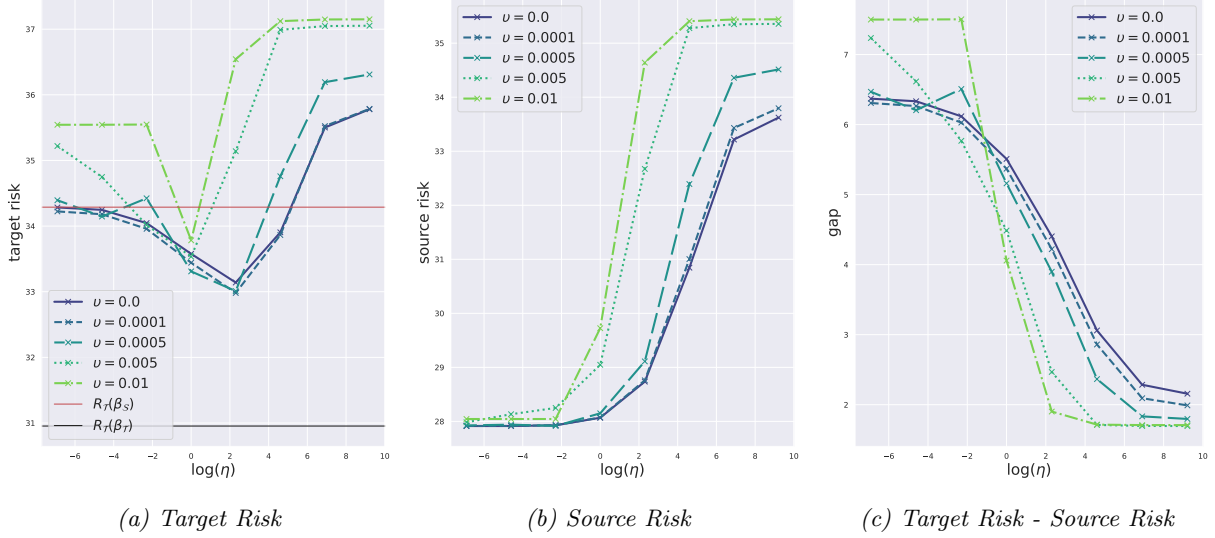
(a) Target Risk *(b) Source Risk* *(c) Target Risk - Source Risk*

Figure 4: *Bike Sharing Data: Performance of the linear subspace predictor $V\alpha$ obtained by solving* (3.6) *for different values of v and η, where $d = 5$ and $\ell = 4$. The subplots (a)-(c) plot the same quantities as in Figure 3.*

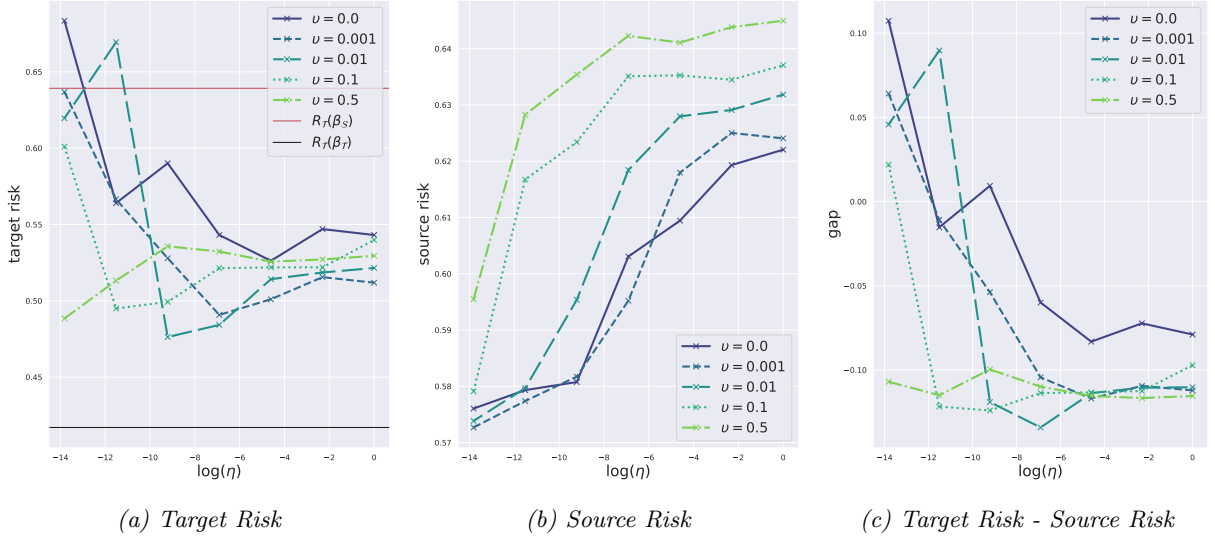*(a) Target Risk* *(b) Source Risk* *(c) Target Risk - Source Risk*

Figure 5: *Wine Quality Data: Performance of the linear subspace predictor $V\alpha$ obtained by solving* (3.6) *for different values of v and η, where $d = 11$ and $\ell = 7$. The subplots (a)-(c) plot the same quantities as in Figure 3.*

iterations. That said, Figure 5(a) still shows that the target risk of the obtained predictor is smaller than that of the vanilla source risk minimization for certain values of v and η. This indicates that taking gradient steps can be beneficial despite an absence of convergence. Meanwhile, the non-monotonic relationship between target risk and η in Figure 4(a) is due to the conflicting monotone behaviors of source risk shown in (b) and risk gap shown in (c), which is explored in Theorem 2. Figures 3(c), 4(c), and 5(c) numerically validate the stability bound in Theorem 2: as η increases, the risk gap, $R_{\mathcal{T}}(\beta^{\eta,v}) - R_{\mathcal{S}}(\beta^{\eta,v})$, improves; while for small v, the risk gap typically deteriorates.

