# ESTIMATING STRUCTURAL SHIFTS IN GRAPH DOMAIN ADAPTATION VIA PAIRWISE LIKELIHOOD MAXIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Graph domain adaptation (GDA) emerges as an important problem in graph machine learning when the distribution of the source graph data used for training is different from that of the target graph data used for testing. While much of the prior work on GDA has focused on the idea of aligning node representations across source and target domains, recent studies show that such approaches can be suboptimal in the presence of conditional structure shift (CSS), where the distribution of graph edges conditioned on labels changes across domains. In this work, we develop a unified framework to solve CSS and show that existing GDA methods for CSS arise as special cases of our framework. This framework further allows us to develop a new method, Pairwise-Likelihood maximization for graph Structure Alignment (PLSA), which uses rich information from pairwise nodes and edges to improve the estimation of target connection probabilities. We establish conditions under which our method is identifiable and introduce a simple edge reweighting scheme based on importance weights to align the source and target graphs. Theoretically, under the contextual stochastic block model (CSBM), we derive finite-sample guarantees using recent results in matrix concentration inequalities for U-statistics. We complement our theoretical results with empirical studies that demonstrate the effectiveness of our method.

## 1 INTRODUCTION

With the growing prevalence of graph-structured data across domains, graph neural networks (GNNs) have emerged as powerful tools for achieving remarkable performance in many graph machine learning tasks (Kipf & Welling, 2017; Zhang & Chen, 2018; Chami et al., 2022). Despite their empirical successes, a key challenge arises when the distribution of data available for training (source) is different from that encountered at test time (target) (Wu et al., 2024; You et al., 2023). Such distributional shifts may occur due to changes in node attributes, class proportions, or the graph structure that encodes dependencies between nodes. These discrepancies can result in significant degradation in model performance, limiting the reliability of GNNs in real-world deployments (Liu et al., 2024b;c; Zhu et al., 2021a). Graph domain adaptation (GDA) seeks to address this challenge by transferring knowledge from a source domain with sufficient supervision to a target domain with no labels (Cai et al., 2024; Liu & Ding, 2024; Ma, 2024).

Unlike the classical domain adaptation (DA) problem which typically involves (marginal or conditional) feature shift or label shift, GDA is more complicated because it must also account for the shift in the graph structure. Existing GDA methods are largely motivated by domain-invariant representation learning (Ganin et al., 2016; Hoffman et al., 2018) and generally aim to align the source and target distributions of node representations after aggregating neighborhood information in GNNs (Zhu et al., 2021a; Xiao et al., 2022; You et al., 2023; Liu et al., 2024a;b). However, it remains unclear under which assumptions these approaches succeed; in particular, since classical domain-invariant representation methods are known to fail in the presence of label-flipping features (Zhao et al., 2019; Johansson et al., 2019; Wu et al., 2025) or label shift (Wu et al., 2019; Chen & Bühlmann, 2021), it is natural to expect that GDA methods based on the similar principle may also fail in such scenarios.

Recent studies further show that when there is a conditional structure shift (CSS)—that is, when the conditional edge probabilities connecting nodes change across domains—aligning the marginal node representations across source and target becomes inefficient and yields suboptimal prediction performance in the target domain (Liu et al., 2023; 2024c). Motivated by this observation, several methods have been proposed, including reweighting the source graph (Liu et al., 2023) and using a pairwise moment-matching based estimator to correct CSS (Liu et al., 2024c). The latter approach, called Pairwise Alignment (Pair-Align), further integrates existing label shift correction methods (Lipton et al., 2018) to simultaneosuly solve the CSS and label shift problems for GDA.

In this work, we focus primarily on CSS, assuming that the joint distribution of features and labels are invariant across source and target domains. This assumption allows us to isolate CSS from other types of shifts and study CSS more directly. In this setting, we propose Pairwise-Likelihood maximization for graph Structure Alignment (PLSA), a new method for estimating and correcting CSS in node classification tasks. We show that existing methods for CSS arise as special cases of a broader framework to solve CSS, and that PLSA is another instantiation of this framework that exploits more information from the data to improve estimation accuracy. Theoretically, when data are generated from the contextual stochastic block model (CSBM), we give upper bounds on the estimation error of PLSA using recently developed matrix concentration inequalities for U-statistics, even when dependencies exist between node attributes and edges. Our main contributions are summarized as follows.

- We develop a unified framework for correcting CSS from a distribution matching point of view and show that existing methods can be viewed as special cases of this framework.
- We propose PLSA, a new instantiation of this framework based on pairwise likelihood maximization with a calibrated predictor, and provide conditions to ensure its identifiability.
- We introduce an edge reweighting scheme for the source graph to improve downstream node classification and establish finite-sample guarantees under CSBM. We further validate our method through empirical studies.

## 2 Related work

**Graph domain adaptation** Graph domain adaptation (GDA) extends classical DA to the setting where data are graph-structured. One popular way to formulate DA is to assume the existence of invariant representations across domains (Ben-David et al., 2010; Ganin et al., 2016), which has inspired many GDA methods that adapt this idea to graph setting. For instance, Zhu et al. (2022) use central moment discrepancy to align node representations of GNNs, and You et al. (2023) propose a spectral regularization framework that enforces invariance by controlling spectral smoothness and maximum frequency response of GNNs. Liu et al. (2024b) provide the GDABench benchmark and show that simple GNN-based baselines with vanilla DA often outperform more sophisticated GDA methods. For a comprehensive review of invariance-based GDA methods, we refer readers to recent surveys (Liu & Ding, 2024; Ma, 2024). Beyond invariance-based approaches, other directions in GDA include causal-based methods (Wu et al., 2022; Luo et al., 2025), generative modeling approaches (Cai et al., 2024), and methods that align homophilic signals (Fang et al., 2025).

**Conditional structure shift** Conditional structure shift (CSS) refers to the type of shift where the conditional distribution of edges given node labels is different across domains (Zhu et al., 2023; Liu et al., 2023). Unlike covariate or label shift, CSS is unique to the graph-structured data since the connectivity patterns between nodes, given labels, can vary even when the node features and label distributions are invariant. Recent studies show that ignoring CSS can make marginal alignment of node representations ineffective. To mitigate such issue, Liu et al. (2023) propose Structural Reweighting, which reweights source graph so that the neighborhood statistics of source nodes mimic those in the target domain. Building on this idea, Liu et al. (2024c) introduce Pairwise Alignment (Pair-Align), a method that simultaneously accounts for both CSS and label shift by formulating the estimation of edge and label shift weights as solutions to linear systems. Following this line of work, we provide a unified framework for CSS with a principled approach and finite-sample guarantees.

**Label shift** In recent years, label shift has been extensively studied in the anticausal setting (Lipton et al., 2018; Azizzadenesheli et al., 2019), where the label distribution changes while the conditional

distribution $x \mid y$ is invariant. Label shift can also arise in GDA, where existing correction methods have been used to address it (Liu et al., 2024c). In label shift, two dominant approaches are Black Box Shift Estimation (BBSE) (Lipton et al., 2018; Azizzadenesheli et al., 2019) and Maximum Likelihood Label Shift (MLLS) (Saerens et al., 2002; Garg et al., 2020). BBSE uses a black box classifier $h$ trained on the source data to estimate the confusion matrix and construct a linear system, whose solution provides an estimate of the importance weights. In contrast, MLLS formulates the label shift problem as a maximum likelihood estimation and directly optimizes the likelilhood of target predictions to recover the importance weights. Garg et al. (2020) show that BBSE is roughly equivalent to MLLS under coarse calibration, explaining MLLS's superior empirical performance. At a high level, our method builds on the idea of MLLS but is developed in the context of CSS.

