Estimating and Explaining Model Performance When Both Covariates and Labels Shift

Lingjiao Chen[1], **Matei Zaharia**[1], **James Zou**[1,2]
[1]Department of Computer Sciences, [2]Department of Biomedical Data Science

Stanford University

Abstract

Deployed machine learning (ML) models often encounter new user data that differs from their training data. Therefore, estimating how well a given model might perform on the new data is an important step toward reliable ML applications. This is very challenging, however, as the data distribution can change in flexible ways, and we may not have any labels on the new data, which is often the case in monitoring settings. In this paper, we propose a new distribution shift model, Sparse Joint Shift (SJS), which considers the joint shift of both labels and a few features. This unifies and generalizes several existing shift models including label shift and sparse covariate shift, where only marginal feature or label distribution shifts are considered. We describe mathematical conditions under which SJS is identifiable. We further propose SEES, an algorithmic framework to characterize the distribution shift under SJS and to estimate a model's performance on new data without any labels. We conduct extensive experiments on several real-world datasets with various ML models. Across different datasets and distribution shifts, SEES achieves significant (up to an order of magnitude) shift estimation error improvements over existing approaches.

1 Introduction

Deployed machine learning (ML) models often face new data different from their training data. For example, mismatch of deployment-development data in geographical locations [21], demographic features [16], and label balance [20] is widely observed and known to affect model performance. Thus, estimating and explaining how a model's performance changes on the new data is an important step toward reliable ML applications.

Estimating and explaining performance shift is challenging for several reasons, however. One major challenge is that the data distribution might shift in flexible ways. Another obstacle is that we often do not have labels on the new data, especially in ML monitoring applications. Without any assumption on the distribution shift, it's impossible to estimate how well the model would perform on the unlabeled new data. Previous work often assumes (i) label shift [22], where feature distributions conditional on the labels are fixed, or (ii) covariate shift [32], where label distributions conditional on features stay the same. However, we often do not know whether the real data shift is limited to label or covariate shift, and naively applying estimation methods designed for one shift may produce inaccurate assessments [26]. Moreover, labels and features may shift simultaneously in practice, invalidating these common assumptions.

To tackle the above challenges, here we propose a new distribution shift model, Sparse Joint Shift (SJS), to consider the joint shift of both labels and a few features. SJS assumes labels and a few features shift, but the remaining features' distribution conditional on the shifted features and labels is fixed. This unifies and generalizes sparse covariate shift and label shift: both of them are SJS,

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

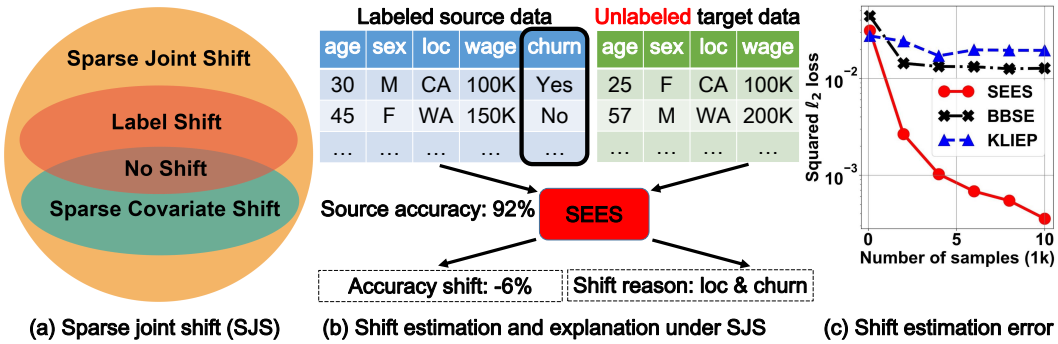


Figure 1: Overview of sparse joint shift (SJS). (a) Both label shift and sparse covariate shift are SJS, but SJS contains additional shifts as well. (b) illustrates SEES, a framework for performance shift estimation and explanation under SJS. Given labeled source and unlabeled target data, SEES exploits the joint shift modeled by SJS to estimate the model performance change and explain which factors drive the shift. In this example, the goal is to predict *churn*. (c) SEES significantly reduces the shift estimation error over existing methods when both labels and covariates shift.

but some SJS is not label or sparse covariate shift (Figure 1). We describe mathematical conditions under which SJS is provably identifiable: if the non-shifted features are weakly correlated, then the marginal feature distribution uniquely determines the joint distribution under SJS. This makes it possible to quantify the shift and estimate model performance on new data without any labels.

When SJS occurs: a motivating example. Consider Alice, a data scientist who built an ML model for customer churn prediction. Two years later, churn rate rose in some states but dropped in others, while the distribution of other features given label (churn) and location (state) remained. This shift in customer distribution is a natural SJS and challenging to estimate without labels on the new data.

Furthermore, we propose SEES, an algorithmic framework for performance shift estimation and explanation under SJS. SEES exploits correlation shifts between features and labels modeled by SJS to improve performance estimation accuracy over existing methods, such as BBSE [22] for label shift and KLIEP [32] for covariate shift (Figure 1). It uses the identified shifted features and labels as a natural explanation of the performance shift. We present an extensive empirical evaluation on real world datasets with both simulated SJS and real shifts. Our experiments validate the effectiveness of SEES: the estimation error of SEES is often an order of magnitude smaller than that of previous state-of-the-art methods. All together, **our contributions are**:

1. We formulate and study sparse joint shift (SJS), a new distribution shift model that considers the joint shift of both labels and features. We show how it unifies and generalizes existing shift models and when it is identifiable with unlabeled data.
2. We propose SEES, a general framework for performance shift estimation and explanation under SJS. We design efficient substantiations of SEES for different data types.
3. We provide comprehensive empirical analysis of our methods. On real world datasets with natural distribution shifts, we found SEES leads to up to 66% performance estimation error reduction over standard approaches.

Related Work. **Label shift:** Label shift has been observed and studied in various domains, such as epidemiologic [5], economics [23], and data mining [13]. Recently, there is an increasing interest in model evaluation and development under under label shift. For example, [22, 3] study how to quantify label shift and evaluate an ML model's performance under label shift accordingly. [40] gives an algorithm for active learning under label shift. Online adaptation to label shift [37] is also shown to be feasible. Label shift can be viewed as a subset of our proposed shift model SJS. Although simpler and easier to estimate, label shift does not capture the joint shift of labels and features. Thus, methods developed for label shift may not work well under SJS.

Covariate shift: Covariate shift [30, 38, 24] is perhaps the most widely adopted assumption in data distribution shift. Since studied in the seminal work [30], various methods have been proposed to estimate covariate shift, including KLIEP [33], KMM [15], and IWCV [31]. Adaptation to covaraite shift has been found useful in many applications, such as spam filtering [4], emotion recognition [19] and human activity detection [18]. More recently, covariate shift adaptation is jointly optimized with model robustness [29], fairness [28], and conformal prediction [35].

Unsupervised model performance evaluation: Model performance evaluation without labels has received relatively limited attention. Domain-specific models' performance can be estimated via certain statistics, such as confidence score [17], rotation prediction [9] and feature statistics of the datasets sampled from a meta-dataset [10] for image recognition. General model evaluation often relies on different assumptions and accessibility [14, 7, 6, 12, 8, 36]. [7] assumes covariate shift and requires users to provide an approximation (slice) of the shifted features, while [6] needs white-box access to the ML models to train an ensemble as a reference. [12] assumes the label distributions are known, while [8] needs a feature independence structure given the labels. When a small number of labels can be obtained, [36] proposes an active model evaluation approach. To our knowledge, this is the first paper that explicitly models the joint shift of both labels and a few features with provable identification guarantees. Moreover, we do not require access to side information such as model design or metadata.

2 Preliminaries and Problem Statement

We start by giving the preliminaries and the problem of estimating and explaining performance shift.

Prediction tasks and ML models. In this paper we consider the standard classification task: given a d-dimensional feature vector $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ from the feature space $\mathcal{X}$, the goal is to predict its associated label $y \in \mathcal{Y}$ in the label space $\mathcal{Y}$. Let $f(\cdot) : \mathbb{R}^d \mapsto \mathcal{Y}$ denote an ML model designed for such a task. For simplicity, we assume that $\mathcal{Y} = \{1, 2, \cdots, L\}$. Given the model's prediction $f(\boldsymbol{x})$ and its true label y, its performance is quantified by some loss function $\ell(\cdot, \cdot)$. A popular choice is the standard 0-1 loss: $\ell(a, b) = \mathbb{1}_{a=b}$, which we focus on, but other losses are also applicable.

Joint distribution shift. The training and inference data for ML often come from two different distributions, referred to as source domain and target domain. Here, we consider the general case when the joint distribution vary across the source and target domains, and call this *joint distribution shift*. Formally, let $\mathbb{P}_s, \mathbb{P}_t : \mathcal{X} \times \mathcal{Y} \mapsto [0, 1]$ denote the source and target domains, and p_s, p_t be their probability density (or mass) functions. Then joint distribution shift means $p_s(\boldsymbol{x}, y) \neq p_t(\boldsymbol{x}, y)$.

Problem statement. Suppose we are given a labeled dataset $D_s \triangleq \{(\boldsymbol{x}^{s,i}, y^{s,i})\}_{i=1}^{n_s}$ from the source distribution $\mathbb{P}_s$, an unlabeled dataset $D_t \triangleq \{(\boldsymbol{x}^{t,i})\}_{i=1}^{n_t}$ from the target distribution $\mathbb{P}_t$, and an ML model $f(\cdot)$ predicting the associated label given any feature vector $\boldsymbol{x}$. Our goal is to estimate how much performance changes from the source domain to the target domain. More formally, we aim at estimating the performance shift $\Delta \triangleq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_t}[\ell(f(\boldsymbol{x}), y)] - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_s}[\ell(f(\boldsymbol{x}), y)]$. This is challenging as we do not observe labels on the target domain. In many applications, attributing the performance shift to certain features is also desired. Thus, we are also interested in identifying a set of features to explain the performance shift.

3 SJS: A Tractable Unification of Label Shift and Sparse Covariate Shift

At a first glance, estimating the performance shift under joint distribution shift without observing labels from target domain seems hopeless: if the marginal feature distributions are identical for both domains, then observing the features alone should give 0 as the estimated performance shift. However, the label distribution given any feature on the target domain is arbitrary, and thus the estimated shift can be arbitrarily bad. In other words, joint distribution shift is not identifiable with no target labels.

To mitigate nonidentifiability, it's necessary to make additional assumptions. The most popular assumptions in literature are label shift [22] and covariate shift [30]. Label shift assumes that only label distribution may change, but the feature distribution given a label remains, i.e., $p_s(\boldsymbol{x}|y) = p_t(\boldsymbol{x}|y)$. On the other hand, covariate shift assumes that feature distribution can shift, but the label

distribution given the features is fixed, i.e., $p_s(y|\boldsymbol{x}) = p_t(y|\boldsymbol{x})$. However, those assumptions disallow simultaneous changes of both features and labels, which often happen in real-world data [21, 27, 34]. To enable joint feature and label estimation which is tractable, we introduce a subclass of joint distribution shift, *Sparse Joint Shift* (SJS), as follows.

Definition 1 (Sparse Joint Shift (SJS))**.** *Suppose for an integer $m \leq d$ and an index set $\mathcal{I} \subset [d]$ with size at most m (i.e., $|I| \leq m$), $p_s(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I, y) = p_t(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I, y)$. Then we say the source and target pair (p_s, p_t) is under m-Sparse Joint Shift, or m-SJS. Here, $I^c \triangleq [d] - I$. We call I the shift index set.*

Roughly speaking, SJS allows both labels and a few features to shift, but assumes the remaining features' conditional distribution to stay the same. Section 1 gives one example when SJS occurs, and more examples and discussions can be found in the appendix. Next, we will study when this assumption allows tractable performance shift estimation. All proofs are left to the appendix.

3.1 When is sparse joint shift identifiable?

Recall that additional assumptions are needed because the general joint distribution shift is not identifiable. However, when $m = d$, m-SJS simply becomes general joint distribution shift. Thus, it is worthy understanding when m-SJS resolves the identifiability issue. To do so, let us first formally introduce the notation of identifiability.