13

3.4 Interpretation: Stability vs. Predictability

It was suggested in Section 3.2 that the penalized objective (3.6) is an instantiation of a robustness and accuracy tradeoff. Observe that under the linear SCM in Definition 1, two notions of stability/invariance coincide: optimizing over a *stable representation* amidst covariate shifts will align with searching for the *causal, invariant relationship*, stable across environments. We make the stability vs. predictability tradeoff explicit now and, by doing so, relate the method to a class of procedures recently popularized in machine learning. Our procedure seeks a two-part solution: a linear projection V, enforcing a similarity between $\mathcal{P}_\mathcal{S}$ and $\mathcal{P}_\mathcal{T}$, and a linear estimator, α, built atop. Selecting penalization then translates to a balancing of stability and predictability. This is transparent once we decouple $F_{v,\eta}$ into loss and regularization,

$$F_{v,\eta}(V,\alpha) := \overbrace{\tfrac{1}{2}\mathbb{E}_\mathcal{S}[(Y - \langle X, V\alpha \rangle)^2]}^{Predictivity} \quad + \quad \overbrace{\tfrac{v}{2}\|\alpha\|^2 + \tfrac{\eta}{4}\left\|V^\top \left(\Sigma_\mathcal{T} - \Sigma_\mathcal{S}\right) V\right\|_{\mathrm{F}}^2}^{Stability} .$$

The source *Predictivity* term corresponds to prediction under squared loss $Y \sim V^\top X$, confined to a subspace V; it extracts a linear relationship α between the labels and a predictive subspace of the data.

The *Stability* term quantifies two things. First, a notion of distributional stability determined by the penalty $\eta \left\|V^\top \left(\Sigma_\mathcal{T} - \Sigma_\mathcal{S}\right) V\right\|_{\mathrm{F}}^2$ for the representation subspace V. The penalty can be interpreted as a distribution distance; see (3.2). Second, a standard ridge penalization quantifies the stability of the linear relationship α given the subspace V.

Conceptually, as $\eta \to \infty$, V captures the linear subspace over which the covariate second moments stay invariant. Moreover, such an invariant linear subspace corresponds to the contribution from the exogenous Z that identifies the invariant causal relationship. Indeed, $\Theta^\top \left(\mathbb{E}_\mathcal{T}[XX^\top] - \mathbb{E}_\mathcal{S}[XX^\top]\right)\Theta = 0$, therefore provided $l > k$, $\Theta \in \mathbb{R}^{d\times k}$ is an element of the non-empty solution set. In this sense, the case when $\eta \to \infty$ coincides with the invariant estimator explored in Proposition 3. More generally, and as seen in Figure 4, a choice of η in an intermediary region may balance stability and source predictability to harness improvements in target risk, the primary object in domain adaptation.

It is easy to see how the framework can be generalized to deal with different notions of distributional discrepancy, loss function, and model class. Specifically, a family of methods is based on finding a stable and predictive representation $\phi \colon \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a suitable feature space. Let $\phi_\#\mu$ and $\phi_\#\nu$ denote the source and target covariate distributions pushed forward to the mapped space $\mathcal{Z}$. The goal is to find a representation $\phi \colon \mathcal{X} \to \mathcal{Z}$ and a composite predictor $g \colon \mathcal{Z} \to \mathcal{Y}$ simultaneously by solving the following:

$$\min_{g,\phi} R_\mathcal{S}(g \circ \phi) + \eta \cdot d(\phi_\#\mu, \phi_\#\nu) ,$$

where $d(\cdot,\cdot)$ is a suitable discrepancy between probability distributions on the mapped space $\mathcal{Z}$. Several choices for d have been proposed in the machine learning literature. For example, [36] and [19] use the kernel maximum mean discrepancy with a suitable kernel $k(\cdot,\cdot) \colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, [13] uses a suitable hypothesis-based discrepancy introduced in [4, 5], and [32] uses the Wasserstein-1 distance, to name a few.

4 Theory: Invariance Alignment and Risk Stability

We now study two theoretical properties of the methodology proposed in Section 3: alignment with the invariant subspace and risk across environments. The first result in Theorem 1 is a characterization of the landscape of the non-convex manifold optimization. We show almost all local optima exhibit favorable alignment with the invariant linear subspace as long as the regularization parameter is sufficiently large. The second result speaks to domain adaptation. Building on the landscape result, we derive a stability bound between the source and target risk in Theorem 2, which provides theoretical underpinnings for the phenomenon observed in the numerics section.

4.1 Alignment with the Invariant Subspace

In Propositions 2 and 3, the invariant subspace Θ is shown to be advantageous. The following theorem proves that the data-driven manifold optimization method given in Section 3 can identify linear subspaces that are approximately orthogonal to the contribution of the endogenous, confounding variable E. We employ the canonical correlation (or angle) between linear subspaces to quantify the notion of approximate orthogonality. By dodging the endogenous subspace where E influences X, the learned subspace V approximately aligns with the invariant subspace—crucial for concept invariance and distributional stability.