# 3 Preliminaries and problem setup

## 3.1 Contextual stochastic block model (CSBM)

In this work, we consider the Contextual Stochastic Block Model (CSBM) introduced by Deshpande et al. (2018). CSBM is an extension of the classical stochastic block model (SBM) by coupling each node with a feature vector, and has been widely used to study the generalization performance of GNNs (Baranwal et al., 2021; Wang & Wang, 2024) as well as different types of distributional shifts in GDA (Zhu et al., 2023; Liu et al., 2023; 2024c). Concretely, in CSBM, each node $u \in [n]$ is assigned a label $y_u \in \mathcal{Y} = \{1, \dots, L\}$ drawn i.i.d. from a categorical distribution $p(y)$. Conditioned on the labels, the edge $a_{uv}$ between node $u$ and $v$ ($u < v$) is generated independently according to $a_{uv} \mid (y_u, y_v) \sim \mathrm{Ber}(q(y_u, y_v))$, where $q : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is the symmetric connection probability function. We then define the adjacency matrix $A = (a_{uv}) \in \mathbb{R}^{n \times n}$ by setting $a_{vu} = a_{uv}$ for $u < v$ and $a_{uu} = 0$ for all $u$, i.e., $A$ is symmetric with zero diagonal. Given its label, each node $u \in [n]$ is also associated with a feature vector $x_u \in \mathcal{X}$ drawn independently from the class-conditional distribution $x_u \sim p(x \mid y_u)$. Hence, CSBM is fully specified by the class prior $p(y)$, the conditional connection probabilities $q(y, y')$, and the class-conditional distribution $p(x \mid y)$.

## 3.2 Graph domain adaptation setup

We describe the GDA setting that we consider in this work. For the source domain, we observe a labeled source graph with $n^{(0)}$ nodes, $\{(x_u^{(0)}, y_u^{(0)})_{u=1}^{n^{(0)}},\ (a_{uv}^{(0)})_{1 \le u < v \le n^{(0)}}\}$, which is generated from a CSBM with class prior $p^{(0)}(y)$, class-conditional distribution $p^{(0)}(x \mid y)$, and connection probability function $q^{(0)}(y, y')$. Independently of the source dataset, in the target domain, we are given an unlabeled target graph with $n^{(1)}$ nodes, $\{(x_u^{(1)})_{u=1}^{n^{(1)}},\ (a_{uv}^{(1)})_{1 \le u < v \le n^{(1)}}\}$, drawn from a CSBM with class prior $p^{(1)}(y)$, class-conditional distribution $p^{(1)}(x \mid y)$, and connection probability function $q^{(1)}(y, y')$. The target labels are unobserved.

In DA, assumptions relating the source and target distributions are needed to make the DA problem tractable. In the setting of GDA, Liu et al. (2023) introduce the notion of graph structure shift, where the distributions of labels and edges change across domains while the class-conditional distribution is invariant. This shift can be further decomposed into two types: label shift (changes in the marginal class prior), and conditional structure shift (CSS; changes in the edge distribution given labels). While label shift has been widely studied in the DA literature, CSS has received relatively little attention. To address this gap, in this work we focus on CSS by making the following assumption.

**Assumption 3.1 (Conditional structure shift (CSS))** *The class prior and the class-conditional features are invariant across domains, while the conditional edge distributions changes. Formally,* $p^{(0)}(y) = p^{(1)}(y)$ *and* $p^{(0)}(x \mid y) = p^{(1)}(x \mid y)$*, while* $q^{(0)}(y, y') \neq q^{(1)}(y, y')$.

Under Assumption 3.1, the only shift between source and target is in the label-conditioned edge structure. While this assumption may seem restrictive, it isolates CSS from other types of shift and allows for simplified analysis, though our theory can be extended to settings where both CSS and label shift exist (see Appendix B). A convenient way to represent the structural mismatch between source and target graphs is via the importance weight matrix $W_{\mathrm{iw}}^{\star} \in \mathbb{R}^{2 \times L \times L}$, where the entries of $W_{\mathrm{iw}}^{\star}$ are defined as $(W_{\mathrm{iw}}^{\star})_{a,y,y'} := p^{(1)}(a, y, y')/p^{(0)}(a, y, y')$. Because the class prior $p(y)$ is

invariant, this ratio is equivalent to $p^{(1)}(a \mid y, y')/p^{(0)}(a \mid y, y')$. In Section 4.3, we show how the importance weight matrix can be used to correct structural mismatches between the source and target graphs.

**Notation** Throughout, we use $p^{(0)}$ and $p^{(1)}$ to denote the source and target distributions. When a distribution is invariant across domains, we omit the superscript and simply write $p$, e.g., under Assumption 3.1, $p^{(0)}(y) = p^{(1)}(y) = p(y)$ and $p^{(0)}(x \mid y) = p^{(1)}(x \mid y) = p(x \mid y)$. Whenever these distributions involve both discrete and continuous variables, they are understood as densities with respect to the product of counting measure $\mu_C$ and Lebesgue measure $\mu_L$; for instance, $p^{(1)}(x, x', a)$ (two node features $x, x'$ and their connecting edge $a$) is a density with respect to $\mu_L \otimes \mu_L \otimes \mu_C$.

# 4 Proposed method

In this section, we introduce Pairwise-Likelihood maximization for graph Structure Alignment (PLSA), a method to estimate the target connection probabilities under CSS. Before presenting our method, we recall the definition of calibration. A predictor $f : \mathcal{X} \to \Delta^{L-1} := \{z \in \mathbb{R}^L : z \geq 0, \sum_i z_i = 1\}$ is called canonically calibrated (Vaicenavicius et al., 2019) if

$$\mathbb{P}(y = j \mid f(x)) = f_j(x) \quad \text{for all } x \in \mathcal{X} \text{ and for all } j \in \mathcal{Y}. \tag{1}$$

In words, the predicted probabilities match the true conditional probabilities given the prediction score vector (Guo et al., 2017; Kumar et al., 2019). In our analysis, we assume access to a predictor $f$ that is (approximately) calibrated on the source domain—this assumption ensures that PLSA correctly identifies the target connection probabilities under CSS.

## 4.1 Distribution matching for graph structure estimation

Let $f : \mathcal{X} \to \mathcal{Z}$ be a feature map and write the latent variable $z = f(x)$. Let $\mathbb{S}^L$ denote the set of $L \times L$ symmetric matrices, and define $\mathcal{W} := \{W \in \mathbb{S}^L : 0 \leq W \leq 1\}$. For any node pair $u < v$, consider the family of distributions on $(z_u, z_v, a_{uv}) \in \mathcal{Z} \times \mathcal{Z} \times \{0, 1\}$,

$$\mathcal{P} = \Big\{p_W(z_u, z_v, a_{uv}) = \sum_{y_u, y_v = 1}^{L} p(z_u, z_v, y_u, y_v) \left[(1-a_{uv})(1-W_{y_u y_v})+a_{uv}W_{y_u y_v}\right] : W \in \mathcal{W}\Big\}.$$

Under CSS, $p(z_u, z_v, y_u, y_v)$ is invariant across source and target (since $p(y)$ and $p(x \mid y)$ are invariant), and for each $W \in \mathcal{W}$, $(1 - a_{uv})(1 - W_{y_u y_v}) + a_{uv}W_{y_u y_v}$ is the pmf of Bernoulli for $a_{uv}$ with parameter $W_{y_u y_v}$. Hence every $p_W \in \mathcal{P}$ is a valid density on $\mathcal{Z} \times \mathcal{Z} \times \{0, 1\}$.