Definition 2 (Identifiable)**.** *Suppose the source-target tuple (p_s, p_t) is under m-SJS. (p_s, p_t) is identifiable if and only if for any alternative distribution $p_a(\boldsymbol{x}, y)$, if $p_a(\boldsymbol{x}) = p_t(\boldsymbol{x})$ and $\exists \mathcal{J} \subset [d], |\mathcal{J}| \leq m$, such that $p_a(\boldsymbol{x}_{\mathcal{J}^c}|\boldsymbol{x}_{\mathcal{J}}, y) = p_s(\boldsymbol{x}_{\mathcal{J}^c}|\boldsymbol{x}_{\mathcal{J}}, y)$, then $p_a(\boldsymbol{x}, y) = p_t(\boldsymbol{x}, y)$.*

The identifiability can be easily interpreted in words: If a joint feature and label distribution matches the target feature distribution and satisfies the m-SJS requirement together with the source distribution, it has to be the target distribution. The following statement shows when (p_s, p_t) is identifiable.

Theorem 1. *Suppose (p_s, p_t) is under m-SJS. Assume for any set $\mathcal{J} \subset [d], |\mathcal{J}| \leq m$ and any fixed $\bar{\boldsymbol{x}} \in \mathcal{X}$, the probability density (or mass) functions $\{p_s(\boldsymbol{x}_{J^c \cap I^c}, \boldsymbol{x}_{\mathcal{J} \cup I} = \bar{\boldsymbol{x}}_{\mathcal{J} \cup I}, y = i)\}_{i=1}^L$ are linearly independent. Then (p_s, p_t) is identifiable.*

This statement sheds light on why uniquely identifying the target distribution without target label is feasible under sparse joint shift. Roughly speaking, m-SJS requires that given the shifted features and labels, the remaining features' distribution remains the same on both domains. If those remaining features are different enough (linear independence), they can uniquely determine the distribution of the shifted features and labels. We stress that the linear independence is necessary: if it does not hold, then for any m, we can always find some source-target pair (p_s, p_t) which is not identifiable. Linear independence implicitly requires sparsity: if $m > d/2$, then $J^c \cap I^c$ can be empty and the linear independence does not hold. In other words, the sparsity is necessary for the shift to be identifiable.

3.2 How does SJS relate to label shift and covariate shift?

A natural question is how does SJS relates to standard label shift and covariate shift. To answer this, let us first introduce label and sparse covariate shift formally.

Definition 3. *The source and target (p_s, p_t) is under Label Shift iff $p_s(\boldsymbol{x}|y) = p_t(\boldsymbol{x}|y)$, and under m-Sparse Covariate Shift iff $p_s(\boldsymbol{x}_{I^c}, y|\boldsymbol{x}_I) = p_t(\boldsymbol{x}_{I^c}, y|\boldsymbol{x}_I)$ for some index set I with size $m < d$.*

Now we are ready to answer the above question.

Theorem 2. *If (p_s, p_t) is under label shift, then it is also under 0-SJS. If (p_s, p_t) is under m-sparse covariate shift, then it is also under m-SJS. In addition, there exists (p_s, p_t) under m-SJS such that it is under neither label shift or covariate shift.*

There are several takeaways. First, label shift implies SJS without additional requirements. In fact, as certain distribution pairs are under SJS but not label shift, SJS is strictly more general than label shift. Second, SJS also includes sparse covariate shift. When $m = d$, SJS completely unifies both label shift and covariate shift, though it is not identifiable. Identifiable SJS, on the other hand, unifies label shift and sparse covariate shift. Finally, SJS also allows shifts not covered by label shift and covaraite shift: the correlation between label and (a set of) features can be shifted.

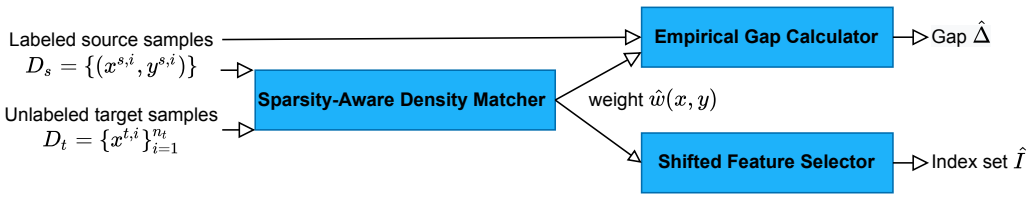


Figure 2: How SEES works. Given labeled source and unlabeled target data, SEES uses a sparsity-aware density matcher to learn a weight function $\hat{w}(\bar{\boldsymbol{x}}, y)$. Next, an empirical gap calculator computes the performance gap $\hat{\Delta}$ by weighing the source samples with the learned $\hat{w}(\bar{\boldsymbol{x}}, y)$. The shifted feature selector extracts the features $\hat{I}$ on which the weight function depends heavily.

4 Shift Estimation and Explanation under Sparse Joint Shift

Now we present SEES, an algorithmic framework to estimate and explain the performance shift Δ when the source and target domain is under m-SJS. As shown in Figure 2, it consists of three components: a sparsity-aware density matcher, an empirical gap calculator, and a shifted feature selector. Given the labeled source samples and unlabeled target samples, we first adopt the sparsity-aware density matcher to obtain an estimated ratio of the target and source density functions, denoted by $\hat{w}(\boldsymbol{x}, y)$. Next, the empirical gap calculator is responsible to estimate the performance shift Δ via appropriately reweighting source samples with ratio $\hat{w}(\boldsymbol{x}, y)$. Finally, the shifted feature selector picks a set of features as the explanation for the shift. We explain each component as follows.

4.1 Sparsity-aware density matcher

A key component of SEES is the sparsity-aware density matcher. Here we want to find some weight function $\hat{w}(\boldsymbol{x}, y)$ to induce an estimated target distribution $\hat{p}_t(\boldsymbol{x}, y) \triangleq \hat{w}(\boldsymbol{x}, y) \cdot p_s(\boldsymbol{x}, y)$. Our goals include (i) a small distance between the estimated target distribution and the true target distribution, (ii) m-SJS between the source and the estimated target distributions, and (iii) flexible parameterization of the weight function. To achieve those goals, we propose the following optimization framework

$$\begin{aligned} &\min_{w(\boldsymbol{x},y)\in\mathcal{W}} D(p_t(\boldsymbol{x}), \hat{p}_t(\boldsymbol{x})) \\ &\quad \text{s.t. } \hat{p}_t(\boldsymbol{x}) = \sum_{y=1}^{L} w(\boldsymbol{x}, y) \cdot p_s(\boldsymbol{x}, y), \text{ and } w(\boldsymbol{x}, y) \text{ depends on at most } m \text{ features of } \boldsymbol{x}. \end{aligned} \tag{4.1}$$

Here, $D(\cdot, \cdot)$ is some distance metric that measures the difference between two density functions. We minimize the distance between the induced feature density $\hat{p}_t(\boldsymbol{x})$ and the target feature density $p_t(\boldsymbol{x})$. The minimization is not over joint label and feature distributions since target labels are not available. The induced feature density function can be easily derived from source density function and the weight function, encoded in the first constraint. m-SJS is enforced by the second constraint: m-SJS means given m features and labels, the distributions of remaining features are fixed across source and the induced domain, which holds if and only if their density ratio $w(\boldsymbol{x}, y)$ only depends on those m features. $\mathcal{W}$ represents the set of all feasible weight functions. Different parameterization can be easily realized by adopting different $\mathcal{W}$. Assume access to density functions $p_s(\boldsymbol{x}, y)$ and $p_t(\boldsymbol{x})$, and a weight function set $\mathcal{W}$ containing the true weight $w^*(\boldsymbol{x}, y) \triangleq \frac{p_t(\boldsymbol{x},y)}{p_s(\boldsymbol{x},y)}$. One can easily show the above optimization returns the true weight function $w^*(\boldsymbol{x}, y)$ for identifiable m-SJS.

One benefit of the above framework is the flexibility of concrete instantiations. Different design choices, including the distance metric $D(\cdot, \cdot)$ and the weight parameterization space $\mathcal{W}$, can fit different feature types, incorporate domain knowledge, and tradeoff different sample and computational complexity. We give two instantiations of the above optimization: SEES-c for continuous features, and SEES-d for discrete features.

SEES-c: SEES for continuous features. For continuous features, we use KL-divergence $D_{KL}(\cdot, \cdot)$ as the distance metric, and initialize the parameterization space $\mathcal{W}$ by linear combinations of K fixed basis functions, $\phi_1(\boldsymbol{x}, y), \phi_2(\boldsymbol{x}, y), \cdots, \phi_K(\boldsymbol{x}, y)$. That is to say, $\mathcal{W} = \{w(\boldsymbol{x}, y) | w(\boldsymbol{x}, y) =$

$\sum_{k=1}^{K} a_{k,y}\phi_k(\boldsymbol{x}, y), a_{k,y} \geq 0, \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_s}\left[\sum_{k=1}^{K} a_{k,y}\phi_k(\boldsymbol{x}, y)\right] = 1, \alpha_{k,y} \geq 0\}$. The last two constraints ensure $w(\boldsymbol{x}, y) \cdot p_s(\boldsymbol{x}, y)$ is a valid probability density. Those basis functions encode users' prior knowledge about the shift. A simple choice, for example, is linear functions (when $\boldsymbol{x}_k \geq 0$): setting $K = d$ and $\phi_k(\boldsymbol{x}, y) = \boldsymbol{x}_k$. To model the dependence on different features, let e_i denote all indexes k such that $\phi_k(\cdot)$ depends on feature $\boldsymbol{x}_i$, and introduce a vector $\boldsymbol{\beta} \in \mathbb{R}^d$ such that $\boldsymbol{\beta}_i \triangleq \sqrt{\sum_{k\in e_i}\sum_{y=1}^{L} a_{k,y}^2}$. The feature dependence requirement in Problem 4.1 is equivalent to sparsity constraint $\|\boldsymbol{\beta}\|_0 \leq s$. We can relax the 0-norm by 1-norm and obtain one instantiation as

$$\max_{a_{1,1},a_{1,2},\cdots,a_{K,L}} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_t}\left[\log\sum_{y=1}^{L} p_s(y|\boldsymbol{x})\sum_{k=1}^{K} a_{k,y}\phi_k(\boldsymbol{x}, y)\right] + \eta\sum_{i=1}^{d}\sqrt{\sum_{k\in e_i}\sum_{y=1}^{L} a_{k,y}^2}$$
$$s.t.\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_s}\left[\sum_{k=1}^{K} a_{k,y}\phi_k(\boldsymbol{x}, y)\right] = 1, \alpha_{k,y} \geq 0$$

where $\eta > 0$ controls the trade-off between sparsity and the KL distance. One benefit of this instantiation is computational efficiency: the constraint is linear in the optimization variables, and the objective is convex. Thus, the problem is convex and can be efficiently solved. The label distribution given feature $p_s(y|\boldsymbol{x})$ is unknown but can be approximated by the ML model $f(\cdot)$ trained on the source domain. Given finite samples, the expectations can be replaced by their empirical estimation.

SEES-d: SEES for discrete features. Another interesting instantiation exists for discrete features. With no prior knowledge, we parameterize $\mathcal{W}$ to include all possible m-SJS: specifically, $\mathcal{W}$ contains all tuple $(J, w_J(\boldsymbol{x}_J, y))$, where index set $J \subset [d], |J| = m$ represents the shifted features, and weight function $w_J(\boldsymbol{x}_J, y)$ only depends on $\boldsymbol{x}_J$ and y.

Features only take finite values, so we can view the density (mass) functions as vectors with finite dimensions. Thus, we adopt the squared ℓ_2 distance, i.e., $D(\boldsymbol{z}, \boldsymbol{z}') = \sum_{i=1}^{|\boldsymbol{z}|}(\boldsymbol{z}_i - \boldsymbol{z}'_i)^2$ to measure distance. However, naively measuring ℓ_2 distance between $p_t(\boldsymbol{x})$ and $\hat{p}_t(\boldsymbol{x})$ leads to a computational complexity exponential in d. Instead, we measure the distance on a set of marginal densities: given an index set J, for every index set with size $2s$ that contains J, denoted by κ, we measure the squared ℓ_2 distance between $p_t(\boldsymbol{x}_\kappa, f(\boldsymbol{x}))$ and $\hat{p}_t(\boldsymbol{x}_\kappa, f(\boldsymbol{x}))$, and then aggregate over κ. This design leads to the following instantiation of Problem 4.1

$$\min_{J, w_J(\boldsymbol{x},y)} \sum_{\kappa: J\subseteq\kappa, |\kappa|=2m}\sum_{\bar{f}=1}^{L}\sum_{\bar{\boldsymbol{x}}_\kappa\in\mathcal{X}_\kappa} \|p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2,\ s.t. |J| = m \quad (4.2)$$

where $p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f})$ and $p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})$ are short for $p_t(\boldsymbol{x}_\kappa = \bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{x}) = \bar{f})$ and $p_s(\boldsymbol{x}_\kappa = \bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{x}) = \bar{f}, y = \bar{y})$, respectively.