Theorem 1 (Alignment with the Invariant Subspace). *Consider the model in Definition 1 that satisfies Assumptions 1 and 3, and that $\Delta^\top \Delta = I_r$. Denote the matrix of covariance shift $D := \Sigma_\mathcal{T} - \Sigma_\mathcal{S}$.*

For any $V \in \mathrm{St}(d, \ell)$ that satisfies,

1. *V is a first-order stationary point of the Stiefel optimization procedure (3.6) with $\mathrm{grad}\, \Phi_{v,\eta}(V) = 0$, and*

2. *V is in the admissible set $\mathcal{S}_\Delta(\delta)$, where $\mathcal{S}_\Delta := \{V \in \mathrm{St}(d, \ell) \ : \ \|V^\top \Delta\|_{\mathrm{op}}^2 \leq 1 - \delta\}$,*

then

$$\|V^\top \Delta\|_{\mathrm{op}}^6 \leq \frac{\delta^{-1} \lambda_{\max}(\Sigma_\mathcal{S}) \|\Sigma_\mathcal{S}^{-1/2} \mathbb{E}_\mathcal{S}[XY]\|^4}{\lambda_{\min}(\Delta^\top D \Delta)^4} \frac{1}{4v\eta^2} \ .$$

Remark. The above result characterizes the optimization landscape on the Stiefel manifold and should be interpreted when the regularization parameter η is large.

The result has two noteworthy features. First, we establish a structural result for *all local optima* for the non-convex manifold optimization, which can be found efficiently in practice using Algorithm 1; we do not require the solution V to be the global optima.

Second, the result operates even when the invariant subspace overlaps with the endogenous space, namely $\Theta^\top \Delta \neq 0$. In such a case, the learned subspace V will still be approximately orthogonal to Δ, in the sense that for any $v \in V$ and $w \in \Delta$,

$$\max_{v \in V, w \in \Delta} \cos \angle(v, w) \leq \mathrm{const.} \times \frac{1}{(v\eta^2)^{1/6}} \ .$$

Conceptually, the optimization procedure will identify a strict subspace inside the invariant subspace to approximately dodge the endogenous subspace Δ, so long as the stability regularization parameter η is large and ridge regularization parameter ν is not too small.

4.2 Stability and Target Risk Bound

Given the alignment with the invariant subspace, what remains to be shown is its implications for the domain adaptation problem, a task for this section. In that regard, the following theorem derives a stability bound between the target and the source risks.

Recall covariance matrix shift $D = \mathbb{E}_\mathcal{T}[XX^\top] - \mathbb{E}_\mathcal{S}[XX^\top] = \Sigma_\mathcal{T} - \Sigma_\mathcal{S}$, and the estimator

$$\beta^{\eta, v} = V \alpha_\mathcal{S}^V \ .$$

Here V is any first-order stationarity of the $\Phi_{v,\eta}(V) := F_{v,\eta}(V, \alpha_\mathcal{S}^V)$, with

$$F_{v,\eta}(V, \alpha) = \tfrac{1}{2} \left\{ R_\mathcal{S}(V\alpha) + v\|\alpha\|^2 + \tfrac{\eta}{2} \|V^\top (\Sigma_\mathcal{T} - \Sigma_\mathcal{S}) V\|_F^2 \right\} \ .$$

The following theorem shows the stability of any stationary point of the above non-convex Stiefel manifold optimization $\Phi_{v,\eta}(V)$.

Theorem 2 (Stability between Source and Target). *Consider the setting as in Theorem 1. Let $V \in \mathrm{St}(d, \ell)$ be any first-order stationary point of the optimization procedure (3.6) with regularization parameters η, υ, and $\beta^{\eta,\upsilon}$ be the corresponding linear subspace estimator restricted to V. For any $\epsilon > 0$, define*

$$\mathsf{S}_{\epsilon,\delta} := (1 + \epsilon^{-1})\delta^{-1/3}\lambda_{\max}(\Sigma_{\mathcal{S}})^{1/3} \frac{\lambda_{\max}(\Delta^{\top} D\Delta)}{\lambda_{\min}(\Delta^{\top} D\Delta)^{4/3}} \|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^{10/3} \,,$$

the following generalization bound holds

$$R_{\mathcal{T}}(\beta^{\eta,\upsilon}) - R_{\mathcal{S}}(\beta^{\eta,\upsilon}) \leq (1 + \epsilon) \cdot \langle \beta^{\star} + \Delta\gamma, D(\beta^{\star} + \Delta\gamma) \rangle + \mathsf{S}_{\epsilon,\delta} \cdot \frac{1}{(4\upsilon)^{4/3}\eta^{2/3}} \,.$$

Remark. Note the first term is necessary. Even for the oracle V perfectly invariant to the environment shift, the oracle gap between the target and source risks equals $\langle \beta^{\star} + \Delta\gamma, D(\beta^{\star} + \Delta\gamma) \rangle$, in view of Proposition 4. The above result proves an approximate oracle to the gap between target and source risk, quantified by ϵ.

As $\eta \to \infty$ (holding all else fixed), the stability bound $S_{\epsilon,\delta} \cdot \frac{1}{(4\upsilon)^{4/3}\eta^{2/3}}$ decreases with η, which confirms the numerical findings in Section 3.3. Conceptually, as η increases, the source risk $R_{\mathcal{S}}(\cdot)$ will increase as one trades off prediction power for stability; on the plus side, instability $R_{\mathcal{T}}(\cdot) - R_{\mathcal{S}}(\cdot)$ is reduced due to invariance.