Now let $W^\star \in \mathbb{S}^L$ denote the matrix of target connection probabilities, with entries $W^\star_{yy'} := q^{(1)}(y, y')$. Clearly $W^\star \in \mathcal{W}$, and because $(z_u, z_v) \perp a_{uv} \mid (y_u, y_v)$ and $p(z_u, z_v, y_u, y_v)$ is shared across domains, the target distribution satisfies $p^{(1)}(z_u, z_v, a_{uv}) = p_{W^\star}(z_u, z_v, a_{uv})$. Thus, $W^\star$ is the solution to the following distribution matching equation

$$p^{(1)}(z_u, z_v, a_{uv}) = p_W(z_u, z_v, a_{uv}) \text{ for all } (z_u, z_v, a_{uv}) \in \mathcal{Z} \times \mathcal{Z} \times \{0, 1\}. \tag{2}$$

Although multiple $W \in \mathcal{W}$ may also satisfy this equation, the following proposition shows that under mild conditions on $p(y)$, $q^{(1)}(y, y')$, and $p(z \mid y)$, $W^\star$ is the unique solution to (2).

**Proposition 4.1 (Identifiability)** *Suppose Assumption 3.1 holds. Assume $p(y) > 0$ for all $y \in \mathcal{Y}$ and $0 < q^{(1)}(y, y') < 1$ for all $y, y' \in \mathcal{Y}$. Then any $W \in \mathcal{W}$ satisfying (2) equals $W^\star$ if and only if $\{p(z \mid y), y = 1, \ldots, L\}$ is linearly independent (as functions on $\mathcal{Z}$).*

The conditions in Proposition 4.1 ensure that every class should appear with positive probability, and for each label pair, both edges and non-edges occur with positive probability. We assume these conditions throughout the paper. Under this setting, the linear independence condition rules out the possibility that any class-conditional distribution can be expressed as a nontrivial linear combination of the others, which is precisely what guarantees identifiability. When $f$ is chosen as a calibrated predictor, Garg et al. (2020, Proposition 1) show that linear independence holds if $\mathbb{E}[zz^\top]$ is invertible; we return to this point in Section 5 (c.f. Proposition 5.4). Under these assumptions, equation (2)

suggests we can estimate $W^\star$ by aligning $p_W(z_u, z_v, a_{uv})$ to the target distribution $p^{(1)}(z_u, z_v, a_{uv})$. Following Garg et al. (2020) for label shift, we use a KL-divergence criterion and estimate $W^\star$ by minimizing the KL-divergence between $p^{(1)}(z_u, z_v, a_{uv})$ and $p_W(z_u, z_v, a_{uv})$, which leads to the pairwise likelihood formulation (see Section 4.2).

#### 4.1.1 EDGE-CONDITIONED DISTRIBUTION MATCHING

Here, we derive a conditional variant of the distribution matching (2) by conditioning on the presence of an edge. Define the family

$$\mathcal{P}_{\text{con}} = \big\{p_W(z_u, z_v \mid a_{uv} = 1) = \sum_{y_u, y_v = 1}^{L} p^{(0)}(z_u, z_v, y_u, y_v \mid a_{uv} = 1) \cdot W_{y_u y_v} : W \in \mathcal{W}_{\text{con}}\big\},$$

where the parameter space is $\mathcal{W}_{\text{con}} = \{W \in \mathbb{S}^L : \sum_{y_u, y_v = 1}^{L} p^{(0)}(y_u, y_v \mid a_{uv} = 1) \cdot W_{y_u y_v} = 1, W \geq 0\}$, so that $p_W(\cdot, \cdot \mid a_{uv} = 1)$ integrates to one (we use the notation $p^{(0)}$ since conditioned on $a_{uv} = 1$, the pairs $(z_u, z_v)$ and $(y_u, y_v)$ are not invariant). Let $W^\star_{\text{con}} \in \mathcal{W}_{\text{con}}$ denote the edge-conditioned importance weight matrix, with $(y_u, y_v)$ entry defined as $p^{(1)}(y_u, y_v \mid a_{uv} = 1)/p^{(0)}(y_u, y_v \mid a_{uv} = 1)$. Since $z_u \mid y_u$ is invariant across domains, we can check that $p_{W^\star_{\text{con}}}(z_u, z_v \mid a_{uv} = 1) = p^{(1)}(z_u, z_v \mid a_{uv} = 1)$. Therefore, in order to estimate $W^\star_{\text{con}}$, we can find $W \in \mathcal{W}_{\text{con}}$ that satisfies the edge-conditioned matching equation

$$p^{(1)}(z_u, z_v \mid a_{uv} = 1) = p_W(z_u, z_v \mid a_{uv} = 1) \text{ for all } (z_u, z_v) \in \mathcal{Z} \times \mathcal{Z}. \tag{3}$$

Now let $h : \mathcal{X} \to \mathcal{Y}$ be a black-box classifier and set $z = h(x) \in \mathcal{Y}$ (so $\mathcal{Z} = \mathcal{Y}$). Define $\Sigma \in \mathbb{R}^{L^2 \times L^2}$ and $\nu \in \mathbb{R}^{L^2}$ by $\Sigma_{(\hat{y}, \hat{y}'), (y, y')} = p^{(0)}(h(x_u) = \widehat{y}, h(x_v) = \widehat{y}', y_u = y, y_v = y' \mid a_{uv} = 1)$ and $\nu_{(\hat{y}, \hat{y}')} = p^{(1)}(h(x_u) = \widehat{y}, h(x_v) = \widehat{y}' \mid a_{uv} = 1)$. Then the equation (3) becomes the linear system $\Sigma \cdot \text{vec}(W) = \nu$ subject to $W \in \mathcal{W}$, where $\text{vec}(W)$ is the vectorization of $W$. This is precisely the formulation of Pair-Align introduced in Liu et al. (2024c). Furthermore, Liu et al. (2024c) observe that Structural Re-weighting (StruRW) (Liu et al., 2023) is a special case of Pair-Align under the additional assumptions of no label shift and perfect prediction on the target graph. Hence both StruRW and Pair-Align can be viewed as edge-conditioned distribution matching in the latent space with black-box classifier $h$.

Finally, comparing the two formulations (2) and (3), we make the following observations: (i) unlike (2), the edge-conditioned version (3) uses only connected node pairs while discarding unconnected pairs. This can substantially reduce effective sample size and increase variance, particularly when graph is sparse. (ii) the unconditioned equation (2) with $W = W^\star$ holds under CSS (both $p(y)$ and $p(x \mid y)$ invariant), whereas equation (3) with $W = W^\star_{\text{con}}$ requires only the invariance of $p(x \mid y)$ and thus it is still valid under label shift; however, an alternative parameterization of $\mathcal{P}$ also allows for extending the unconditioned matching equation to incorporate label shift; see Appendix B for details.