Compared to the naive approach, the above formulation is much more computationally efficient: the number of parameters in the above objective is only polynomial in the feature dimension d. For fixed J, the problem is simply a linear regression over the weight $w_J(\boldsymbol{x}_J, y)$ and thus can be efficiently solved. In practice, one can estimate $p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f})$ and $p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})$ via labeled source and unlabeled target samples, and then solve the empirical version of the above problem. Compared to using KL-divergence, solving the empirical version produces the correct shifted index set and a weight function close to the true weight $w^*(\boldsymbol{x}, y)$ (under mild conditions). This is formally stated as follows.

Theorem 3. *Consider when all features are discrete, i.e., for each i, $\boldsymbol{x}_i \in \{1, 2, \cdots, v\}$. Suppose (i) the source and target are under exact m-SJS, (ii) for any set $\mathcal{J} \subset [d], |\mathcal{J}| \leq m$ and any $\bar{\boldsymbol{x}} \in \mathcal{X}$, the marginal probability density (or mass) functions $\{p_s(f(\boldsymbol{x}), \boldsymbol{x}_{\mathcal{J}\cup I} = \bar{\boldsymbol{x}}_{\mathcal{J}\cup I}, y = i)\}_{i=1}^{d}$ are linearly independent, and (iii) $w(\boldsymbol{x}, y)$ is bounded by a constant M. Then there exists some constant c (independent of d, n_s and n_t), such that if $\sqrt{\frac{1}{2n_s}} + LM\sqrt{\frac{1}{2n_t}} < c/\sqrt{2m\log d + m\log v + 2\log L + \log 1/\delta}$, then with probability $1 - \delta$, (i) the index set $\hat{J}$ learned by Problem 4.2 matches the true shift index set I, and (ii) the produced weights $w_{\hat{J}}(\boldsymbol{x}_{\hat{J}}, y)$ satisfies $\left|w_{\hat{J}}(\boldsymbol{x}_{\hat{J}}, y) - w^*(\boldsymbol{x}, y)\right| \leq O\left(\sqrt{2m\log d + m\log v + 2\log L + \log 1/\delta}\left(\sqrt{\frac{1}{2n_s}} + LM\sqrt{\frac{1}{2n_t}}\right)\right)$.*

Roughly speaking, this statement ensures that, when source and target sample sizes are large enough, with high probability, the true shift index set can be identified with finite samples, and the error rate of the learned weight function is approximately the inverse of sample sizes' square root.

Comparisons of SEES-c and SEES-d. SEES-d enjoys mathematical guarantees, but SEES-c can be computationally more efficient. In practice, we can discretize continuous features to use SEES-d.

4.2 Empirical gap calculator and shifted feature selector

Now we explain how the other two components of SEES work. The empirical gap calculator computes the performance shift $\hat{\Delta}$ via three steps. First, it estimates the source performance by $\frac{1}{n_s}\sum_{i=1}^{n_s}\ell(\boldsymbol{x}^{s,i}, y^{s,i})$. Next, it estimates the performance on the induced target distribution. Note that the performance on the induced target domain is $\int \hat{p}_t(\boldsymbol{x}, y)\ell(\boldsymbol{x}, y)d\boldsymbol{x}dy = \int \hat{w}(\boldsymbol{x}, y)p_s(\boldsymbol{x}, y)\ell(\boldsymbol{x}, y)d\boldsymbol{x}dy = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_s}[\hat{w}(\boldsymbol{x}, y)\ell(\boldsymbol{x}, y)]$. Thus, we use the weighted loss on the source samples $\frac{1}{n_s}\sum_{i=1}^{n_s}\hat{w}(\boldsymbol{x}^{s,i}, y^{s,i})\ell(\boldsymbol{x}^{s,i}, y^{s,i})$ as the estimation. Finally, their difference, i.e., $\hat{\Delta} = \frac{1}{n_s}\sum_{i=1}^{n_s}(\hat{w}(\boldsymbol{x}^{s,i}, y^{s,i}) - 1)\ell(\boldsymbol{x}^{s,i}, y^{s,i})$ is returned as the estimated performance shift.

The shifted feature selector picks a set of features as the shift explanation. For discrete data, the weight function $\hat{w}(\boldsymbol{x}, y)$ learned by the density matcher's instantiation is parameterised as a shifted index $\hat{J}$ and the corresponding weight $\hat{w}_{\hat{J}}(\boldsymbol{x}_{\hat{J}}, y)$. Thus, a natural choice is to return $\hat{I} = \hat{J}$ as the explanation. For continuous data, the weight function is $\hat{w}(\boldsymbol{x}, y) = \sum_{k=1}^{K}\hat{a}_{k,y}\phi_k(\boldsymbol{x}, y)$, where $\hat{a}_{k,y}$ is learned by the corresponding instantiation. Recall that e_i denotes all basis functions that depends on feature i. Then $\hat{\boldsymbol{\beta}}_i \triangleq \sqrt{\sum_{k\in e_i}\sum_{y=1}^{L}\hat{a}_{k,y}^2}$ can be viewed as the total contribution of feature i. Thus, a simple choice is to pick features with the m largest contributions. Formally, we use $\hat{I} = \{i|\hat{\boldsymbol{\beta}}_i > \hat{\boldsymbol{\beta}}_{(d-m)}\}$, where $\hat{\boldsymbol{\beta}}_{(d-m)}$ is the $d-m$ smallest value in $\hat{\boldsymbol{\beta}}$.

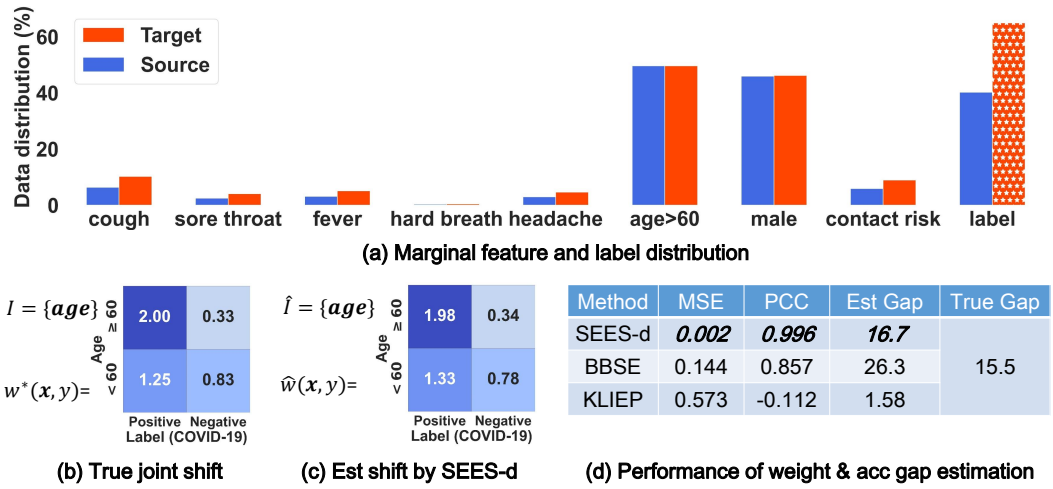
5 Experiments

In this section, we study the performance of SEES on several real world datasets with synthetic and natural distribution shifts. Our goal is four-fold: (i) understand when and how SEES estimates the performance shift, (ii) evaluate the trade-offs between the estimation performance reached by SEES and the required dataset sizes, (iii) explore the effects of shift sparsity on SEES' performance, and (iv) validate the effectiveness of SEES on datasets with real world distribution shifts.

Datasets, ML models and baselines. Six datasets are used for evaluation purposes. we first simulate various SJS on BANKCHURN [1], COVID-19 [2], and CREDIT [39] to systematically understand the performance of SEES. Next, we apply SEES on EMPLOY, INCOME, and INSURANCE [11] with real world distribution shifts and perform an in-depth analysis. We use a gradient boosting tree model as the ML model, and results for more models can be found in the Appendix. For comparison, we adopt two state-of-the-art methods for comparison: BBSE [22] for label shift and KLIEP [32] for covariate shift. More details on the experiments can be found in the Appendix.

Case study. We start with a case study on the dataset COVID-19. The task is to predict whether a person tests positive for COVID-19. We simulate a joint shift of the feature aged and label. Specifically, both source and target data contain 5000 young and aged individuals. The positive rate is 40% for both young and aged group from the source. In the target data, we raise the positive rate to 80% for aged group and 50% for young group. This simulates a shift due to a COVID variant more harmful to the elder than its ancestor. We adopt SEES-d as all features are categorical.

Figure 3 summarizes this case study. First note that identifying which feature is shifted is not obvious. As shown in Figure 3(a), the marginal distribution of most features except age and gender has changed from the source to the target. The actual joint shift (Figure 3(b)) is, on the other hand, due to age group and the labels. Identifying shifted age is challenging as the label on the target (last red bar in Figure 3(a)) also shifts but cannot be observed. On the other hand, SEES-d correctly identifies the shifted feature age, and produces a weight function close to the true weight (Figure 3(b) and (c)). This is primarily because SEES-d explicitly exploits the joint shift modeled by SJS. In fact, SEES-d's performance is significantly better than existing methods. As shown in Figure 3(d), the



Method	MSE	PCC	Est Gap	True Gap
SEES-d	*0.002*	*0.996*	*16.7*	15.5
BBSE	0.144	0.857	26.3	
KLIEP	0.573	-0.112	1.58	

Figure 3: A case study on the COVID-19 dataset. (a) The marginal distribution of labels and all features. The label on the target domain (the last red bar) is not observable. (b) The actual joint shift between source and target data. (c) The mean square error (MSE), pearson correlation coefficient (PCC) between learned and true weights, and the the estimated accuracy gap. Overall, SEES-d significantly improves estimation performance over existing methods.

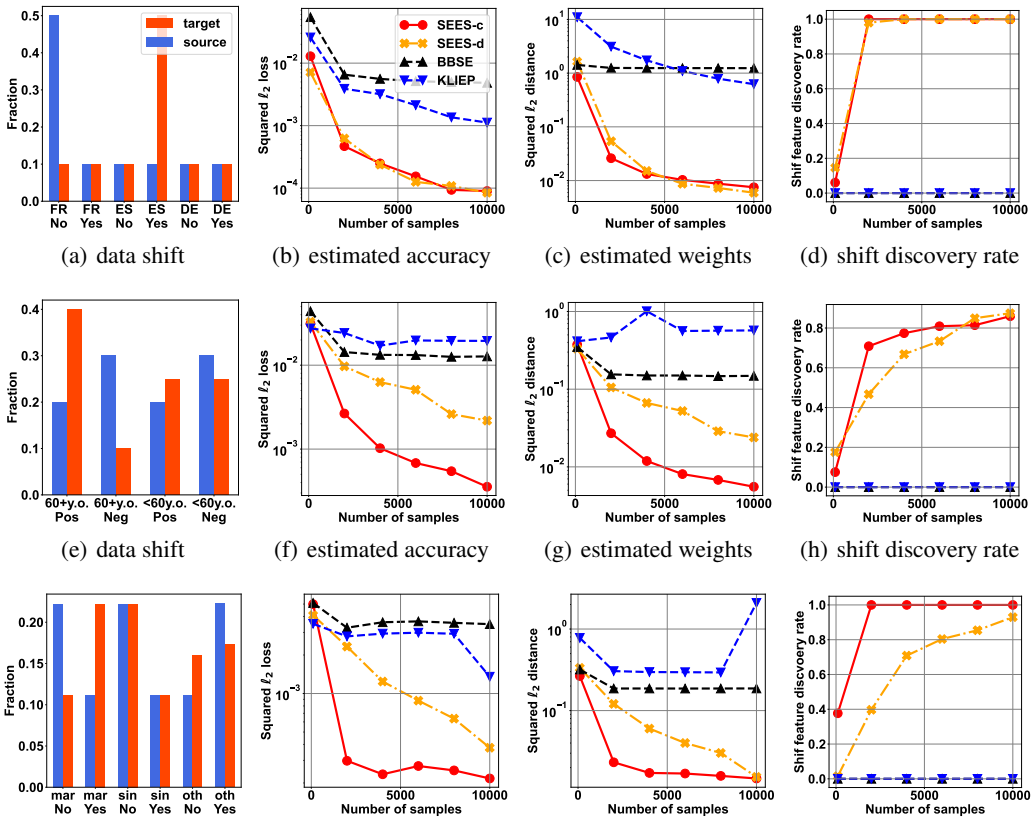
mean square error (MSE) between the true weights and learned weights is only 0.002 when adopting SEES-d, but 0.144 and 0.573 when using BBSE and KLIEP, respectively. The Pearson correlation coefficient (PCC) between the true weights and weights learned by SEES-d is 0.996, indicating a strong correlation. The weight estimation performance directly affects how precise the estimated accuracy gap is. The estimated gap $\hat{\Delta}$ of SEES-d is 16.7%, which is close to the true gap (15.5%). By contrast, BBSE tends to overestimate the gap (26.3$) while KLIEP underestimates it (1.58%).