5 Conclusion and Future Work

This paper investigates domain adaptation for observational data, directly confronting the unobserved confounding issue. Unobserved confounding complicates domain adaptation in two ways: (i) by shifting the concept of the best statistical model across environments and (ii) by extrapolating the model to the unseen domain of covariates. Both pose robustness challenges to conventional statistical learning. To progress, we postulate a causal model for distribution shifts in the presence of confounding. We identify the undesired theoretical properties of vanilla source risk minimization and further characterize the benefits of leveraging an invariant subspace to enforce stable learning and improve the target risk. On the methodological front, we propose a new, concrete domain adaptation algorithm. The algorithm learns a representation parametrized by a lower-dimensional projection and solves a non-convex manifold optimization problem using now-standard first-order Riemannian optimization techniques. The new algorithm can also be derived—using first principles—as a distributional-stability/invariance-regularized source risk minimization that anchors the target risk, thereby conforming to the domain adaptation literature on the stability and predictability trade-off. On the empirical front, we demonstrate the practical use of the method on three real-world data sets. On the theoretical front, we study the landscape of the non-convex manifold optimization and further establish provable guarantees for nearly all local optima. With sufficient regularization, local optima identified by first-order Riemannian optimization will align to an exogenous, invariant linear subspace resilient to both the concept shift and the covariate shift. Regarding predictability in domain adaptation, we prove a target risk bound by showing a predictive model via the learned lower-dimensional subspace could incur a nearly ideal gap between target and source risk. There are several directions we left for future work.

Non-Linear Models/Representations This paper discusses a linear SCM and optimizes over linear representations parametrized by lower-dimensional projections. In the presence of confounding, extensions to non-linear models and to provable non-linear representation learning are left unexplored.

Notions of Distributional Stability and Invariance In our SCM setting, we derive from first principles that the notion of distributional stability should be based on the shift in the second-moment matrix from the source to the target. This notion of distance between probability distribution suffices in the context of a finite-dimensional linear model. What should be the correct notion of distributional stability for a generic domain adaptation problem? How to solve the corresponding distributional-stability-regularized source risk minimization?

Higher-Order Methods for Manifold Optimization We only study the first-order optimization method over the Stiefel manifold. How can accelerated or higher-order optimization methods be readily incorporated into representation learning in domain adaptation? By leveraging higher-order differential information, can we provide a stronger analysis of the optimization landscape?

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[2] Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430): 431–442, 1995.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

[5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[6] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128. Association for Computational Linguistics, 2006.

[7] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

[8] David E. Card and Alan B. Krueger. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4):772–793, 1994.

[9] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2022.

[10] Paulo Cortez and Aníbal Morais. A data mining approach to predict forest fires using meteorological data. In *Portuguese Conference on Artificial Intelligence*, 2007.

[11] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

[12] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127, 2014.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[14] Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*, 2023.

[15] Christina Heinze-Deml, J. Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2017.

[16] Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference on Learning Theory*, 2018.

[17] Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, 2021.

[18] Tengyuan Liang. Blessings and curses of covariate shifts: Adversarial learning dynamics, directional convergence, and equilibria. *Journal of Machine Learning Research*, 25(140):1–27, 2024.

[19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015.

[20] Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

[21] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[22] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[23] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[24] Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[25] Judea Pearl. *Causality*. Cambridge University Press, second edition, 2009.

[26] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[27] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[28] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

[29] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[30] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

[31] Bernhard Scholkopf, Dominik Janzing, J. Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *International Conference on Machine Learning*, 2012.

[32] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[33] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.

[34] James H. Stock and Francesco Trebbi. Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194, 2003.

[35] James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016.

[36] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

A Remaining Proofs

A.1 Proofs in Section 2

Proof of Proposition 1. Under Assumption 1, as $\tau > 0$ implies that $\mathbb{E}_{\mathcal{E}}[XX^\top]$ is invertible, we can deduce that $\beta_{\mathcal{E}} = (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\mathbb{E}_{\mathcal{E}}[XY]$ is uniquely defined. One can verify that

$$\mathbb{E}_{\mathcal{E}}[XY] = \mathbb{E}_{\mathcal{E}}[XX^\top]\beta^\star + \Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma + \Theta\mathbb{E}[Z](\mathbb{E}_{\mathcal{E}}[E])^\top\gamma = \mathbb{E}_{\mathcal{E}}[XX^\top]\beta^\star + \Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma. \tag{A.1}$$

Hence, we have

$$\beta_{\mathcal{E}} = (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\mathbb{E}_{\mathcal{E}}[XY] = \beta^\star + (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma.$$

Under Assumption 1, note that

$$\mathbb{E}_{\mathcal{E}}[XX^\top] = \begin{bmatrix} \Theta & \Delta \end{bmatrix}\begin{bmatrix} \mathbb{E}[ZZ^\top] & 0 \\ 0 & \mathbb{E}_{\mathcal{E}}[EE^\top] \end{bmatrix}\begin{bmatrix} \Theta^\top \\ \Delta^\top \end{bmatrix} + \tau^2 I_d. \tag{A.2}$$

Under Assumption 2, denote $\Gamma \in O_{d-(k+r)}$ whose column space spans the orthogonal complement of $[\Theta, \Delta]$. Then, from (A.2), we have

$$\mathbb{E}_{\mathcal{E}}[XX^\top] = \begin{bmatrix} \Theta & \Delta & \Gamma \end{bmatrix}\begin{bmatrix} \mathbb{E}[ZZ^\top] + \tau^2 I_k & 0 & 0 \\ 0 & \mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r & 0 \\ 0 & 0 & \tau^2 I_{d-(k+r)} \end{bmatrix}\begin{bmatrix} \Theta^\top \\ \Delta^\top \\ \Gamma^\top \end{bmatrix}, \tag{A.3}$$

where we use $\Gamma\Gamma^\top = I_d - \Theta\Theta^\top - \Delta\Delta^\top$. Therefore,

$$(\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1} = \begin{bmatrix} \Theta & \Delta & \Gamma \end{bmatrix}\begin{bmatrix} (\mathbb{E}[ZZ^\top] + \tau^2 I_k)^{-1} & 0 & 0 \\ 0 & (\mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r)^{-1} & 0 \\ 0 & 0 & (\tau^2 I_{d-(k+r)})^{-1} \end{bmatrix}\begin{bmatrix} \Theta^\top \\ \Delta^\top \\ \Gamma^\top \end{bmatrix}. \tag{A.4}$$