### 4.2 PAIRWISE LIKELIHOOD MAXIMIZATION FOR GRAPH STRUCTURE ALIGNMENT

We begin with the population formulation of PLSA. By Proposition 4.1, the target connection probabilities $W^\star$ minimize the KL-divergence $D_{\text{KL}}\left(p^{(1)}(z_u, z_v, a_{uv}) \,\|\, p_W(z_u, z_v, a_{uv})\right) = \mathbb{E}[\log(p^{(1)}(z_u^{(1)}, z_v^{(1)}, a_{uv}^{(1)})/p_W(z_u^{(1)}, z_v^{(1)}, a_{uv}^{(1)}))]$. Since $p(z_u, z_v, y_u, y_v) = p(y_u \mid z_u)\, p(y_v \mid z_v)\, p(z_u)\, p(z_v)$, substituting this into $p_W$ and ignoring terms that do not depend on $W$ yields the equivalent maximization problem

$$W^\star = \arg\max_{W \in \mathcal{W}} \mathbb{E}\Big[\log \sum_{y, y' = 1}^{L} p(y \mid z_u^{(1)}) p(y' \mid z_v^{(1)}) [(1 - a_{uv}^{(1)})(1 - W_{yy'}) + a_{uv}^{(1)} W_{yy'}]\Big]. \tag{4}$$

In practice, $p(y \mid z)$ is unknown, so we approximate it with a probabilistic predictor $f : \mathcal{X} \to \Delta^{L-1}$ trained on the labeled source data. If $f$ is (canonically) calibrated on the source (equation (1)) (so $\mathcal{Z} = \Delta^{L-1}$), then $p(y \mid z) = p(y \mid f(x)) = f_y(x)$. Plugging this into (4) gives

$$W_{\text{f}} := \arg\max_{W \in \mathcal{W}} \mathbb{E}\Big[\log \sum_{y, y' = 1}^{L} f_y(x_u^{(1)}) f_{y'}(x_v^{(1)}) [(1 - a_{uv}^{(1)})(1 - W_{yy'}) + a_{uv}^{(1)} W_{yy'}]\Big]. \tag{5}$$

The objective in (5) is well defined for any predictor $f$. When $f$ is calibrated, it coincides with (4). Hence, if $W^\star$ is the unique solution to (4), we must have $W_{\mathrm{f}} = W^\star$. This is formalized in the following proposition.

**Proposition 4.2** *Suppose $f : \mathcal{X} \to \Delta^{L-1}$ is canonically calibrated on the source distribution. If $\{p(z \mid y), y = 1, \ldots, L\}$ is linearly independent, then $W^\star = W_{\mathrm{f}}$.*

Having introduced the population formulation, we now define the finite-sample PLSA estimator based on the unlabeled target data $\{(x_u^{(1)})_{u=1}^{n^{(1)}},\ (a_{uv}^{(1)})_{1\le u<v\le n^{(1)}}\}$:

$$\widehat{W}_{\mathrm{f}} := \operatorname*{arg\,max}_{W\in\mathcal{W}} \frac{1}{\binom{n^{(1)}}{2}} \sum_{u<v} \log\Big( \sum_{y,y'=1}^{L} f_y(x_u^{(1)}) f_{y'}(x_v^{(1)}) [(1-a_{uv}^{(1)})(1-W_{yy'}) + a_{uv}^{(1)} W_{yy'}]\Big). \quad (6)$$

The feasible set $\mathcal{W}$ is convex, and the objective in (6) is concave in $W$, so equation 6 is a convex program, which can be solved using any convex optimization algorithms. We use the projected gradient descent method in our numerical experiments. Note that the source data is not directly used in formulating the objective (6) while it is only used to pre-train the calibrated predictor $f$.

Our proposed PLSA with source-calibrated predictor is inspired by Garg et al. (2020) where they develop maximum likelihood estimation procedure with calibrated predictor $f$ for label shift estimation (MLLS). In contrast, Pair-Align adopts a moment-matching approach based on the (edge-conditioned) confusion matrix that is analogous to BBSE (Lipton et al., 2018). As noted by Garg et al. (2020), MLLS typically outperforms BBSE due to the information loss by coarse calibration in BBSE. Hence, we can expect a similar advantage for PLSA over Pair-Align in the GDA setting.

### 4.3 Reweighting the source graph

We describe a simple sampling-based procedure to adjust the source graph so that under CSS, its edge distribution matches the target graph. The idea is to reweight source edges using importance ratios between target and source connection probabilities. For labels $y, y' \in [L]$, let $r_{yy'} := (W^\star_{\mathrm{iw}})_{1,y,y'} = q^{(1)}(y,y')/q^{(0)}(y,y')$. Here $q^{(1)}$ is estimated by PLSA, and $q^{(0)}$ can be estimated by the empirical edge ratio in the labeled source graph, $\widehat{q}^{(0)}(y,y') := \frac{\sum_{u<v} \mathbf{1}\{y_u^{(0)}=y,\, y_v^{(0)}=y', a_{uv}^{(0)}=1\}}{\sum_{u<v} \mathbf{1}\{y_u^{(0)}=y,\, y_v^{(0)}=y'\}}$. Thus $r_{yy'}$ is observable from labeled source and unlabeled target graph.

Fix a pair $u < v$ with $(y_u, y_v) = (y, y')$ and write $a = a_{uv}^{(0)} \sim \mathrm{Ber}(q^{(0)}(y,y'))$. (i) Case $r_{yy'} < 1$: draw $z \sim \mathrm{Ber}(r_{yy'})$ independently and set $\widetilde{a} = a \cdot z$. Then $\mathbb{P}(\widetilde{a} = 1 \mid y, y') = \mathbb{P}(a = 1 \mid y, y')\mathbb{P}(z = 1) = q^{(0)}(y,y')\, r_{yy'} = q^{(1)}(y,y')$. (ii) Case $r_{yy'} > 1$: draw $z \sim \mathrm{Ber}(\alpha_{yy'})$ with

$$\alpha_{yy'} := \frac{q^{(1)}(y,y') - q^{(0)}(y,y')}{1 - q^{(0)}(y,y')} \in [0,1],$$

independently and set $\widetilde{a} = \max\{a, z\}$. Then $\mathbb{P}(\widetilde{a} = 1 \mid y, y') = q^{(0)}(y,y') + (1 - q^{(0)}(y,y'))\alpha_{yy'} = q^{(1)}(y,y')$. (iii) Case $r_{yy'} = 1$: set $\widetilde{a} = a$. By construction, for every $(y,y')$, we have $\widetilde{a}_{uv}^{(0)} \mid (y_u = y, y_v = y') \sim \mathrm{Ber}(q^{(1)}(y,y'))$. Under CSS, since $p(y)$ and $p(x \mid y)$ are invariant, replacing $a_{uv}^{(0)}$ by $\widetilde{a}_{uv}^{(0)}$ ensures that the joint distribution of $(x_u^{(0)}, x_v^{(0)}, \widetilde{a}_{uv}^{(0)}, y_u^{(0)}, y_v^{(0)})$ matches that of the target. Consequently, any GNN trained on the adjusted source graph data is aligned with the target distribution and generalizes to the target data.

## 5 Theoretical results

We now develop theoretical error bounds for the PLSA estimator when the data are generated from CSBM. Our analysis proceeds by identifying two sources of errors in estimating $W^\star$: (i) the finite-sample error (i.e., the gap between optimizing the population objective (5) and the empirical objective (6)), and (ii) the error due to miscalibration of the predictor $f$ (i.e., the objectives (4) and (5) are different when $f$ is not perfectly calibrated).

To facilitate our analysis, we impose the following assumption on the pairwise likelihood objective. For a given predictor $f$, define $S_{\mathrm{f}}(W; x, x', a) := f(x)^\top((1-a)(1-W) + aW)f(x')$.

**Assumption 5.1** *There exists a constant $\tau_{\min} > 0$ such that for all $(x, x', a) \in \mathcal{X} \times \mathcal{X} \times \{0, 1\}$ in the support of $p^{(1)}(x_u, x_v, a_{uv})$, we have $S_{\mathrm{f}}(W_{\mathrm{f}}; x, x', a) \geq \tau_{\min}$, and $S_{\mathrm{f}}(W^\star; x, x', a) \geq \tau_{\min}$.*

Assumption 5.1 is analogous to Condition 1 in Garg et al. (2020) for label shift: if $f$ is perfectly calibrated, we have $W^\star = W_{\mathrm{f}}$. In this case, whenever the objective $\mathbb{E}[\log S_{\mathrm{f}}(W_{\mathrm{f}}; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})]$ is finite, $S_{\mathrm{f}}(W_{\mathrm{f}}; x, x', a)$ (and hence $S_{\mathrm{f}}(W^\star; x, x', a)$) must be bounded away from zero with high probability, since it is always upper bounded. When $f$ is miscalibrated but sufficiently close to a calibrated predictor, Assumption 5.1 is still reasonable to make because in practice post-hoc recalibration on the source data is performed to improve the calibration of $f$.