Trade-offs between estimation performance and sample size. Next, we study the trade-offs between estimation performance and the number of available samples. For simplicity, we simulate various sparse joint shifts with $s = 1$ via (i) specifying marginal distribution of the shifted feature and labels first, and (ii) then drawing random samples from the original dataset conditional on values of specified labels and shifted feature. Same number of samples are allocated to both source and target datasets. Figure 4 shows the simulated data shift (column 1), the squared ℓ_2 loss of accuracy estimation (column 2), weight estimation (column 3), and the shifted feature discovery rate (column 4) for three datasets. Overall, we observe that the estimation error of SEES-c and SEES-d diminishes as the number of samples increases, while that of BBSE and KLIEP is almost flat.

Table 1: Root mean square error of estimated accuracy gap (%) under real shifts for a gradient boosting model. The numbers are averaged over all source-target pairs in each dataset. Results for other models (e.g., a neural network) can be found in the Appendix. For each dataset and ML model, SEES provides significant estimation error reduction over baselines.

EMPLOY				INCOME				INSURANCE			
SEES-c	SEES-d	BBSE	KLIEP	SEES-c	SEES-d	BBSE	KLIEP	SEES-c	SEES-d	BBSE	KLIEP
2.90	3.00	5.20	5.20	**1.90**	2.40	3.00	3.40	**1.70**	2.20	2.10	5.00

Effects of shift sparsity. We have focused on 1-SJS for simplicity, but how the sparsity s affects the estimation performance remains unknown. To answer this, we fix the number of samples to be 10,000, and then measure the performance of all compared methods for joint distribution of label and different number of features. Overall, we observe that the estimation error often grows as more features shift jointly with the labels. This is because more shifted features often implies higher complexity in the distribution shifts and thus more parameters to estimate. It is worthy noting that 0-SJS degenerates to label shift, and the performance of SEES-c and SEES-d is slightly worse



(a) data shift (b) estimated accuracy (c) estimated weights (d) shift discovery rate

(e) data shift (f) estimated accuracy (g) estimated weights (h) shift discovery rate

(i) data shift (j) estimated accuracy (k) estimated weights (l) shift discovery rate

Figure 4: Trade-offs between shift estimation and sample size. We vary the shifted features (first column), and measure performance of the estimated accuracy shift (second column), estimated weights (third column), and how often the true shifted features are discovered (last column). The first, second, and third row corresponds to dataset BANKCHURN, COVID-19, and CREDIT, respectively. Overall, both SEES-c and SEES-d consistently outperforms existing estimation approaches.

than BBSE (designed for label shift) under 0-SJS. This is a trade-off for covering more general distribution shifts. More details can be found in the appendix.

Accuracy gap estimation on real shifts. Finally we validate the effectiveness of SEES on accuracy estimation with real world shifts. EMPLOY and INCOME are partitioned by geography (states) and INSURANCE is partitioned by time (year). For each partition pair, we train a gradient boosting model on one and estimate its performance on the other. Table 1 shows the estimation error averaged over all partition pairs for each dataset. SEES consistently outperforms BBSE and KLIEP, and reduces the estimation error by up to 66% (1-1.7/5.0). We also evaluate other models (including a neural network) and observe similar results. More results can be found in the Appendix.

6 Conclusion

In this paper, we propose Sparse Joint Shift (SJS), a new distribution shift model that accounts for both label and covariate shifts. We show how SJS unifies and generalizes existing distribution shift models and remains identifiable under reasonable assumptions. We develop SEES, an algorithmic framework for unsupervised model performance estimation and explanation under SJS. Both theoretical analysis and empirical study validates the effectiveness of SEES. Our work contributes to making ML more reliable when data can change. A natural next step is how to improve estimation performance under SJS when a small number of target labels can be queried. This paper focuses on model performance estimation and explanation. Developing ML models robust to different SJS is also an open question.

Acknowledgement

This work was supported in part by a Google PhD Fellowship, a Sloan Fellowship, NSF CCF 1763191, NSF CAREER AWARD 1651570 and 1942926, NIH P30AG059307, NIH U01MH098953, grants from the Chan-Zuckerberg Initiative, Sutherland, and affiliate members and other supporters of the Stanford DAWN project, including Meta, Google, and VMware. We also thank anonymous reviewers for helpful discussion and feedback.

References

[1] The BANKCHURN dataset. `https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction/notebook`. [Accessed 2022].

[2] The COVID-19 government dataset. `https://data.gov.il/dataset/covid-19`. [Accessed 2022].

[3] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020.

[4] Steffen Bickel and Tobias Scheffer. Dirichlet-enhanced spam filtering based on biased samples. *Advances in neural information processing systems*, 19, 2006.

[5] Alfred A Buck and John J Gart. Comparison of a screening test and a reference test in epidemiologic studies: I. indices of agreement and their relation to prevalence. *American Journal of Epidemiology*, 83(3):586–592, 1966.

[6] Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. Mandoline: Model evaluation under distribution shift. In *International Conference on Machine Learning*, pages 1617–1629. PMLR, 2021.

[8] Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. *arXiv preprint arXiv:2007.03511*, 2020.

[9] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International Conference on Machine Learning*, pages 2579–2589. PMLR, 2021.

[10] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15069–15078, 2021.

[11] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[12] Pinar Donmez, Guy Lebanon, and Krishnakumar Balasubramanian. Unsupervised supervised learning i: Estimating classification and regression errors without labels. *Journal of Machine Learning Research*, 11(4), 2010.

[13] George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.

[14] Saurabh Garg and Sivaraman Balakrishnan. Leveraging unlabeled data to predict out-of-distribution performance. *ICLR*, 2022.

[15] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[16] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test part 3: Demographic effects, 2019-12-19 2019.

[17] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1134–1144, 2021.

[18] Hirotaka Hachiya, Masashi Sugiyama, and Naonori Ueda. Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. *Neurocomputing*, 80:93–101, 2012.

[19] Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014, 2014.

[20] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.*, 52(4):79:1–79:36, 2019.

[21] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[22] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[23] Charles F Manski and Steven R Lerman. The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, pages 1977–1988, 1977.

[24] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[25] Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR, 2020.

[26] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[28] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9419–9427, 2021.

[29] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020.

[30] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[31] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

[32] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

[33] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[34] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[35] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[36] Peter Welinder, Max Welling, and Pietro Perona. A lazy man's approach to benchmarking: Semisupervised classifier evaluation and recalibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3269, 2013.

[37] Ruihan Wu, Chuan Guo, Yi Su, and Kilian Q Weinberger. Online adaptation to label distribution shift. *Advances in Neural Information Processing Systems*, 34, 2021.

[38] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Advances in neural information processing systems*, 24, 2011.

[39] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

[40] Eric Zhao, Anqi Liu, Animashree Anandkumar, and Yisong Yue. Active learning under label shift. In *International Conference on Artificial Intelligence and Statistics*, pages 3412–3420. PMLR, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] . See Appendix A.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] . See Appendix A.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] .
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] .
   (b) Did you include complete proofs of all theoretical results? [Yes] .See Appendix B
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] . Please see Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] . Please see Appendix D.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] .
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] .
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] .
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] .

A Broader Impact and Limitation Discussion

Monitoring, estimating, and explaining performance of deployed ML models is a growing area with significant economic and social impact. In this paper, we propose SJS, a new data distribution shift model to consider when both labels and features shift after model deployment. We show how SJS generalizes existing data shift models, and further propose SEES, a generic framework that efficiently explains and estimates an ML model's performance under SJS. This may serve as a monitoring tool to help ML practitioners recognize performance changes, discover potential fairness issues and take appropriate business decisions (e.g., switching to other models or debugging the existing ones). One limitation in general is adaption to continuously changing data streams. An online algorithm for performance estimation and explanation under SJS is in need to address this challenge. We will also open source our prototype of SEES serving as a resource for broad community to use.

B Missing Proofs

We provide all missing proofs in this section.

B.1 Proof of Theorem 1

Proof. We prove the statement by contradiction. Suppose not. That is to say, there exists a distribution $p_a(\boldsymbol{x}, y)$, such that $p_a(\boldsymbol{x}) = p_t(\boldsymbol{x})$, and there exists some set $\mathcal{J} \subset [d], |\mathcal{J}| \leq m$, such that $p_a(\boldsymbol{x}_{\mathcal{J}^c}|\boldsymbol{x}_{\mathcal{J}}, y) = p_s(\boldsymbol{x}_{\mathcal{J}^c}|\boldsymbol{x}_{\mathcal{J}}, y)$ and $p_a(\boldsymbol{x}, y) \neq p_t(\boldsymbol{x}, y)$. Let $w_a(\boldsymbol{x}, y) \triangleq \frac{p_a(\boldsymbol{x},y)}{p_s(\boldsymbol{x},y)}$ denote the ratio between this alternative distribution p_a and the source distribution. Recall that $w^*(\boldsymbol{x}, y) = \frac{p_t(\boldsymbol{x},y)}{p_s(\boldsymbol{x},y)}$. To show the contradiction, we simply need to show $p_a(\boldsymbol{x}, y) = p_t(\boldsymbol{x}, y)$, which is equivalent to show $w_a(\boldsymbol{x}, y) = w^*(\boldsymbol{x}, y)$.

By the m-SJS assumption, $p_t(\boldsymbol{x}_I^c|\boldsymbol{x}_I, y) = p_s(\boldsymbol{x}_I^c|\boldsymbol{x}_I, y)$, we have

$$w^*(\boldsymbol{x}, y) = \frac{p_t(\boldsymbol{x}, y)}{p_s(\boldsymbol{x}, y)} = \frac{p_t(\boldsymbol{x}_I, y)p_t(\boldsymbol{x}_I^c|\boldsymbol{x}_I, y)}{p_s(\boldsymbol{x}_I, y)p_s(\boldsymbol{x}_I^c|\boldsymbol{x}_I, y)} = \frac{p_t(\boldsymbol{x}_I, y)}{p_s(\boldsymbol{x}_I, y)}$$

Similarly, by the assumption $p_a(\boldsymbol{x}_J^c|\boldsymbol{x}_J, y) = p_s(\boldsymbol{x}_J^c|\boldsymbol{x}_J, y)$, we have

$$w_a(\boldsymbol{x}, y) = \frac{p_a(\boldsymbol{x}, y)}{p_s(\boldsymbol{x}, y)} = \frac{p_a(\boldsymbol{x}_J, y)p_a(\boldsymbol{x}_J^c|\boldsymbol{x}_J, y)}{p_s(\boldsymbol{x}_J, y)p_s(\boldsymbol{x}_J^c|\boldsymbol{x}_I, y)} = \frac{p_a(\boldsymbol{x}_J, y)}{p_s(\boldsymbol{x}_J, y)}$$

Thus, we only need to show

$$\frac{p_t(\boldsymbol{x}_I, y)}{p_s(\boldsymbol{x}_I, y)} = \frac{p_a(\boldsymbol{x}_J, y)}{p_s(\boldsymbol{x}_J, y)}$$

Our approach is to show that the two ratios all satisfy a system of linear equations, which, however, should have only a unique solution. To see this, let us first note that, for any $\bar{\bar{\boldsymbol{x}}}$, we have

$$\begin{aligned}
p_t(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}) &= \sum_{y=1}^{d} p_t(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}, y = \bar{y}) \\
&= \sum_{y=1}^{d} p_t(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{\mathcal{J}} = \bar{\boldsymbol{x}}_{\mathcal{J}}|\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})p_t(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y}) \\
&= \sum_{y=1}^{d} p_s(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{\mathcal{J}} = \bar{\boldsymbol{x}}_{\mathcal{J}}|\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})p_t(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y}) \\
&= \sum_{y=1}^{d} p_s(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}, y = \bar{y})\frac{p_t(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})}{p_s(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})}
\end{aligned}$$