Therefore, we have

$$\begin{aligned} \beta_{\mathcal{E}} &= \beta^\star + (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma \\ &= \beta^\star + \Delta\left(\mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r\right)^{-1}\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma \\ &= \beta^\star + \Delta\gamma - \tau^2\Delta\left(\mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r\right)^{-1}\gamma, \end{aligned} \tag{A.5}$$

where the second equality uses $(\mathbb{E}_{\mathcal{E}}[XX])^{-1}\Delta = \Delta(\mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r)^{-1}$ deduced from (A.4). Now, if $\mathbb{E}_{\mathcal{S}}[EE^\top] \neq \mathbb{E}_{\mathcal{T}}[EE^\top]$, there must exist a nonzero coefficient γ such that

$$\tau^2\Delta\left(\mathbb{E}_{\mathcal{S}}[EE^\top] + \tau^2 I_r\right)^{-1}\gamma \neq \tau^2\Delta\left(\mathbb{E}_{\mathcal{T}}[EE^\top] + \tau^2 I_r\right)^{-1}\gamma,$$

which implies that $\beta_{\mathcal{S}} \neq \beta_{\mathcal{T}}$. □

Proof of Proposition 2. Note that $\alpha_{\mathcal{E}}^{\Theta} = (\Theta^\top\mathbb{E}_{\mathcal{E}}[XX^\top]\Theta)^{-1}\Theta^\top\mathbb{E}_{\mathcal{E}}[XY]$. From (A.1), we have $\Theta^\top\mathbb{E}_{\mathcal{E}}[XY] = \Theta^\top\mathbb{E}_{\mathcal{E}}[XX^\top]\beta^\star$. From (A.3), we have $\Theta^\top\mathbb{E}_{\mathcal{E}}[XX^\top] = (\mathbb{E}[ZZ^\top] + \tau^2 I_k)\Theta^\top$. Therefore,

$$\alpha_{\mathcal{E}}^{\Theta} = (\mathbb{E}[ZZ^\top] + \tau^2 I_k)^{-1}(\mathbb{E}[ZZ^\top] + \tau^2 I_k)\Theta^\top\beta^\star = \Theta^\top\beta^\star.$$

As a result, $\beta_{\mathcal{E}}^{\Theta} = \Theta\Theta^\top\beta^\star$, meaning that $\beta_{\mathcal{S}}^{\Theta} = \beta_{\mathcal{T}}^{\Theta}$ always holds regardless of γ. □

Proof of Proposition 3. Observe that

$$R_{\mathcal{T}}(\beta) = \|\beta_{\mathcal{T}} - \beta\|_{\mathbb{E}_{\mathcal{T}}[XX^\top]}^2 + \mathbb{E}_{\mathcal{T}}[Y^2] - \|\beta_{\mathcal{T}}\|_{\mathbb{E}_{\mathcal{T}}[XX^\top]}^2.$$

Hence, $R_\mathcal{T}(\beta_\mathcal{S}^\Theta) < R_\mathcal{T}(\beta_\mathcal{S})$ is equivalent to $\|\beta_\mathcal{T} - \beta_\mathcal{S}^\Theta\|^2_{\mathbb{E}_\mathcal{T}[XX^\top]} < \|\beta_\mathcal{T} - \beta_\mathcal{S}\|^2_{\mathbb{E}_\mathcal{T}[XX^\top]}$. From (A.5) and $\beta_\mathcal{S}^\Theta = \Theta\Theta^\top \beta^\star$, we have

$$\beta_\mathcal{T} - \beta_\mathcal{S}^\Theta = (I_d - \Theta\Theta^\top)\beta^\star + \Delta(\tau^2 I_r + \Lambda_\mathcal{T})^{-1}\Lambda_\mathcal{T}\gamma = \Delta\left[\Delta^\top\beta^\star + (\tau^2 I_r + \Lambda_\mathcal{T})^{-1}\Lambda_\mathcal{T}\gamma\right],$$
$$\beta_\mathcal{T} - \beta_\mathcal{S} = \Delta\left[(\tau^2 I_r + \Lambda_\mathcal{T})^{-1}\Lambda_\mathcal{T}\gamma - (\tau^2 I_r + \Lambda_\mathcal{S})^{-1}\Lambda_\mathcal{S}\gamma\right],$$

where the first equation uses $I_d = \Theta\Theta^\top + \Delta\Delta^\top$. From (A.3), we can derive the following and thus complete the proof,

$$\|\beta_\mathcal{T} - \beta_\mathcal{S}^\Theta\|^2_{\mathbb{E}_\mathcal{T}[XX^\top]} = \|\Delta^\top\beta^\star + (\tau^2 I_r + \Lambda_\mathcal{T})^{-1}\Lambda_\mathcal{T}\gamma\|^2_{\tau^2 I_r + \Lambda_\mathcal{T}},$$
$$\|\beta_\mathcal{T} - \beta_\mathcal{S}\|^2_{\mathbb{E}_\mathcal{T}[XX^\top]} = \|(\tau^2 I_r + \Lambda_\mathcal{S})^{-1}\Lambda_\mathcal{S}\gamma - (\tau^2 I_r + \Lambda_\mathcal{T})^{-1}\Lambda_\mathcal{T}\gamma\|^2_{\tau^2 I_r + \Lambda_\mathcal{T}}.$$