We now state our main theoretical results. For a symmetric $W \in \mathbb{S}^L$, let $\mathrm{vech}(W) \in \mathbb{R}^{L(L+1)/2}$ denote the half-vectorization (Magnus & Neudecker, 2019, Chapter 3.8), obtained by stacking the upper-triangular entries of $W$. Define $\ell_{\mathrm{f}}(W) := \mathbb{E}[\log S_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})]$ and let $\lambda_{\min,\mathrm{f}} > 0$ denote the minimum eigenvalue of $-\nabla^2 \ell_{\mathrm{f}}(W_{\mathrm{f}})$ where the Hessian is taken with respect to $\mathrm{vech}(W)$.

**Theorem 5.2** *Suppose the target data is generated according to CSBM and the predictor $f$ satisfies Assumption 5.1. Then there exist universal constants $c, c' > 0$ such that for $\delta \in (0, 1/2)$, if $n^{(1)} \geq \max\left\{c\tau_{\min}^{-4}(\lambda_{\min,\mathrm{f}})^{-2}\log(8L^2/\delta), (\log(3L^2))^2\log(8/\delta)\right\}$, with probability at least $1 - 2\delta$,*

$$\left\|\mathrm{vech}(\widehat{W}_{\mathrm{f}}) - \mathrm{vech}(W_{\mathrm{f}})\right\|_2 \leq c'\tau_{\min}^{-3}(\lambda_{\min,\mathrm{f}})^{-1}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}}.$$

Theorem 5.2 shows that when the parameters $\tau_{\min}, \lambda_{\min,\mathrm{f}} > 0$ are constants, the unlabeled target sample size of $n^{(1)} \gtrsim \mathcal{O}(\log L)$ suffice to guarantee small finite-sample error. In proving Theorem 5.2, the main technical challenge is that the empirical objective in (6) does not fit the classical U-statistics framework, because the pairwise node features $(x_u, x_v)$ and the edge $a_{uv}$ are dependent through the node labels. Consequently, standard concentration tools for U-statistics based on independent random variables do not directly apply. To deal with this, we exploit the conditional independence structure of $(x_u, x_v)$ and $a_{uv}$ given labels under CSBM, together with the matrix-valued concentration bounds for U-statistics (Minsker & Wei, 2019), to obtain the needed bounds.

Next, given $f$, define the new predictor $f^\star(x) = p(y \mid f(x))$. By construction, $f^\star$ is canonically calibrated and can be viewed as the closest calibrated version of $f$ (Vaicenavicius et al., 2019, Equation (4)). Our next theorem controls the error due to the miscalibration of $f$ on the source.

**Theorem 5.3** *Suppose the source and target data are generated according to CSBM, and additionally Assumptions 3.1, 5.1 hold. Then there exists a universal constant $c > 0$ such that*

$$\|\mathrm{vech}(W_{\mathrm{f}}) - \mathrm{vech}(W^\star)\|_2 \leq c\tau_{\min}^{-4}(\lambda_{\min,\mathrm{f}})^{-1} \cdot \mathrm{MC}(f),$$

*where $\mathrm{MC}(f) := \mathbb{E}_{x\sim p^{(0)}(x)}\left[\|f(x) - f^\star(x)\|_1\right]$ is the miscalibration of $f$ in terms of $\ell_1$ norm.*

For binary classification problems ($L = 2$), $\mathrm{MC}(f)$ is also known as expected calibration error (ECE) (Guo et al., 2017). For multiclass problems ($L > 2$), $\mathrm{MC}(f)$ has been used as a miscalibration metric in the literature (e.g. Vaicenavicius et al. (2019); Popordanoska et al. (2022)).

In both Theorem 5.2 and Theorem 5.3, the error bounds crucially depend on $\lambda_{\min,\mathrm{f}}$. The next theorem provides a sufficient condition to ensure $\lambda_{\min,\mathrm{f}}$ is strictly positive.

**Proposition 5.4** *Suppose that there exists $\overline{\lambda}_{\min} > 0$ such that $\mathbb{E}[f(x_u^{(1)})f(x_u^{(1)})^\top] \succeq \overline{\lambda}_{\min}\mathbb{I}_L$. Then $\lambda_{\min,\mathrm{f}}$ is lower bounded by $\overline{\lambda}_{\min}^2$, i.e., $\lambda_{\min,\mathrm{f}} \geq \overline{\lambda}_{\min}^2$.*

Under CSS, if the condition $\mathbb{E}[f(x_u^{(1)})f(x_u^{(1)})^\top] \succeq \overline{\lambda}_{\min}\mathbb{I}_L$ is satisfied, $\mathbb{E}[f(x_u^{(0)})f(x_u^{(0)})^\top]$ is also invertible. According to Garg et al. (2020, Proposition 1), if $f$ is perfectly calibrated, this condition implies that $\{p(f(x) \mid y), y = 1, \ldots, L\}$ is linearly independent, and therefore $W^\star$ is the unique maximizer of $\ell_{\mathrm{f}}(W)$ due to Proposition 4.1. Theorem 5.3 and Proposition 5.4 show that when $f$ is miscalibrated, the same condition further guarantees that the population-based estimator $W_{\mathrm{f}}$ is close to $W^\star$ where the difference depends on the miscalibration error.

Combining together Theorem 5.2, Theorem 5.3, and Proposition 5.4, it follows that the estimation error of the finite-sample PLSA estimator is bounded as (assuming $\tau_{\min}$ and $L$ are constants)

$$\overline{\lambda}_{\min}^{-2} \cdot \left(\mathcal{O}\left(1/\sqrt{n^{(1)}}\right) + \mathrm{MC}(f)\right).$$

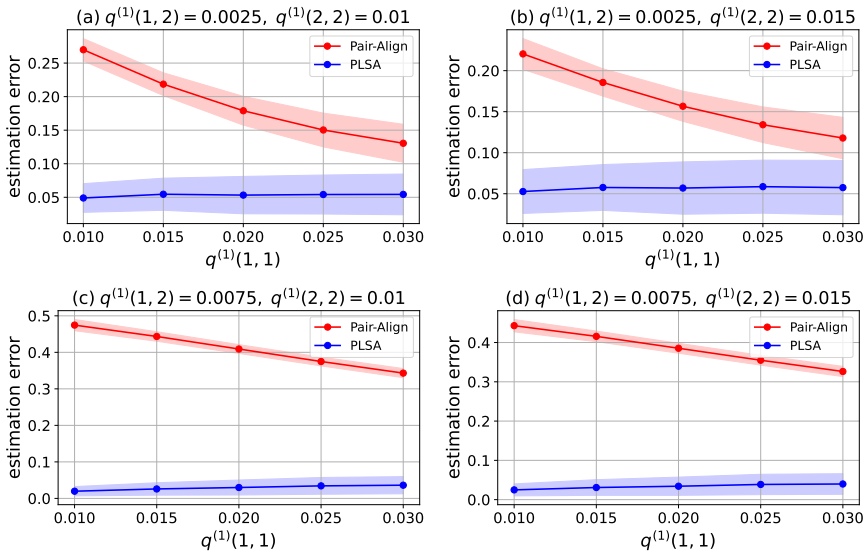


Figure 1: Estimation error of the importance weights as the target connection probability varies under CSBM with binary classes and a uniform class prior. The results are averaged over 10 trials.