Here, the first equation is by the total probability rule, the second equation is by the conditional probability rule, the third equation is by the s-SJS assumption, and the last equation is by conditional

probability rule again. Similarly, we can obtain

$$
\begin{aligned}
p_a(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}) &= \sum_{y=1}^{d} p_a(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}, y = \bar{y}) \\
&= \sum_{y=1}^{d} p_a(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_I = \bar{\boldsymbol{x}}_I | \boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y}) p_t(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y}) \\
&= \sum_{y=1}^{d} p_s(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_I = \bar{\boldsymbol{x}}_I | \boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y}) p_t(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y}) \\
&= \sum_{y=1}^{d} p_s(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}, y = \bar{y}) \frac{p_t(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y})}{p_s(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y})}
\end{aligned}
$$

where the first equation is by the total probability rule, the second equation is by the conditional probability rule, the third equation is by the assumption that $p_a(\boldsymbol{x}_{J^c}|\boldsymbol{x}_J, y) = p_s(\boldsymbol{x}_J|\boldsymbol{x}_J, y)$, and the last equation is by conditional probability rule again. Note that we have assumed that $p_a(\boldsymbol{x}) = p_t(\boldsymbol{x})$. Thus, we must have

$$
\begin{aligned}
&\sum_{y=1}^{d} p_s(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}, y = \bar{y}) \frac{p_t(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})}{p_s(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})} \\
=&\sum_{y=1}^{d} p_s(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}, y = \bar{y}) \frac{p_a(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y})}{p_s(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y})}
\end{aligned}
$$

which is simply

$$
\sum_{y=1}^{d} p_s(\boldsymbol{x}_{I^c\cap\mathcal{J}^c}, \boldsymbol{x}_{I\cup\mathcal{J}} = \bar{\boldsymbol{x}}_{I\cup\mathcal{J}}, y = \bar{y}) \left(\frac{p_t(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})}{p_s(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})} - \frac{p_a(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y})}{p_s(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y})} \right) = 0
$$

By the assumption that $\{p_s(\boldsymbol{x}_{\mathcal{J}^c\cap\mathcal{K}^c}, \boldsymbol{x}_{\mathcal{J}\cup I} = \bar{\boldsymbol{x}}_{\mathcal{J}\cup I}, y = \bar{y})\}_{y=1}^{d}$ are linearly independent, the above system of equations implies $\frac{p_t(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})}{p_s(\boldsymbol{x}_I = \bar{\boldsymbol{x}}_I, y = \bar{y})} - \frac{p_a(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y})}{p_s(\boldsymbol{x}_J = \bar{\boldsymbol{x}}_J, y = \bar{y})} = 0$. Note that this holds for any $\bar{\boldsymbol{x}}$. Thus, it is simply

$$
\frac{p_t(\boldsymbol{x}_I, y)}{p_s(\boldsymbol{x}_I, y)} = \frac{p_a(\boldsymbol{x}_J, y)}{p_s(\boldsymbol{x}_J, y)}
$$

That is to say, $w^*(\boldsymbol{x}, y) = w_a(\boldsymbol{x}, y)$ and thus $p_a(\boldsymbol{x}, y) = p_t(\boldsymbol{x}, y)$, which is a contradiction. Thus, the assumption is incorrect and (p_s, p_t) is identifiable, which finishes the proof. □

B.2 Proof of Theorem 2

Proof. Proving the first half statement is straightforward: suppose (p_s, p_t) is under label shift. Then by definition, $p_t(\boldsymbol{x}|y) = p_s(\boldsymbol{x}|y)$. That is basically $p_t(\boldsymbol{x}_{[d]}|y, \boldsymbol{x}_\varnothing) = p_s(\boldsymbol{x}_{[d]}|y, \boldsymbol{x}_\varnothing)$, which corresponds to 0-SJS with $I = \varnothing$.

Next we show the proof for the second half. Suppose (p_s, p_t) is under sparse covariate shift, i.e., $p_t(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I) = p_s(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I)$ for some I with size $m < d$, and $p_t(y|\boldsymbol{x}) = p_s(y|\boldsymbol{x})$. Adopting the definition of conditional probability, $p_t(y|\boldsymbol{x}) = p_s(y|\boldsymbol{x})$ can be rewritten as

$$
\frac{p_t(y, \boldsymbol{x})}{p_t(\boldsymbol{x})} = \frac{p_s(y, \boldsymbol{x})}{p_s(\boldsymbol{x})}
$$

By definition of conditional probability, $p_s(\boldsymbol{x}) = p_s(\boldsymbol{x}_I) \cdot p_s(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I)$ and $p_t(\boldsymbol{x}) = p_t(\boldsymbol{x}_I) \cdot p_t(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I)$, and thus we have

$$
\frac{p_t(y, \boldsymbol{x})}{p_t(\boldsymbol{x}_I) \cdot p_t(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I)} = \frac{p_s(y, \boldsymbol{x})}{p_s(\boldsymbol{x}_I) \cdot p_s(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I)}
$$

By the assumption $p_t(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I) = p_s(\boldsymbol{x}_{I^c}|\boldsymbol{x}_I)$, we can simplify this as

$$\frac{p_t(y, \boldsymbol{x})}{p_t(\boldsymbol{x}_I)} = \frac{p_s(y, \boldsymbol{x})}{p_s(\boldsymbol{x}_I)}$$

By definition of conditional probability, this is basically

$$p_t(y, \boldsymbol{x}_{I^c}|\boldsymbol{x}_I) = p_s(y, \boldsymbol{x}_{I^c}|\boldsymbol{x}_I)$$

which means (p_s, p_t) is under m-SJS.

Now we show the last piece of the statement by construction. Consider the case of $d = 2$. $\boldsymbol{x}_1, \boldsymbol{x}_2, y$ are all binary variables. $\mathbb{P}_s$ is generated as follows: y is first generated from Bernoulli distribution $Bern(0.5)$. If $y = 0$, $\boldsymbol{x}_1, \boldsymbol{x}_2$ are independently generated from $Bern(0.7)$ and $Bern(0.6)$. If $y = 1$, $\boldsymbol{x}_1, \boldsymbol{x}_2$ are independently generated from $Bern(0.1)$ and $Bern(0.2)$. $\mathbb{P}_t$ is generated as follows: y is first generated from $Bern(0.6)$. If $y = 0$, $\boldsymbol{x}_1, \boldsymbol{x}_2$ are independently generated from $Bern(0.5)$ and $Bern(0.6)$. If $y = 1$, $\boldsymbol{x}_1, \boldsymbol{x}_2$ are independently generated from $Bern(0.5)$ and $Bern(0.2)$.

It is easy to see that (p_s, p_t) is under 1-SJS with associated shift index set $I = \{1\}$, since $\boldsymbol{x}_2$ is independent of $\boldsymbol{x}_1$ and only depends on y. However,

$$p_t(\boldsymbol{x}_1 = 1|y = 1) = 0.5 \neq 0.1 = p_s(\boldsymbol{x}_1 = 1|y = 1)$$

and thus (p_s, p_t) is not under label shift. In addition,

$$p_t(y = 1|\boldsymbol{x}_1 = 1, \boldsymbol{x}_2 = 1) = \frac{0.6 \times 0.5 \times 0.2}{0.6 \times 0.5 \times 0.2 + 0.4 \times 0.5 \times 0.6} = \frac{1}{3}$$

$$p_s(y = 1|\boldsymbol{x}_1 = 1, \boldsymbol{x}_2 = 1) = \frac{0.5 \times 0.1 \times 0.2}{0.5 \times 0.1 \times 0.2 + 0.5 \times 0.7 \times 0.6} = \frac{1}{22} \neq p_t(y = 1|\boldsymbol{x}_1 = 1, \boldsymbol{x}_2 = 1)$$

and thus (p_s, p_t) is not under covaraite shift, which completes the proof. □

B.3 Proof of Theorem 3

Proof. We prove the statement via three main steps. First, we show that given full access to the distribution, the optimization over the marginal mass functions are sufficient to obtain the correct shifted features and weights. Next, we demonstrate that a large enough number of samples ensures the identified index set stays the same as when full access to the distribution is given with high probability. Finally, we can prove that with high probability, the learned weight function with large number of samples is close to the learned weight function when full distribution is known.

To proceed, let us introduce a few more notations for convenience.

- The distance used by SEES-d: $dd(J, w_J) \triangleq \sum_{\kappa: J \subseteq \kappa, |\kappa| = 2m} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa \in \mathcal{X}_\kappa} \|p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2$.
- The empirical distance used by SEES-d with finite samples: $\hat{dd}(J, w_J) \triangleq \sum_{\kappa: J \subseteq \kappa, |\kappa| = 2m} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa \in \mathcal{X}_\kappa} \|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot \hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2$
- The optimal weight function when the shifted feature set is fixed to J: $w_J^* \triangleq \arg\min_{w_J(\boldsymbol{x}, y)} \sum_{\kappa: J \subseteq \kappa, |\kappa| = 2m} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa \in \mathcal{X}_\kappa} \|p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2$
- The optimal weight function when the shifted feature set is fixed to J and full distribution is available: $w_J^* \triangleq \arg\min_{w_J(\boldsymbol{x}, y)} \sum_{\kappa: J \subseteq \kappa, |\kappa| = 2m} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa \in \mathcal{X}_\kappa} \|p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2$
- The optimal weight function when the shifted feature set is fixed to J and only finite samples are available: $\hat{w}_J^* \triangleq \arg\min_{w_J(\boldsymbol{x}, y)} \sum_{\kappa: J \subseteq \kappa, |\kappa| = 2m} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa \in \mathcal{X}_\kappa} \|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot \hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2$

Next we offer a few useful lemmas before giving the full proof.

Lemma 4. *Let Ω be a compact set, and $\boldsymbol{z}_1^*, \boldsymbol{z}_2^*$ be the optimal solution to the problems*

$$\min_{\boldsymbol{z}\in\Omega} f_1(\boldsymbol{z})$$

and

$$\min_{\boldsymbol{z}\in\Omega} f_2(\boldsymbol{z})$$

where $f_1(\cdot), f_2(\cdot)$ are two functions defined on Ω such that $|f_1(\boldsymbol{z}) - f_2(\boldsymbol{z})| \leq \Delta, \forall \boldsymbol{z} \in \Omega$. Then we have $|f_1(\boldsymbol{z}_1^*) - f_2(\boldsymbol{z}_2^*)| \leq \Delta$. If f_1 is strongly convex with parameter λ and Ω is the full real vector space, then $\|\boldsymbol{z}_1^* - \boldsymbol{z}_2^*\|_2^2 \leq \frac{4\Delta}{\lambda}$.