$\square$

A.2 Proofs in Section 3

Proof of Proposition 4. By definition of $R_\mathcal{E}$, one can deduce that

$$R_\mathcal{T}(\beta) - R_\mathcal{S}(\beta) = \langle\beta, (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})\beta\rangle - 2\langle\mathbb{E}_\mathcal{T}[XY] - \mathbb{E}_\mathcal{S}[XY], \beta\rangle + \mathbb{E}_\mathcal{T}[Y^2] - \mathbb{E}_\mathcal{S}[Y^2]. \tag{A.6}$$

From (A.1), we have

$$\mathbb{E}_\mathcal{T}[XY] - \mathbb{E}_\mathcal{S}[XY] = (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})\beta^\star + \Delta(\Lambda_\mathcal{T} - \Lambda_\mathcal{S})\gamma = (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})(\beta^\star + \Delta\gamma), \tag{A.7}$$

where the second equality uses $\Delta^\top\Delta = I_r$ from Assumption 1 and

$$\Sigma_\mathcal{T} - \Sigma_\mathcal{S} = \Delta(\Lambda_\mathcal{T} - \Lambda_\mathcal{S})\Delta^\top \tag{A.8}$$

from (A.2). Meanwhile, from (1.1) and $\mathbb{E}[U] = 0$, we have

$$\mathbb{E}_\mathcal{E}[Y^2] = \langle\beta^\star, \Sigma_\mathcal{E}\beta^*\rangle + \langle\gamma, \Lambda_\mathcal{E}\gamma\rangle + 2\langle\beta^\star, \mathbb{E}_\mathcal{E}[XE^\top]\gamma\rangle + \mathbb{E}[U^2].$$

Therefore,

$$\begin{aligned}
\mathbb{E}_\mathcal{T}[Y^2] - \mathbb{E}_\mathcal{S}[Y^2] &= \langle\beta^\star, (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})\beta^\star\rangle + \langle\gamma, (\Lambda_\mathcal{T} - \Lambda_\mathcal{S})\gamma\rangle + 2\langle\beta^\star, (\mathbb{E}_\mathcal{T}[XE^\top] - \mathbb{E}_\mathcal{S}[XE^\top])\gamma\rangle \\
&= \langle\beta^\star, (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})\beta^\star\rangle + \langle\gamma, (\Lambda_\mathcal{T} - \Lambda_\mathcal{S})\gamma\rangle + 2\langle\beta^\star, \Delta(\Lambda_\mathcal{T} - \Lambda_\mathcal{S})\gamma\rangle \\
&= \langle\beta^\star, (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})\beta^\star\rangle + \langle\Delta\gamma, (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})\Delta\gamma\rangle + 2\langle\beta^\star, (\Sigma_\mathcal{T} - \Sigma_\mathcal{S})\Delta\gamma\rangle,
\end{aligned} \tag{A.9}$$

where the second equality is due to (1.2) and Assumption 1 and the third equality uses $\Delta^\top\Delta = I_r$ and (A.8) again. Combining (A.6), (A.7), and (A.9), we have (3.1). $\square$

Proof of Proposition 5. By Proposition 4 and Young's inequality, $\forall\xi > 0$,

$$0 \leq R_\mathcal{T}(\beta) - R_\mathcal{S}(\beta) \leq (1 + \xi)\langle\beta, D\beta\rangle + (1 + \xi^{-1})\langle\beta^\star + \Delta\gamma, D(\beta^\star + \Delta\gamma)\rangle,$$

where the second term on the right-hand side does not depend on β. To bound the first term, plug in $\beta = V\alpha$

$$\langle\beta, D\beta\rangle = \langle\alpha\alpha^\top, V^\top DV\rangle \leq \|\alpha\|^2\|V^\top DV\|_\mathrm{F}.$$

Apply Young's inequality again, $\forall\zeta > 0$,

$$\|\alpha\|^2\|V^\top DV\|_\mathrm{F} \leq \frac{\zeta}{2}\|\alpha\|^4 + \frac{1}{2\zeta}\|V^\top DV\|^2_\mathrm{F}.$$

The proof follows. $\square$

Proof of Proposition 6. Let $\xi_V = \text{grad }\Phi_{v,\eta}(V)$. By (3.9), we know grad $\bar{\Phi}_{v,\eta}(V) = P_V\text{grad }\Phi_{v,\eta}(V) = P_V(\xi_V)$, where

$$P_V(\xi_V) = (I_d - VV^\top)\xi + V\text{skew}(V^\top\xi) \in T_X\text{St}(d,\ell) . \tag{A.10}$$

We calculate ξ_V with an application of the Envelope Theorem, see for example [23]. Indeed, the inner minimization, $\min_{\alpha \in \mathbb{R}^\ell} F_{v,\eta}(V,\alpha)$, admits a unique minimizer $\alpha_V := (V^\top\Sigma_\mathcal{S}V + vI_\ell)^{-1}V^\top\mathbb{E}_\mathcal{S}[XY]$. And so,

$$\begin{aligned}
\xi_V = \text{grad }\Phi_{v,\eta}(V) &= \text{grad }F_{v,\eta}(V,\alpha)\big|_{\alpha=\alpha_V} \\
&= (\Sigma_\mathcal{S}V\alpha_V - \mathbb{E}_\mathcal{S}[XY])\,\alpha_V^\top + \eta DVV^\top DV .
\end{aligned}$$