If $f$ assigns nonvanishing probability mass to each class, $\overline{\lambda}_{\min}$ is bounded away from zero and does not significantly degrade the rate of estimation error.

## 6 Numerical Experiments

In this section, we evaluate the empirical performance of our method on both simulated data from CSBM and the Airport dataset. In all experiments, we solve the convex program (6) via projected gradient descent and apply post-hoc recalibration on a held-out source calibration dataset using Bias-Corrected Temperature Scaling (BCTS) (Alexandari et al., 2020).

**CSBM experiments** We first study the behavior of our method on simulated data. Both source (training and calibration) and target data are generated from CSBM with 5000 nodes each. The node labels are sampled uniformly, and for each node, we generate 20-dimensional Gaussian features from $\mathcal{N}(\mu_y, \sigma^2\mathbb{I})$, where $\sigma = 1$ and for each label $y \in \mathcal{Y}$, $\mu_y$ is drawn from $\mathcal{N}(0, \mathbb{I}/20)$. In the first setting, we consider binary classes ($L = 2$) and vary the target connection probabilities so that CSS is present between source and target graphs. Specifically, for the source graph, we fix $q^{(0)}(1,1) = q^{(0)}(2,2) = 0.02$ and $q^{(0)}(1,2) = q^{(0)}(2,1) = 0.005$, while for the target graph, we vary $q^{(1)}(1,1) \in \{0.01, 0.015, \ldots, 0.03\}$, $q^{(1)}(1,2) \in \{0.0025, 0.0075\}$, and $q^{(1)}(2,2) \in \{0.01, 0.015\}$.

In the second setting, we vary the number of nodes in both the source and target graphs with $n^{(0)}, n^{(1)} \in \{1000, 1250, \ldots, 10000\}$. We consider both binary and three-class ($L = 3$) cases where we fix the source connection probabilities as $q^{(0)}(y, y') = 0.02$ for $y = y'$ and $q^{(0)}(y, y') = 0.005$ for $y \neq y'$, while the target connection probabilities are set differently so that the source and target graphs are different. Additional details are given in Appendix C.

The results for these two settings are shown in Figure 1 and Figure 2, respectively. In Figure 1, PLSA consistently outperforms Pair-Align in estimating the importance weights $((W^{\star}_{\text{iw}})_{1,y,y'})_{y,y'=1}^{L}$, measured by the relative $\ell_2$ norm error on the half-vectorized weights. The performance gap is especially pronounced when $q^{(1)}(1,1)$ is small. Since Pair-Align only uses connected edges to estimate the importance weights, its accuracy significantly degrades when the graph becomes sparser; whereas PLSA exhibits stable performance due to its use of all pairwise nodes. Overall the performance of both methods also improve when cross-class connection probabilities are small (compare panels (a),(b) vs. (c),(d)) and when within-class connection probabilities increase (compare panels (a),(c)

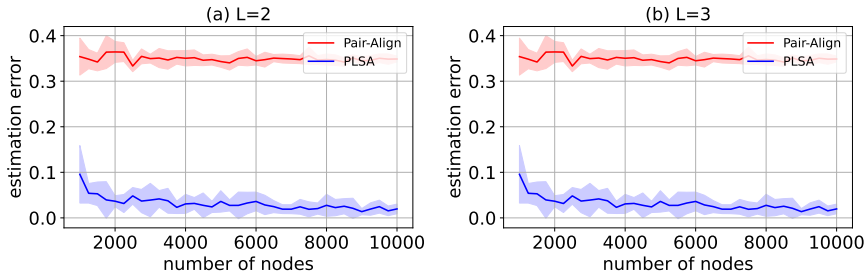


Figure 2: Estimation error of the importance weights as the number of source and target graph nodes varies under CSBM with binary or three classes and a uniform class prior. The results are averaged over 10 trials.

vs. (b),(d)). In Figure 2, we observe that as the number of nodes increases, the error of PLSA decreases rapidly, while Pair-Align shows limited improvement. This indicates that for Pair-Align, the connection probabilities (i.e., edge density) are more critical than the graph size, whereas PLSA benefits from larger number of nodes as predicted by our theory. Additional results are also provided in Appendix D.

**Airport experiments** To illustrate the application of our method, we consider the Airport dataset (Zhu et al., 2021b). This dataset contains three domains, Brazil (B), Europe (E), and the USA (U), where nodes represent airports and edges denote flight connections. Labels correspond to airport activity levels, measured by flight counts and passenger numbers. Since the original dataset does not contain node features, we synthetically generate 32-dimensional features from a Gaussian distribution $\mathcal{N}(\mu_y \mathbf{1}, \sigma^2 \mathbb{I})$, where $\mu_y \in \{-0.5, 0, 0.5, 1\}$ and $\sigma = 1$. As noted by Liu et al. (2024b), the Airport dataset is dominated by structural shift, which makes it well-suited for studying CSS.

The dataset contains 131 nodes for Brazil, 399 nodes for Europe, and 1190 nodes for the USA. Since a sufficiently large source data is needed to train a calibrated predictor, we use the USA as the source domain and Brazil and Europe as the target domains. We split the source nodes into 80% for training and 20% for post-hoc recalibration. After we estimate the importance weights, we apply the edge reweighting scheme in Section 4.3 to adjust the source graph and then train GNNs on the adjusted source graph data for target label prediction. The results are shown in Table 1, where ERM denotes a GNN trained on the source data and applied to the target without reweighting. We find that GNNs trained with PLSA-based graph reweighting achieve the best target performance on both target domains, demonstrating that with sufficient source data and calibrated predictors, PLSA is an effective approach for correcting CSS.

| Method | U $\to$ B | U $\to$ E |
|---|---|---|
| ERM | $53.36 \pm 6.45$ | $45.06 \pm 6.80$ |
| Pair-Align | $48.86 \pm 4.85$ | $45.16 \pm 6.66$ |
| PLSA | $60.92 \pm 4.29$ | $50.20 \pm 3.54$ |

Table 1: Target accuracy for Airport where the results are averaged over 10 trials.

## 7 Discussion

In this paper, we present a unified framework for addressing CSS by formulating CSS estimation as distribution matching over node-pair features and edges in the latent space. This framework provides a principled way to view existing methods for CSS as special cases, while also motivating our new method, PLSA, which benefits from calibrated source predictors for more accurate estimation. Our theoretical and empirical results demonstrate that PLSA accurately estimates CSS and enables source graph reweighting, allowing downstream GNNs to achieve strong target prediction performance. An interesting future direction is to extend PLSA beyond CSBM to richer random graph models, such as graphon models, where heterogeneity and sparsity better reflect real networks.

## REFERENCES


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 (ICML)*, pp. 222–232. PMLR, 2020.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations (ICLR)*, 2019.

Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization. In *International Conference on Machine Learning (ICML)*, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Ruichu Cai, Fengzhu Wu, Zijian Li, Pengfei Wei, Lingling Yi, and Kun Zhang. Graph domain adaptation: A generative view. *ACM Transactions on Knowledge Discovery from Data*, 18(3): 1–24, 2024.

Ines Chami, Sami Abu-El-Haija, Bryan Perozzi, Christopher Ré, and Kevin Murphy. Machine learning on graphs: A model and comprehensive taxonomy. *Journal of Machine Learning Research*, 23(89):1–64, 2022.

Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *Journal of Machine Learning Research*, 22(261):1–80, 2021.

Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. Contextual stochastic block models. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Ruiyi Fang, Bingheng Li, Jingyu Zhao, Ruizhi Pu, Qiuhao Zeng, Gezheng Xu, Charles Ling, and Boyu Wang. Homophily enhanced graph domain adaptation. *arXiv preprint arXiv:2505.20089*, 2025.