Proof. We start with the first half statement. By $|f_1(\boldsymbol{z}) - f_2(\boldsymbol{z})| \leq \Delta, \forall \boldsymbol{z} \in \Omega$, we have $f_1(\boldsymbol{z}_1^*) \leq \Delta + f_2(\boldsymbol{z}_1^*)$. Subtracting $f_2(\boldsymbol{z}_2^*)$ from both sides gives

$$f_1(\boldsymbol{z}_1^*) - f_2(\boldsymbol{z}_2^*) \leq \Delta + f_2(\boldsymbol{z}_1^*) - f_2(\boldsymbol{z}_2^*)$$

Observe that $\boldsymbol{z}_2^*$ is the optimal solution to minimizing $f_2(\cdot)$ on the set Ω. Thus, $f_2(\boldsymbol{z}_1^*) \leq f_2(\boldsymbol{z}_2^*)$ must hold. Thus, the above inequality becomes

$$f_1(\boldsymbol{z}_1^*) - f_2(\boldsymbol{z}_2^*) \leq \Delta$$

By symmetry of the two functions, we can obtain

$$f_2(\boldsymbol{z}_2^*) - f_1(\boldsymbol{z}_1^*) \leq \Delta$$

Combining the two inequalities gives

$$|f_2(\boldsymbol{z}_2^*) - f_1(\boldsymbol{z}_1^*)| \leq \Delta$$

Next let us turn to the second half. We first notice that $|f_1(\boldsymbol{z}_1^*) - f_2(\boldsymbol{z}_2^*)| \leq 2\Delta$. To see this, we can decompose the difference $f_1(\boldsymbol{z}_2^*) - f_2(\boldsymbol{z}_1^*)$ as

$$f_1(\boldsymbol{z}_2^*) - f_1(\boldsymbol{z}_1^*) = f_1(\boldsymbol{z}_2^*) - f_2(\boldsymbol{z}_2^*) + f_2(\boldsymbol{z}_2^*) - f_2(\boldsymbol{z}_1^*) + f_2(\boldsymbol{z}_1^*) - f_1(\boldsymbol{z}_1^*)$$

Here, $f_1(\boldsymbol{z}_2^*) - f_2(\boldsymbol{z}_2^*) \leq \Delta$ and $f_2(\boldsymbol{z}_1^*) - f_1(\boldsymbol{z}_1^*) \leq \Delta$ by the assumption $|f_1(\boldsymbol{z}) - f_2(\boldsymbol{z})| \leq \Delta, \forall \boldsymbol{z} \in \Omega$. $\boldsymbol{z}_2^*$ is the optimal solution to minimizing $f_2(\cdot)$, and thus $f_2(\boldsymbol{z}_2^*) - f_2(\boldsymbol{z}_1^*) \leq 0$. Therefore, combining all those leads to

$$f_1(\boldsymbol{z}_2^*) - f_1(\boldsymbol{z}_1^*) = f_1(\boldsymbol{z}_2^*) - f_2(\boldsymbol{z}_2^*) + f_2(\boldsymbol{z}_2^*) - f_2(\boldsymbol{z}_1^*) + f_2(\boldsymbol{z}_1^*) - f_1(\boldsymbol{z}_1^*) \leq 2\Delta$$

Meanwhile, $f_1(\boldsymbol{z}_2^*) - f_1(\boldsymbol{z}_1^*) \geq 0$ as $\boldsymbol{z}_1^*$ is the optimal solution to minimizing $f_1(\cdot)$. That is to say,

$$|f_1(\boldsymbol{z}_2^*) - f_1(\boldsymbol{z}_1^*)| \leq 2\Delta$$

Note that $f_1(\cdot)$ is a strongly convex function with parameter λ. Thus, we have

$$f_1(\boldsymbol{z}') \geq f_1(\boldsymbol{z}) + < \frac{\partial f_1(\boldsymbol{z})}{\partial \boldsymbol{z}}, \boldsymbol{z}' - \boldsymbol{z} > + \frac{\lambda}{2}\|\boldsymbol{z}' - \boldsymbol{z}\|_2^2$$

for any $\boldsymbol{z}', \boldsymbol{z}$. Now let us set $\boldsymbol{z} = \boldsymbol{z}_1^*, \boldsymbol{z}' = \boldsymbol{z}_2^*$. Remember that $\boldsymbol{z}_1^*$ is the optimal solution to minimizing $f_1(\cdot)$ and Ω is the full space. Thus, $\frac{\partial f_1(\boldsymbol{z}_1^*)}{\partial \boldsymbol{z}} = 0$. As a result, the above inequality becomes

$$f_1(\boldsymbol{z}_2^*) - f_1(\boldsymbol{z}_1^*) \geq \frac{\lambda}{2}\|\boldsymbol{z}_2^* - \boldsymbol{z}_1^*\|_2^2$$

By $|f_1(\boldsymbol{z}_2^*) - f_1(\boldsymbol{z}_1^*)| \leq 2\Delta$, we obtain

$$2\Delta \geq \frac{\lambda}{2}\|\boldsymbol{z}_2^* - \boldsymbol{z}_1^*\|_2^2$$

Rearranging the terms gives

$$\|\boldsymbol{z}_2^* - \boldsymbol{z}_1^*\|_2^2 \leq \frac{4\Delta}{\lambda}$$

which completes the proof. □

Lemma 5. *Suppose the source and target are under exact s-SJS, and for any set $\mathcal{J} \subset [d], |\mathcal{J}| \leq s$ and any $\bar{\boldsymbol{x}} \in \mathcal{X}$, the marginal probability mass functions $\{p_s(f(\boldsymbol{x}), \boldsymbol{x}_{\mathcal{J}\cup I} = \bar{\boldsymbol{x}}_{\mathcal{J}\cup I}, y = \bar{y})\}_{y=1}^d$ are linearly independent. Then $dd(J, w_J) = 0$ if and only if $J = I$ and $w_J(\boldsymbol{x}_J, y) = w^*(\boldsymbol{x}, y)$.*

Proof. Let us first relate the marginal density functions on the target domain to those on the source domain. Recall that

$$w^*(\boldsymbol{x}, y) = \frac{p_t(\boldsymbol{x}, y)}{p_s(\boldsymbol{x}, y)}$$

denote the true weights between the target and source density functions. Since the shift is only due to $\boldsymbol{x}_I$, $w^*(\boldsymbol{x}, y)$ only depends on $\boldsymbol{x}_I$. Abusing the notation a little bit, we use $w^*(\boldsymbol{x}_I, y)$ to denote $w^*(\boldsymbol{x}, y)$. Now we can write $p_t(\boldsymbol{x}, y) = w^*(\boldsymbol{x}_I, y) \cdot p_s(\boldsymbol{x}, y)$. Thus, the marginal distribution of $(\boldsymbol{x}_\kappa, f(\boldsymbol{x}))$ on the target domain can be written as

$$\begin{aligned} p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) =& \sum_{\bar{y}=1}^{L} \sum_{\boldsymbol{z}:\boldsymbol{z}_\kappa=\bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{z})=\bar{f}} p_t(\boldsymbol{x} = \boldsymbol{z}, y = \bar{y}) \\ =& \sum_{\bar{y}=1}^{L} \sum_{\boldsymbol{z}:\boldsymbol{z}_\kappa=\bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{z})=\bar{f}} w^*(\boldsymbol{z}_I, \bar{y}) \cdot p_s(\boldsymbol{x} = \boldsymbol{z}, y = \bar{y}) \end{aligned}$$

Now let us consider the two directions of the statement separately.

- $J = I$ and $w_J(\boldsymbol{x}_J, y) = w^*(\boldsymbol{x}, y) \implies dd(J, w_J) = 0$: In this case, $I = J \subseteq \kappa$ and thus $\boldsymbol{z}_I$ is forced to be $\bar{\boldsymbol{x}}_\kappa$. Hence, the marginal distribution becomes

$$\begin{aligned} p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) =& \sum_{\bar{y}=1}^{L} \sum_{\boldsymbol{z}:\boldsymbol{z}_\kappa=\bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{z})=\bar{f}} w^*(\boldsymbol{z}_I, \bar{y}) \cdot p_s(\boldsymbol{x} = \boldsymbol{z}, y = \bar{y}) \\ =& \sum_{\bar{y}=1}^{L} w^*(\bar{\boldsymbol{x}}_I, \bar{y}) \sum_{\boldsymbol{z}:\boldsymbol{z}_\kappa=\bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{z})=\bar{f}} p_s(\boldsymbol{x} = \boldsymbol{z}, y = \bar{y}) \\ =& \sum_{\bar{y}=1}^{L} w^*(\bar{\boldsymbol{x}}_I, \bar{y}) p_s(\boldsymbol{x}_\kappa = \bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{x}) = \bar{f}, y = \bar{y}) \\ =& \sum_{\bar{y}=1}^{L} w^*(\bar{\boldsymbol{x}}_J, \bar{y}) p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) \end{aligned}$$

  where the second equation is because $\boldsymbol{z}_I$ is fixed and does not depend on the inner summation, the third is by applying definition of conditional probability, and the last is simply change of notations. The above equation is simply

$$p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w^*(\bar{\boldsymbol{x}}_J, \bar{y}) p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) = 0$$

  which holds for every κ. Since $w_J(\boldsymbol{x}_J, y) = w^*(\boldsymbol{x}_J, y)$, it is equivalent to

$$p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) = 0$$

  Thus, summing over the square of it also leads to 0, i.e.,

$$dd(J, w_J) \triangleq \sum_{\kappa: J \subseteq \kappa, |\kappa|=2s} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa \in \mathcal{X}_\kappa} \|p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2 = 0.$$

- $dd(J, w_J) = 0 \implies I = J, w_J(\boldsymbol{x}_J, y) = w^*(\boldsymbol{x}, y)$: $dd(J, w_J) = 0$ implies that

$$p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) = 0$$

holds for each κ. In particular, consider κ that contains both I and J. Then $\boldsymbol{z}_\kappa = \bar{\boldsymbol{x}}_\kappa$ implies $\boldsymbol{z}_I = \bar{\boldsymbol{x}}_I$. Hence, the marginal distribution can be written as

$$\begin{aligned}
p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) =& \sum_{\bar{y}=1}^{L} \sum_{\boldsymbol{z}:\boldsymbol{z}_\kappa=\bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{z})=\bar{f}} w^*(\boldsymbol{z}_I, \bar{y}) \cdot p_s(\boldsymbol{x}=\boldsymbol{z}, y=\bar{y}) \\
=& \sum_{\bar{y}=1}^{L} w^*(\bar{\boldsymbol{x}}_I, \bar{y}) \sum_{\boldsymbol{z}:\boldsymbol{z}_\kappa=\bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{z})=\bar{f}} p_s(\boldsymbol{x}=\boldsymbol{z}, y=\bar{y}) \\
=& \sum_{\bar{y}=1}^{L} w^*(\bar{\boldsymbol{x}}_I, \bar{y}) p_s(\boldsymbol{x}_\kappa = \bar{\boldsymbol{x}}_\kappa, f(\boldsymbol{x}) = \bar{f}, y=\bar{y}) \\
=& \sum_{\bar{y}=1}^{L} w^*(\bar{\boldsymbol{x}}_I, \bar{y}) p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})
\end{aligned}$$

where the second equation is because $\boldsymbol{z}_I$ is fixed and does not depend on the inner summation, the third is by applying definition of conditional probability, and the last is simply change of notations. Comparing this with the above equation, we end up with

$$\sum_{\bar{y}=1}^{L} w^*(\bar{\boldsymbol{x}}_I, \bar{y}) p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) = 0$$

Or alternatively,

$$\sum_{\bar{y}=1}^{L} [w^*(\bar{\boldsymbol{x}}_I, \bar{y}) - w_J(\bar{\boldsymbol{x}}_J, \bar{y})] p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) = 0$$

which holds for any $\bar{f}, \bar{\boldsymbol{x}}_\kappa$. By the linear independence assumption, this holds if and only if all the coefficients are 0, i.e., $w^*(\bar{\boldsymbol{x}}_I, \bar{y}) - w_J(\bar{\boldsymbol{x}}_J, \bar{y}) = 0$ for all $\bar{\boldsymbol{x}}$ and $\bar{y}$. That is to say, the two functions are identical: $w^*(\boldsymbol{x}_I, y) = w_J(\boldsymbol{x}_J, y)$. As $w_J(\boldsymbol{x}_J, y)$ and $W_*(\boldsymbol{x}_J, y)$ are identical, they can only depend on variables in the set $I \cap J$. That is to say, there exists another importance weights $w_{I\cap J}(\boldsymbol{x}_{I\cap J}, y)$ which results in the same target distribution produced by $W_I(\bar{\boldsymbol{x}}_I, y)$. By the assumption, the shift among the two distribution is exactly s-SJS. Hence, we have $|I \cap J| = s$. $|I| = s$ and $I \cap J \subseteq I$ implies $I \cap J = I$ and thus $J \supseteq I$. $|J| = s$ further implies $J = I$.