We evaluate $\text{skew}(V^\top\xi)$. Note that

$$\begin{aligned}
V^\top\xi &= \left(V^\top\Sigma_\mathcal{S}V\alpha_V - V^\top\mathbb{E}_\mathcal{S}[XY]\right)\alpha_V^\top + \eta V^\top DVV^\top DV \\
&= -v\alpha_V\alpha_V^\top + \eta V^\top DVV^\top DV
\end{aligned}$$

is a symmetric matrix, thus $\text{skew}(V^\top\xi) = 0$. Therefore:

$$\text{grad }\bar{\Phi}_{v,\eta}(V) = P_V(\xi_V) = (I_d - VV^\top)\left((\Sigma_\mathcal{S}V\alpha_V - \mathbb{E}_S[XY])\alpha_V^\top + \eta DVV^\top DV\right)$$

as required. $\qquad\qquad\square$

A.3 Proofs in Section 4

Lemma 1. *For any positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $V \in \text{St}(d,\ell)$, any $v > 0$, the following inequality holds in Loewner ordering,*

$$\tfrac{v}{v+\lambda_{\max}(\Sigma)} \cdot I_d \preceq I_d - \Sigma^{1/2}V(V^\top\Sigma V + vI_\ell)^{-1}V^\top\Sigma^{1/2} \preceq I_d ,$$

$$\Sigma^{1/2}V(V^\top\Sigma V + vI_\ell)^{-2}V^\top\Sigma^{1/2} \preceq \tfrac{1}{4v} \cdot I_d .$$

Proof of Lemma 1. The proof uses singular value decomposition (SVD) of $\Sigma^{1/2}V$. Let $\Sigma^{1/2}V = ADB^\top$, $A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{\ell\times\ell}$, $D \in \mathbb{R}^{d\times\ell}$, $A^\top A = I_d$, $BB^\top = I_\ell$ and D is rectangular diagonal. We tackle the first inequality. Note

$$\begin{aligned}
I_d - \Sigma^{1/2}V(V^\top\Sigma V + vI_\ell)^{-1}V^\top\Sigma^{1/2} &= I_d - ADB^\top(BD^\top DB^\top + vI_\ell)^{-1}BD^\top A^\top \\
&= A(I_d - D(D^\top D + vI_\ell)^{-1}D^\top)A^\top ,
\end{aligned}$$

where the second equality uses the Woodbury matrix identity. Here, $I_d - D(D^\top D + vI_\ell)^{-1}D^\top$ is a diagonal matrix with entries

$$(I_d - D(D^\top D + vI_\ell)^{-1}D^\top)_{ii} = \begin{cases} \frac{v}{v+D_{ii}^2} & \text{for } i \leq \ell \\ 1 & \text{otherwise} \end{cases} .$$

Note $D_{ii}^2 = \lambda_i(V^\top\Sigma V)$ where λ_i denotes the i-th largest eigenvalue. We can verify $\lambda_i(V^\top\Sigma V) \leq \lambda_{\max}(\Sigma)$ as $V \in \text{St}(d,\ell)$, thus we arrive at $\frac{v}{v+\lambda_{\max}(\Sigma)} \leq \frac{v}{v+D_{ii}^2} \leq 1$. The second inequality is established in a similar fashion. Reusing the SVD, we have

$$\Sigma^{1/2}V(V^\top\Sigma V + vI_\ell)^{-2}V^\top\Sigma^{1/2} = AD(D^\top D + vI_\ell)^{-2}D^\top A^\top .$$

The entries of the inner diagonal matrix read,

$$(D(D^\top D + vI_\ell)^{-2}D^\top)_{ii} = \begin{cases} \frac{D_{ii}^2}{(D_{ii}^2+v)^2} & \text{for } i \leq \ell \\ 0 & \text{otherwise} \end{cases} .$$

Then, $\frac{D_{ii}^2}{(D_{ii}^2+v)^2} \leq \max_{d>0}\frac{d^2}{(d^2+v)^2} = \frac{1}{4v}$. $\qquad\qquad\square$

Proof of Theorem 1. Let $M := \Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}}$. Under Assumption 3, M is a positive definite matrix. The first-order condition of the Stiefel manifold optimization (3.6) reads

$$(I_d - VV^\top)(\mathbb{E}_S[XY] - \mathbb{E}_S[XX^\top]V\alpha_V)\alpha_V^\top = \eta(I_d - VV^\top)DVV^\top DV . \tag{A.11}$$

Define the canonical angles $A := V^\top\Delta \in \mathbb{R}^{\ell\times r}$, and then take the Frobenius norm of the right-hand side of (A.11), we have

$$
\begin{aligned}
&\eta^2\mathrm{Tr}\big[V^\top DVV^\top D(I_d - VV^\top)DVV^\top DV\big] \\
&= \eta^2\mathrm{Tr}\big[AMA^\top AM(I_r - A^\top A)MA^\top AMA^\top\big] \\
&\geq \eta^2\lambda_{\min}(I_r - A^\top A)\cdot\lambda_{\min}(M)\cdot\mathrm{Tr}\big[AMA^\top AMA^\top AMA^\top\big] \\
&\geq \eta^2\lambda_{\min}(I_r - A^\top A)\cdot\lambda_{\min}(M)^4\cdot\|AA^\top\|_{\mathrm{op}}^3 .
\end{aligned}
$$

In the inequalities above, we use the fact that if $B \succeq C$, then $DBD^\top \succeq DCD^\top$, in Loewner ordering. The last inequality leverages the additional facts that $\mathrm{Tr}\big[(AMA^\top)^3\big] \geq \|AMA^\top\|_{\mathrm{op}}^3$ and $\|AMA^\top\|_{\mathrm{op}} \geq \lambda_{\min}(M)\|AA^\top\|_{\mathrm{op}}$.