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

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:3290–3300, 2020.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330. PMLR, 2017.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 1989–1998. PMLR, 2018.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 527–536. PMLR, 2019.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

A J Lee. *U-statistics: Theory and Practice*. Routledge, 2019.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, pp. 3122–3130. PMLR, 2018.

Meihan Liu, Zeyu Fang, Zhen Zhang, Ming Gu, Sheng Zhou, Xin Wang, and Jiajun Bu. Rethinking propagation for unsupervised graph domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 13963–13971, 2024a.

Meihan Liu, Zhen Zhang, Jiachen Tang, Jiajun Bu, Bingsheng He, and Sheng Zhou. Revisiting benchmarking and understanding unsupervised graph domain adaptation. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:89408–89436, 2024b.

Shikun Liu, Tianchun Li, Yongbin Feng, Nhan Tran, Han Zhao, Qiang Qiu, and Pan Li. Structural reweighting improves graph domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 21778–21793. PMLR, 2023.

Shikun Liu, Deyu Zou, Han Zhao, and Pan Li. Pairwise alignment improves graph domain adaptation. In *Forty-first International Conference on Machine Learning (ICML)*, 2024c.

Shuhan Liu and Kaize Ding. Beyond generalization: A survey of out-of-distribution adaptation on graphs. *arXiv preprint arXiv:2402.11153*, 2024.

Junyu Luo, Yuhao Tang, Yiwei Fu, Xiao Luo, Zhizhuo Kou, Zhiping Xiao, Wei Ju, Wentao Zhang, and Ming Zhang. Sparse causal discovery with generative intervention for unsupervised graph domain adaptation. *arXiv preprint arXiv:2507.07621*, 2025.

Jing Ma. A survey of out-of-distribution generalization for graph machine learning from a causal view. *AI Magazine*, 45(4):537–548, 2024.

Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

Stanislav Minsker and Xiaohan Wei. Moment inequalities for matrix-valued U-statistics of order 2. *Electronic Journal of Probability*, 24:1–50, 2019.

Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable Lp canonical calibration error estimator. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:7933–7946, 2022.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3459–3467. PMLR, 2019.

Hai-Xiao Wang and Zhichao Wang. Optimal exact recovery in semi-supervised learning: A study of spectral methods and graph convolutional networks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Keru Wu, Yuansi Chen, Wooseok Ha, and Bin Yu. Prominent roles of conditionally invariant components in domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 26(110):1–92, 2025.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations (ICLR)*, 2022.

Shirley Wu, Kaidi Cao, Bruno Ribeiro, James Zou, and Jure Leskovec. GraphMETRO: Mitigating complex graph distribution shifts via mixture of aligned experts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning (ICML)*, pp. 6872–6881. PMLR, 2019.

Jiaren Xiao, Quanyu Dai, Xiaochen Xie, Qi Dou, Ka-Wai Kwok, and James Lam. Domain adaptive graph infomax via conditional adversarial networks. *IEEE Transactions on Network Science and Engineering*, 10(1):35–52, 2022.

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. Graph domain adaptation via theory-grounded spectral regularization. In *The eleventh International Conference on Learning Representations (ICLR)*, 2023.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 7523–7532. PMLR, 2019.

Qi Zhu, Natalia Ponomareva, Jiawei Han, and Bryan Perozzi. Shift-robust GNNs: Overcoming the limitations of localized graph training data. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:27965–27977, 2021a.

Qi Zhu, Carl Yang, Yidan Xu, Haonan Wang, Chao Zhang, and Jiawei Han. Transfer learning of graph neural networks with ego-graph information maximization. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:1766–1779, 2021b.

Qi Zhu, Chao Zhang, Chanyoung Park, Carl Yang, and Jiawei Han. Shift-robust node classification via graph clustering co-training. In *NeurIPS 2022 Workshop: New Frontiers in Graph Learning*, 2022.

Qi Zhu, Yizhu Jiao, Natalia Ponomareva, Jiawei Han, and Bryan Perozzi. Explaining and adapting graph conditional shift. *arXiv preprint arXiv:2306.03256*, 2023.

## A Error bounds for estimating importance weights

For GNNs to perform well on the target domain for downstream tasks, we need to reweight the source graph using the importance weights

$$r_{yy'} = (W^{\star}_{\text{iw}})_{1,y,y'} = \frac{q^{(1)}(y,y')}{q^{(0)}(y,y')},$$

as described in Section 4.3. In this section, we briefly explain how to obtain theoretical error bounds for estimating these importance weights.

The PLSA estimator provides estimates of the target connection probabilities, $W^{\star}_{yy'} = q^{(1)}(y,y')$ for $y, y' \in \mathcal{Y}$. Consider the estimator

$$\widehat{r}_{yy'} := \frac{\widehat{q}^{(1)}(y,y')}{\widehat{q}^{(0)}(y,y')},$$

where $\widehat{q}^{(1)}(y,y') = (\widehat{W}_{\text{f}})_{yy'}$ is the $(y,y')$ entry of the PLSA estimator, and $\widehat{q}^{(0)}(y,y')$ is the empirical edge ratio in the source graph, as defined in Section 4.3, i.e.,

$$\widehat{q}^{(0)}(y,y') = \frac{\sum_{u<v} \mathbf{1}\{y_u^{(0)} = y,\, y_v^{(0)} = y', a_{uv}^{(0)} = 1\}}{\sum_{u<v} \mathbf{1}\{y_u^{(0)} = y,\, y_v^{(0)} = y'\}}.$$

Since we already have error bouds for $\widehat{q}^{(1)}(y,y')$ from Section 5, we only need to control the error between $\widehat{q}^{(0)}(y,y')$ and $q^{(0)}(y,y')$. Both the numerator and denominator of $\widehat{q}^{(0)}(y,y')$ are U-statistics, so concentration bounds (see Lemma E.1 in Appendix E) imply that $\widehat{q}^{(0)}(y,y')$ converges to $q^{(0)}(y,y')$ at the parametric rate $1/\sqrt{n^{(0)}}$.

Combining with the theoretical results in Theorem 5.2 and Theorem 5.3, we can obtain the following (asymptotic) error bound for the estimated importance weights:

$$\mathcal{O}_p\left(\frac{1}{\sqrt{n^{(1)}}}+\frac{1}{\sqrt{n^{(0)}}}\right)+\mathrm{MC}(f).$$

Since the result follows in a straightforward manner, we omit the detailed proof.

## B Distribution matching in the presence of both CSS and label shift

Consider the case where both CSS and label shift are present across domains, i.e., $q^{(0)}(y,y') \neq q^{(1)}(y,y')$ and $p^{(0)}(y) \neq p^{(1)}(y)$ while $p^{(0)}(x \mid y) = p^{(1)}(x \mid y) = p(x \mid y)$. In this setting, because of label shift, $W^\star$ no longer solves the matching equation (2), so the PLSA estimator based on that equation is not consistent for $W^\star$. One way to fix this is to reweight class priors in the considered family of distributions. Specifically, let $(w^\star_{\mathrm{iw}})_y := p^{(1)}(y)/p^{(0)}(y)$ denote the importance weights for the class priors, and define the modified family of distributions

$$\widetilde{p}_W(z_u,z_v,a_{uv}) = \sum_{y_u,y_v=1}^{L} p^{(0)}(z_u,z_v,y_u,y_v)(w^\star_{\mathrm{iw}})_y(w^\star_{\mathrm{iw}})_{y'}\left[(1-a_{uv})(1-W_{y_uy_v})+a_{uv}W_{y_uy_v}\right],$$

for $W \in \mathcal{W}$. Under CSBM, we can verify

$$\begin{aligned}p^{(0)}(z_u,z_v,y_u,y_v)(w^\star_{\mathrm{iw}})_y(w^\star_{\mathrm{iw}})_{y'} &= p(z_u \mid y_u)p(z_v \mid y_v)p^{(0)}(y_u)p^{(0)}(y_v)(w^\star_{\mathrm{iw}})_y(w^\star_{\mathrm{iw}})_{y'}\\ &= p(z_u \mid y_u)p(z_v \mid y_v)p^{(1)}(y_u)p^{(1)}(y_v)\\ &= p^{(1)}(z_u,z_v,y_u,y_v),\end{aligned}$$

so $W^\star$ now becomes a solution to the equation $\widetilde{p}_W(z_u,z_v,a_{uv}) = p^{(1)}(z_u,z_v,a_{uv})$. In the finite-sample setting, the importance weights $w^\star_{\mathrm{iw}}$ can be estimated via BBSE or MLLS and plugged in to approximate this family.