Therefore, $dd(J, w_J) = 0 \Leftrightarrow I = J, w_J(\boldsymbol{x}_J, y) = w^*(\boldsymbol{x}, y)$, which completes the proof. □

Lemma 6. *With probability at least $1-\delta$, for any possible $J \subseteq [d], |J| = s$, we have*

$$\begin{aligned}
&|dd(J, w_J^*) - \hat{dd}(J, \hat{w}_J^*)| \\
\leq& 3 \cdot (2s\bar{v})^s M L^2 \sqrt{2s \log d + s \log \bar{v} + 2 \log L + \log 1/\delta} \left(\sqrt{\frac{1}{2n_s}} + LM \sqrt{\frac{1}{2n_t}} \right).
\end{aligned}$$

and

$$\|w_J^* - \hat{w}_J^*\|_2^2 \leq O\left(ML^2 \left(\sqrt{\frac{\log 1/\delta}{2n_s}} + LM \sqrt{\frac{\log 1/\delta}{2n_t}} \right) \right)$$

Proof. Let us start by considering a fixed J. By definition of w_J^* and $\hat{w}_J^*$, we have

$$dd(J, w_J^*) = \min_{w_J(\boldsymbol{x}, y)} \sum_{\kappa: J \subseteq \kappa, |\kappa|=2s} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa \in \mathcal{X}_\kappa} \|p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2$$

and

$$\hat{dd}(J, \hat{w}_J^*) = \min_{w_J(\boldsymbol{x}, y)} \sum_{\kappa: J \subseteq \kappa, |\kappa|=2s} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa \in \mathcal{X}_\kappa} \|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y}) \cdot \hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\|_2^2$$

Let us first show that the above two objective functions are close for any fixed w_J. To see this, we first apply difference of two squares to obtain

$$\begin{aligned}
&\left(\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)^2 - \left(p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)^2 \\
=&\left(\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) + p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, y)\left(\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) + p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)\right) \\
&\cdot\left(\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, y)\left(\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) - p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)\right)
\end{aligned} \tag{B.1}$$

Note that all estimated probability mass must be bounded by 1, and by assumption, $|w_J| \leq M$. Thus,

$$\left|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) + p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, y)\left(\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) + p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)\right| \leq 2 + 2LM \leq 3LM \tag{B.2}$$

and

$$\begin{aligned}
&\left|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})\left(\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) - p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)\right| \\
\leq&\left|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f})\right| + \left|\sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})\left(\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) - p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)\right| \\
\leq&\left|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f})\right| + \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})\left|\left(\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) - p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)\right| \\
\leq&\left|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f})\right| + M\sum_{\bar{y}=1}^{L}\left|\left(\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) - p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)\right|
\end{aligned}$$

Observe that, $\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f})$ is the standard empirical estimation of $p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f})$. Therefore, applying Hoeffding's inequality, we have with probability $1 - \delta$,

$$|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f})| \leq \sqrt{\frac{\log 1/\delta}{2n_s}}$$

Similarly, with probability $1 - \delta$,

$$|\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) - p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})| \leq \sqrt{\frac{\log 1/\delta}{2n_t}}$$

Thus, with probability $1 - (\bar{v}^s L + \bar{v}^s L^2)\delta$, the above holds for any $\bar{f}, \bar{y}, \bar{\boldsymbol{x}}_\kappa$. Thus we have

$$|\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})(\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}) - p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y}))| \leq \sqrt{\frac{\log 1/\delta}{2n_s}} + LM\sqrt{\frac{\log 1/\delta}{2n_t}}$$

Combing this with inequalities B.1 and B.2, we have

$$\begin{aligned}
&\left(\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)^2 - \left(p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)^2 \\
\leq&3LM\left(\sqrt{\frac{\log 1/\delta}{2n_s}} + LM\sqrt{\frac{\log 1/\delta}{2n_t}}\right)
\end{aligned}$$

Summing over $\kappa, \bar{\boldsymbol{x}}_\kappa, \bar{f}$, we have

$$\sum_{\kappa: J\subseteq\kappa\in[d], |K|=2s} \sum_{\bar{f}=1}^{L} \sum_{\bar{\boldsymbol{x}}_\kappa\in\mathcal{X}_\kappa} \left(\hat{p}_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})\hat{p}_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)^2$$
$$- \left(p_t(\bar{\boldsymbol{x}}_\kappa, \bar{f}) - \sum_{\bar{y}=1}^{L} w_J(\bar{\boldsymbol{x}}_J, \bar{y})p_s(\bar{\boldsymbol{x}}_\kappa, \bar{f}, \bar{y})\right)^2$$
$$\leq (2s)^s \bar{v}^s L \cdot 3LM \left(\sqrt{\frac{\log 1/\delta}{2n_s}} + LM\sqrt{\frac{\log 1/\delta}{2n_t}}\right)$$

That is to say, for all w_J, $|dd(J, w_J) - \hat{dd}(J, w_J)| \leq 3(2s\bar{v})^s ML^2 \left(\sqrt{\frac{\log 1/\delta}{2n_s}} + LM\sqrt{\frac{\log 1/\delta}{2n_t}}\right)$ with probability $1 - (\bar{v}^s L + \bar{v}^s L^2)\delta$. In addition, note that $dd(J, w_J)$ is a quadratic function over w_J. By the linear independence assumption, $dd(J, w_J)$ must be strongly convex. Now applying Lemma 4, we have

$$|dd(J, w_J^*) - \hat{dd}(J, \hat{w}_J^*)| \leq 3(2s\bar{v})^s ML^2 \left(\sqrt{\frac{\log 1/\delta}{2n_s}} + LM\sqrt{\frac{\log 1/\delta}{2n_t}}\right)$$

and

$$\|w_J^* - \hat{w}_J^*\|_2^2 \leq \frac{12}{\lambda}(2s\bar{v})^s ML^2 \left(\sqrt{\frac{\log 1/\delta}{2n_s}} + LM\sqrt{\frac{\log 1/\delta}{2n_t}}\right)$$

where λ is the parameter corresponding to the strongly convexity of $dd(J, \cdot)$. This holds for a fixed J with probability $1 - (\bar{v}^s L + \bar{v}^s L^2)\delta$. There are $\binom{d}{s}$ many possible choices of J. Thus, with probability at least $1 - d^s(\bar{v}^s L + \bar{v}^s L^2)\delta \geq 1 - 2d^s\bar{v}^s L^2\delta$, the above holds. Replacing $2d^s\bar{v}^s L^2\delta$ by δ gives the desired form. □

Finally we are ready to prove the statement. By Lemma 5, $dd(I, w^*) = 0$ and for any $J \neq I$, we have $dd(J, w_J) > dd(I, w^*)$. Let $c_1 = \min_{J\neq I} d(J, w_J) > 0$, $c_2 = \frac{c_1}{6(2m\bar{v})^m ML^2}$, and the constant $c = c_2/c_1$. Now we make progresses in two steps.

- First let us show with high probability, the estimated shifted features match the true shifted features. This is equivalent to show, with high probability, $\hat{dd}(I, \hat{w}_I^*) < \hat{dd}(J, \hat{w}_J^*)$ for any $J \neq I$. To do so, let us note that, for any $J \neq I$,

$$\begin{aligned}
&\hat{dd}(I, \hat{w}_I^*) - \hat{dd}(J, \hat{w}_J^*) \\
=&\hat{dd}(I, \hat{w}_I^*) - dd(I, w_I^*) - (\hat{dd}(J, \hat{w}_J^*) - dd(J, w_J^*)) + dd(I, w_I^*) - dd(J, w_J^*) \\
=&\hat{dd}(I, \hat{w}_I^*) - dd(I, w_I^*) - (\hat{dd}(J, \hat{w}*_J) - dd(J, w_J^*)) + c_1
\end{aligned}$$

  By Lemma 6, with probability $1 - \delta$, for any J,

$$\begin{aligned}
&|dd(J, w_J^*) - \hat{dd}(J, \hat{w}_J^*)| \\
\leq&3(2m\bar{v})^m ML^2\sqrt{2m\log d + m\log\bar{v} + 2\log L + \log 1/\delta}\left(\sqrt{\frac{1}{2n_s}} + LM\sqrt{\frac{1}{2n_t}}\right) \\
=&\frac{1}{3}c_2\sqrt{2m\log d + m\log\bar{v} + 2\log L + \log 1/\delta}\left(\sqrt{\frac{1}{2n_s}} + LM\sqrt{\frac{1}{2n_t}}\right)
\end{aligned}$$

  Therefore, we have

$$\begin{aligned}
&\hat{dd}(I, \hat{w}_I^*) - \hat{dd}(J, \hat{w}_J^*) \\
\leq& -\frac{2}{3}c_2\sqrt{2m\log d + m\log\bar{v} + 2\log L + \log 1/\delta}\left(\sqrt{\frac{1}{2n_s}} + LM\sqrt{\frac{1}{2n_t}}\right) + c_1
\end{aligned}$$

By the assumption, $\sqrt{2m\log d + m\log\bar{v} + 2\log L + \log 1/\delta}\left(\sqrt{\frac{1}{2n_s}} + LM\sqrt{\frac{1}{2n_t}}\right) < c_1/c_2$. Hence, we have

$$\hat{dd}(I, \hat{w}_I^*) - \hat{dd}(J, \hat{w}_J^*) \leq -\frac{2}{3}c_2\frac{c_1}{c-2} + c_1 = \frac{1}{3}c_1 < 0$$

for any $J \neq I$. Thus, the correct shifted features are selected with probability at least $1-\delta$.

- Finally we show the learned $\hat{w}_{\hat{J}}^*$ is close to the true importance weights w^*. By Lemma 6,

$$\|w_J - \hat{w}_J\|_2^2 \leq O\left(ML^2\left(\sqrt{\frac{\log 1/\delta}{2n_s}} + LM\sqrt{\frac{\log 1/\delta}{2n_t}}\right)\right)$$

with high probability for all J. Thus it holds for the selected features $\hat{J}$. We have just shown that with high probability, the correct shifted features are selected, i.e., $\hat{J} = I$. Thus, it simply means

$$\|w^* - \hat{w}_{\hat{J}}^*\|_2^2 \leq O\left(ML^2\left(\sqrt{\frac{\log 1/\delta}{2n_s}} + LM\sqrt{\frac{\log 1/\delta}{2n_t}}\right)\right)$$

which completes the proof.

□

C Additional Discussions

Here we provide additional discussions.

Motivating examples when SJS occurs. In Section 1 we give one example when SJS occurs. Now we give two more examples to show how SJS broadly exists in different scenarios.

- Cancer diagnosis: Suppose we wish to build an ML model to diagnose cancer based on patient health records. The model is developed based on labeled dataset in some developed countries. However, when deploying it to hospitals in a developing country, there might be much more young patients, and the cancer rate for the elderly can also increase. Suppose the other features' distribution remains unchanged given age and cancer diagnosis. Then the distribution shift is naturally an SJS.
- Toxic text recognition: Consider a mobile app that detects and filters toxic texts based on the content and senders' information. Due to unexpected events (for example, disappointing football games), the toxic texts rate, as well as the total number of texts, may both significantly increase in some locations at different time periods. The shift of text locations and toxic text rate is thus another example of SJS.

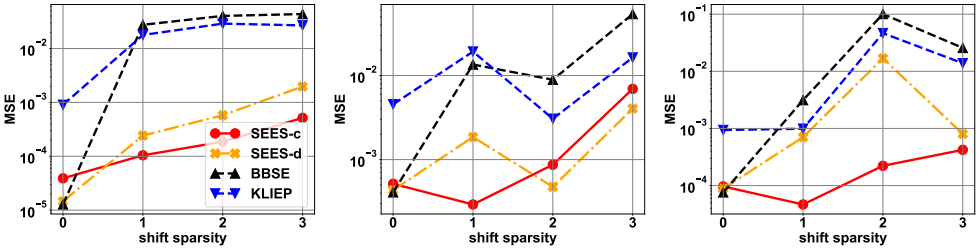
Understanding sparse covariate shifts. Sparse covariate shift is a special case of covariate shift [31]. It occurs when the shifts are caused by a few variables. For example, consider two census datasets collected in two periods. If a large population moved from one city to another between the two periods and everything else remains the same, then there is a sparse covariate shift (location alone). It is also related to Adversarial patches: if adversarial noises are added to a few features (or a small number of pixels in image domains), it also corresponds to the sparse covariate shift.

D Experimental Details

Here we provide additional experimental details.

Table 2: Dataset statistics.

Dataset	# of instances	# of features	Shift types
BANKCHURN	10000	10	Synthetic
COVID-19	660787	8	
CREDIT	29946	23	
EMPLOY	227871	16	Geography (CA,PR,IA,WI)
INCOME	245783	10	Geography (CO,CA,KS,OH)
INSURANCE	32140	19	Temporality (2014,2016,2018)



(a) BANKCHURN (b) COVID-19 (c) CREDIT

Figure 5: Effects of shift sparsity. For each dataset, we measure how the ℓ_2 loss of the estimated accuracy varies as number of shifted features increases given the same sample sizes. Overall, the estimation error of SEES slowly grows as shift sparsity increases, but is consistently lower than BBSE and KLIEP.