The Frobenius norm on the left-hand side of (A.11) is

$$\|(I_d - VV^\top)(\mathbb{E}_S[XY] - \mathbb{E}_S[XX^\top]V\alpha_V)\|^2 \cdot \|\alpha_V\|^2 . \tag{A.12}$$

Recall the two facts in Lemma 1,

$$\frac{\upsilon}{\upsilon+\lambda_{\max}(\Sigma)}\cdot I_d \preceq I_d - \Sigma^{1/2}V(V^\top\Sigma V + \upsilon I_\ell)^{-1}V^\top\Sigma^{1/2} \preceq I_d , \tag{A.13}$$

$$\Sigma^{1/2}V(V^\top\Sigma V + \upsilon I_\ell)^{-2}V^\top\Sigma^{1/2} \preceq \tfrac{1}{4\upsilon}\cdot I_d . \tag{A.14}$$

Now we control each term in (A.12).

For the first term in (A.12), we have

$$
\begin{aligned}
&(I_d - VV^\top)(\mathbb{E}_S[XY] - \mathbb{E}_S[XX^\top]V\alpha_V) \\
&= (I_d - VV^\top)\big(I_d - \mathbb{E}_{\mathcal{S}}[XX^\top]V(V^\top\mathbb{E}_{\mathcal{S}}[XX^\top]V + \upsilon I_\ell)^{-1}V^\top\big)\mathbb{E}_{\mathcal{S}}[XY] .
\end{aligned}
$$

Therefore, denote $\Sigma := \mathbb{E}_{\mathcal{S}}[XX^\top]$

$$
\begin{aligned}
&\|(I_d - VV^\top)(\mathbb{E}_S[XY] - \mathbb{E}_S[XX^\top]V\alpha_V)\|^2 \\
&\leq \|\big(I_d - \mathbb{E}_{\mathcal{S}}[XX^\top]V(V^\top\mathbb{E}_{\mathcal{S}}[XX^\top]V + \upsilon I_\ell)^{-1}V^\top\big)\mathbb{E}_{\mathcal{S}}[XY]\|^2 \\
&= \|\Sigma^{1/2}\big(I_d - \Sigma^{1/2}V(V^\top\Sigma V + \upsilon I_\ell)^{-1}V^\top\Sigma^{1/2}\big)\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \\
&\leq \lambda_{\max}(\Sigma)\cdot\|\big(I_d - \Sigma^{1/2}V(V^\top\Sigma V + \upsilon I_\ell)^{-1}V^\top\Sigma^{1/2}\big)\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \quad \text{by (A.13)} \\
&\leq \lambda_{\max}(\Sigma)\cdot\|\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 .
\end{aligned}
$$

For the second term in (A.12), we know

$$
\begin{aligned}
&\|\alpha_V\|^2 \\
&= \|(V^\top\Sigma V + \upsilon I_\ell)^{-1}V^\top\mathbb{E}_S[XY]\|^2 \\
&\leq \lambda_{\max}\big(\Sigma^{1/2}V(V^\top\Sigma V + \upsilon I_\ell)^{-2}V^\top\Sigma^{1/2}\big)\cdot\|\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \quad \text{by (A.14)} \\
&\leq \frac{1}{4\upsilon}\|\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 .
\end{aligned}
$$

Put things together, for $V \in \mathcal{S}_\Delta(\delta)$, we have that the canonical angles $A = V^\top\Delta \in \mathbb{R}^{\ell\times r}$ satisfies $1 - \lambda_{\max}(A^\top A) \geq 1 - \|A\|_{\mathrm{op}}^2 \geq \delta$, and that

$$\eta^2\delta\lambda_{\min}(M)^4\cdot\|A^\top A\|_{\mathrm{op}}^3 \leq \lambda_{\max}(\Sigma_{\mathcal{S}})\cdot\|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2\frac{1}{4\upsilon}\|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 .$$

Therefore, we have proved

$$\|V^\top \Delta\|_{\mathrm{op}}^6 \leq \frac{\delta^{-1}\lambda_{\max}(\Sigma_{\mathcal{S}})\|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^4}{\lambda_{\min}(\Delta^\top D\Delta)^4}\frac{1}{4\eta^2 \upsilon}\ .$$

□

Proof of Theorem 2. In view of (3.1), by Young's inequality, we have

$$R_{\mathcal{T}}(\beta^{\eta,\upsilon}) - R_{\mathcal{S}}(\beta^{\eta,\upsilon}) \leq (1+\epsilon)\langle \beta^\star + \Delta\gamma, D(\beta^\star + \Delta\gamma)\rangle + (1+\epsilon^{-1})\langle V\alpha_V, DV\alpha_V\rangle\ .$$

Note

$$\begin{aligned}
\langle V\alpha_V, DV\alpha_V\rangle &\leq \|\alpha_V\|^2 \cdot \|V^\top DV\|_{\mathrm{op}} \\
&\leq \|\alpha_V\|^2 \cdot \|V^\top \Delta M\Delta^\top V\|_{\mathrm{op}} \\
&\leq \|\alpha_V\|^2 \cdot \lambda_{\max}(M)\|V^\top \Delta\|_{\mathrm{op}}^2\ .
\end{aligned}$$

Recall the fact that

$$\|\alpha_V\|^2 \leq \frac{1}{4\upsilon}\|\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2$$

and the bound on $\|V^\top \Delta\|_{\mathrm{op}}^6$ established in Theorem 1, we finish the proof after separating out the terms independent of η, υ and defining them as $\mathsf{S}_{\epsilon,\delta}$. □