Datasets and ML tasks. We use six datasets for evaluation, namely, BANKCHURN [1], COVID-19 [2], and CREDIT [39] for various SJS simulations, and EMPLOY, INCOME, and INSURANCE [11] for performance evaluation under real world distribution shifts. BANKCHURN [1] contains 10 features such as gender, age, credit score and balance, and the goal is to predict whether a bank customer may churn. COVID-19 [2] is a subset of the publicly accessible COVID-19 dataset from the Israel government website, containing both demographic and symptom features. Here, we select the subset that contains all tested cases in January, 2022, and aim at predicting whether a person tests positive or negative for COVID-19. CREDIT [39] includes age, gender, education, bill payments and several other features for 29946 individuals. The goal is to predict the default payment. EMPLOY, INCOME, and INSURANCE are subset of the public use microdata samples from the US census [11]. EMPLOY contain 16 features for individual samples from four different states, CA, PR, IA, and WI and our goal is to predict if a person is employed or not. In INCOME, 245,783 anonymous census record samples from four states including CA, CO, KS, and OH are collected. The goal is to predict if a person's income is lower or higher than \$50,000. INSURANCE contains 32140 individual samples from the state IA collected in year 2014, 2016, and 2018. The task is to predict whether an individual is covered by an insurance plan. The dataset statistics can be found in Table 2.

Experiment setups. All experiments were run on a machine with 20 Intel Xeon E5-2660 2.6 GHz cores, 160 GB RAM, and 80 GB disk with Ubuntu 18.04 LTS as the OS. Our prototype was implemented and tested in Python 3.8. To apply SEES-d, we discretized all continuous features. To apply SEES-c, we set the trade-off parameter $\eta = 0.001$. For continuous features, linear functions were used as the basis. For discrete features, indicate functions were adopted. For example, if $\bar{\boldsymbol{x}}_1 \in \mathbb{R}$ and $\bar{\boldsymbol{x}}_2 \in \{0, 1\}$, then the basis functions consist of three components, $\phi_1(\bar{\boldsymbol{x}}, y) = \bar{\boldsymbol{x}}_1, \phi_2(\bar{\boldsymbol{x}}, y) = \mathbb{1}_{\bar{\boldsymbol{x}}_2=0}$, and $\phi_3(\bar{\boldsymbol{x}}, y) = \mathbb{1}_{\bar{\boldsymbol{x}}_2=1}$. For KLIEP, the maximum number of iterations was set as 2,500.

Effects of shift sparsity. Figure 5 shows how the accuracy gap estimation performance vary as the number of shifted features increase. For each dataset, we start with shifting no features (label shift), to shifting 1, 2, and 3 features together with labels. Specifically, for BANKCHURN, shifts occur for (i) labels alone (0-SJS), (ii) then both labels and geography feature (1-SJS), (iii) then labels, geography, and gender (2-SJS), and (iv) finally labels, geography, gender, and credit card

owned before. For COVID-19, we start with shifting label alone (0-SJS), and then incrementally shift features age, gender, and contact risk to simulate 1-SJS, 2-SJS, and 3–SJS, respectively. For credits, 0-SJS, 1-SJS, 2-SJS, and 3-SJS correspond to shift in (i) labels, (ii) labels and marriage status, (iii) labels, marriage status, and gender, and (iv) labels, marriage status, gender, and credit balance. Overall, we observe that the estimation error of SEES-c and SEES-d slightly increases as the number of shifted features grows, but is consistently lower than that of BBSE and KLIEP.

Robustness to data randomness. In Section 5 we mainly focus on the average performance metric (ℓ_2 loss). Here we provide additional robustness measurement: on the COVID-19 dataset, we repeat the case study experiments 200 times with different random seeds to generate the source and target datasets, and report the variance of the estimated accuracy gap. As shown in Table 3, we observe that the variance for all of the methods is small: variance of SEES-c and KLIEP are are less than 0.00003, and the variance for SEES-d is 0.003 and for BBSE is 0.0003.

Table 3: Variance of the estimated accuracy gap. The values were calculated over 200 experimental runs. Overall, the variance of all methods is small.

Method	SEES-c	SEES-d	KLIEP	BBSE
Estimated Accuracy Gap Variance	0.000018	0.0033	0.00005	0.0003

Sensitivity of the sparsity parameter in SEES-d. SEES-d needs knowledge of the sparsity parameter s, and thus a natural question is how sensitive its performance is when the sparsity parameter does not exactly match the true sparsity. To study this, we first generate a source-target pair (each containing 10,000 data points) on the COVID-19 dataset where labels and 3 features (age, contact risk, and gender) shift, and then measure the ℓ_2 loss of the estimated performance gap by SEES-d with sparsity parameter ranging from 0 to 7 (the number of features). Here, sparsity parameter being 0, 1, 2, 4, 5, 6, and 7 corresponds to the mismatched case. Figure 6 summarizes the averaged ℓ_2 loss over 100 experimental runs. The randomness comes from the choices of shifted features and the samples from source and target datasets, and the shaded area indicates the standard deviation. Overall, we observe that SEES-d is robust to small parameter mismatch: there is little change of the estimation error when the sparsity parameter (2, 3, 4, 5) is close to the true number of shifted features (3). When the parameter mismatch is too large, a relatively larger change in the estimation error can be observed (though SEES-d still works better than BBSE even when the model mismatch is large). This is because a too small sparsity parameter restricts the search space, while a too large parameter often incurs an identifiability issue as our theory shows (i.e., different feature-label joint distributions correspond to the same observed target feature distribution). In practice, if the user has a prior belief that the distribution shift is not sparse (i.e. number of shifted features is > $d/2$), then SEES-d may not be appropriate.

Comparison with additional baselines. For more in-depth understanding, we compare the performance of SEES with an additional baseline DLU [25]. DLU basically adopts discriminative learning on the union of the source and target dataset, and then uses the classifier's prediction to reweigh the source data. We adopt it on the COVID-19 dataset for a case study and measure the performance of the weight and accuracy gap estimation for it along with SEES-c, SEES-d, BBSE, and KLIEP. As shown in Table 4, DLU performs better than KLIEP, but is still much worse than SEES-c and SEES-d. For example, the MSE of its estimated weights is 0.320, while that of SEES-d is only 0.002. Hence, SEES-c and SEES-d still lead to the best estimated accuracy gap.

Robustness across different shifts. Real world distribution shifts may vary, and thus it is important to understand how different methods behave when encountering different shifts. To understand this, we study the performance of SEES-c and SEES-d along with all baselines (BBSE, KLIEP, and DLU) on the COVID-19 dataset when different shifts occur. Specifically, we generate source-target pair where (i) label shift, (ii) (sparse) covariate shift (feature age), and (iii) 1-SJS (both label and feature age) occurs, separately. Table 5 gives the squared ℓ_2 loss of the accuracy gap estimation for all methods. Interestingly, we observe that SEES-c and SEES-d produce reliable accuracy gap estimation across different shifts, while all baselines are sensitive to shift types. For example, KLIEP and DLU achieves decent performance when there is only covariate shift, but their estimation is much worse when label shift occurs. Similarly, the estimation error of BBSE is small when labels

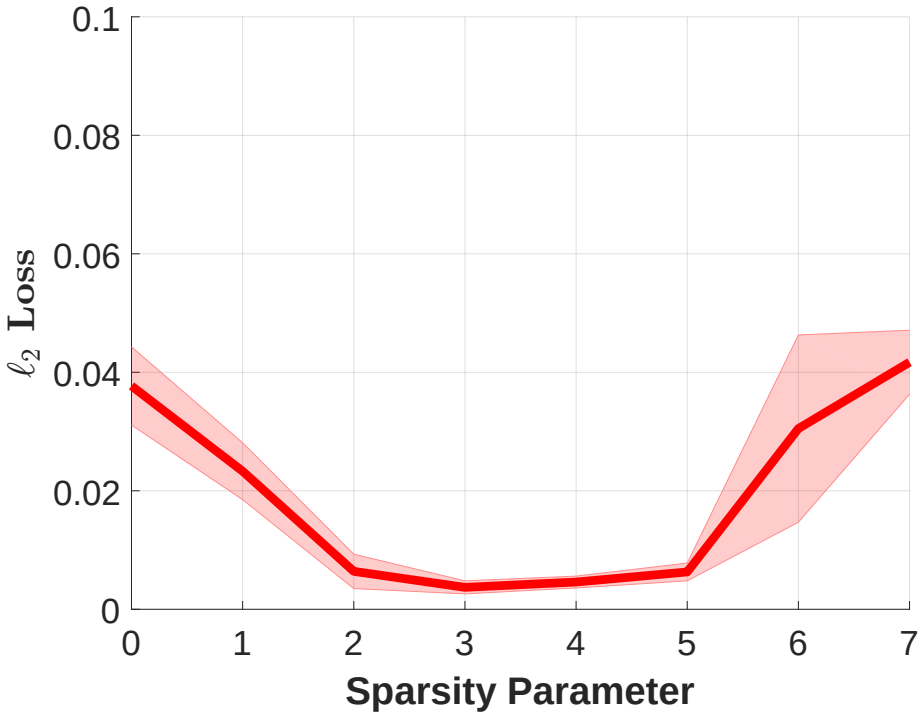


Figure 6: Sensitivity of the sparsity parameter in SEES-d on the COVID-19 dataset. On a source-target pair where labels and 3 features shift, we measure the ℓ_2 loss of the estimated model performance gap produced by SEES-d with sparsity parameter ranging from 0 to 7 (number of features). The performance was averaged over 100 experimental runs, where the randomness comes from choices of shifted features and samples in the source-target pairs and the shaded area indicates the standard deviation.. Overall, SEES-d is robust to small parameter mismatch: there is little change of the estimation error when the sparsity parameter (2,3,4,5) is close to the true number of shifted features (3). In addition, the estimation performance of SEES-d with small parameter mismatch is also consistently better than all baselines.

Table 4: Performance of the weight and accuracy gap estimation on the COVID-19 case study. Overall, SEES-c and SEES-d achieve the smallest mean square error (MSE) and largest Pearson correlation coefficient (PCC). Thus, their estimated accuracy gap is closest to the true gap.

Method	MSE	PCC	Est Gap	True Gap
SEES-c	0.003	0.996	18.4	15.5
SEES-d	***0.002***	***0.996***	***16.7***	
BBSE	0.144	0.857	26.3	
KLIEP	0.573	-0.112	1.58	
DLU	*0.320*	*0.328*	*0.38*	

indeed shift but is much worse than other methods when the shift is due to covariate. On the other hand, the performance of SEES-c and SEES-d is as good as that of the best baseline when label or covariate shifts. For SJS, the proposed methods achieve significantly better estimation than all baselines. Those suggest that SEES-c and SEES-d are more robust to different shifts and thus safer to be deployed in the wild.

Table 5: Squared ℓ_2 loss of the estimated accuracy gap on the COVID-19 dataset. SEES-c and SEES-d are the only approaches providing reliable estimation across all shifts.

Method	Label Shift	Covariate Shift	Joint Shift
SEE-c	0.0005	0.0010	0.0003
SEE-d	0.0004	0.0015	0.0019
BBSE	0.0004	0.0321	0.0135
KLIEP	0.0045	0.0008	0.0193
DLU	0.0029	0.0001	0.0251

More models on datasets with real world shifts. Next we provide the performance estimation for more models on datasets with real world shifts. We study the estimation performance for three models, namely, a gradient boosting, a neutral network, and a decision tree. The maximum depth of the gradient boosting and deicision tree is 50, and the neutral network consists of two layers with 100 hidden units. As shown in Table 6, SEES often provides significant error reduction over the compared baselines BBSE and KLIEP.

Table 6: Root mean square error of estimated accuracy gap (%) under real shifts for gradient boosting, a neural network, and a decision tree. The numbers are averaged over all source-target pairs in each dataset. For each dataset and ML model, SEES provides significant estimation error reduction over baselines.

Dataset	ML model	Accuracy estimation's ℓ_2 error (%)			
		SEES-c	SEES-d	BBSE	KLIEP
EMPLOY	Gradient boosting	**2.9**	3.0	5.2	5.2
	Neural network	**5.2**	6.0	6.2	5.6
	Decision tree	4.0	**3.8**	5.7	5.7
INCOME	Gradient boosting	**1.9**	2.4	3.0	3.4
	Neural network	**4.4**	6.4	9.8	7.6
	Decision tree	**2.4**	2.9	3.0	3.8
INSURANCE	Gradient boosting	**1.7**	2.2	2.1	5.0
	Neural network	**2.2**	4.7	10.8	7.3
	Decision tree	**2.0**	2.5	2.4	5.9