Alternatively, we may also parameterize the family of distributions directly through the importance weights over $(a_{uv},y_u,y_v)$. Define

$$\begin{aligned}\mathcal{W}_{\mathrm{iw}} := \Big\{(W_0,W_1) \in \mathbb{S}^L \times \mathbb{S}^L : \sum_{a_{uv}\in\{0,1\}} \sum_{y_u,y_v=1}^{L} p^{(0)}(a_{uv},y_u,y_v)\cdot\\ \left[(1-a_{uv})(W_0)_{y_uy_v}+a_{uv}(W_1)_{y_uy_v}\right]=1, W_0,W_1 \geq 0\Big\},\end{aligned}$$

and consider the family parameterized by $W \in \mathcal{W}_{\mathrm{iw}}$,

$$p^{\mathrm{iw}}_W(z_u,z_v,a_{uv}) = \sum_{y_u,y_v=1}^{L} p^{(0)}(z_u,z_v,a_{uv},y_u,y_v)\left[(1-a_{uv})(W_0)_{y_uy_v}+a_{uv}(W_1)_{y_uy_v}\right].$$

For the true importance weights $W^\star_{\mathrm{iw}}$, write $W^\star_{\mathrm{iw}} = (W^\star_{\mathrm{iw},0}, W^\star_{\mathrm{iw},1})$ with

$$\begin{aligned}(W^\star_{\mathrm{iw},0})_{y_uy_v} &= p^{(1)}(a_{uv}=0,y_u,y_v)/p^{(0)}(a_{uv}=0,y_u,y_v),\\ (W^\star_{\mathrm{iw},1})_{y_uy_v} &= p^{(1)}(a_{uv}=1,y_u,y_v)/p^{(0)}(a_{uv}=1,y_u,y_v).\end{aligned}$$

Then it is straightforward to check that $W^\star_{\mathrm{iw}} \in \mathcal{W}_{\mathrm{iw}}$, and under CSBM, we have

$$\begin{aligned}&p^{\mathrm{iw}}_{W^\star_{\mathrm{iw}}}(z_u,z_v,a_{uv})\\ &= \sum_{y_u,y_v=1}^{L} p(z_u,z_v \mid y_u,y_v)p^{(0)}(a_{uv},y_u,y_v)\left[(1-a_{uv})(W^\star_{\mathrm{iw},0})_{y_uy_v}+a_{uv}(W^\star_{\mathrm{iw},1})_{y_uy_v}\right]\\ &= \sum_{y_u,y_v=1}^{L} p(z_u,z_v \mid y_u,y_v)p^{(1)}(a_{uv},y_u,y_v) = p^{(1)}(z_u,z_v,a_{uv}).\end{aligned}$$

Therefore, this matching equation provides a direct way to estimate the importance weights when both CSS and label shift occur simultaneously. In principle, one can develop the estimation procedure based on minimizing KL-divergence analogous to PLSA (i.e., minimizing $D_{\mathrm{KL}}\left(p^{(1)} \,\|\, p_W^{\mathrm{iw}}\right)$). Theoretical properties in finite-samples can also be analyzed with proofs similar to those used for the finite-sample PLSA estimator in Section 5, albeit with more involved notation. In this paper, we instead focus on the formulation (6), which estimates the target connection probabilities $W^\star$ under CSS (Assumption 3.1); this allows for simplifying the presentation while conveying the core idea of graph structure shift estimation in GDA.

# C Experimental details

## C.1 Datasets

**CSBM** We gives details of the simulated data generated from CSBM described in Section 6. In both settings, we generate two independent source data from CSBM where one is used for training the source predictor and the other is used for calibrating the predictor via BCTS. We additionally generate target data from CSBM. The node attributes are 20-dimensional Gaussian features generated from $\mathcal{N}(\mu_y, \sigma^2\mathbb{I})$, where for each class $y$, $\mu_y \sim \mathcal{N}(0, \mathbb{I}/20)$ and $\sigma = 1$.

In the first setting, we generate 5000 nodes for each of the source training, source calibration, and target data under the binary class case with a uniform class prior. Let $Q^{(0)} = (q^{(0)}(y, y'))_{y,y' \in [L]} \in \mathbb{R}^{L \times L}$ and $Q^{(1)} = (q^{(1)}(y, y'))_{y,y' \in [L]} \in \mathbb{R}^{L \times L}$ denote the source and target connection probability matrices. For the source, we fix

$$Q^{(0)} = \begin{pmatrix} 0.02 & 0.005 \\ 0.005 & 0.02 \end{pmatrix}.$$

In order to introduce CSS, we vary the entries of the target connection matrix

$$Q^{(1)} = \begin{pmatrix} p_1 & q \\ q & p_2 \end{pmatrix},$$

where $q \in \{0.0025, 0.0075\}$, $p_2 \in \{0.01, 0.015\}$, and $p_1 \in \{0.01, 0.015, 0.02, 0.025, 0.03\}$. The results are shown in Figure 1.

In the second setting, we consider both binary and three-class cases with uniform class priors. For the binary case, we fix

$$Q^{(0)} = \begin{pmatrix} 0.02 & 0.005 \\ 0.005 & 0.02 \end{pmatrix}, \quad Q^{(1)} = \begin{pmatrix} 0.03 & 0.0075 \\ 0.0075 & 0.01 \end{pmatrix}.$$

For the three-class case, we set

$$Q^{(0)} = \begin{pmatrix} 0.02 & 0.005 & 0.005 \\ 0.005 & 0.02 & 0.005 \\ 0.005 & 0.005 & 0.02 \end{pmatrix}, \quad Q^{(1)} = \begin{pmatrix} 0.03 & 0.005 & 0.0025 \\ 0.005 & 0.015 & 0.001 \\ 0.0025 & 0.001 & 0.01 \end{pmatrix}.$$

We then vary the number of nodes in both the source and target graphs simultaneously with $n^{(0)}, n^{(1)} \in \{1000, 1250, \ldots, 10000\}$, and present the results in Figure 2.

**Airport** The Airport dataset[1] is a real-world graph dataset consisting of three domains: Brazil, Europe, and the USA. In each domain, nodes represent airports and edges represent flight connections. Labels categorize airports into four classes according to their activity level, typically measured by the number of flights or passenger throughput. Since the dataset does not contain node features, we generate 32-dimensional features for each node from a Gaussian distribution $\mathcal{N}(\mu_y \mathbf{1}, \sigma^2 I)$, where $\mu_1 = -0.5, \mu_2 = 0, \mu_3 = 0.5, \mu_4 = 1$, and $\sigma = 1$. Here $\mathbf{1}$ denotes the all-ones vector, so each mean $\mu_y$ corresponds to a constant vector.

The dataset statistics of the Aiport dataset are as follows:

- Brazil: 131 nodes, 2,148 edges

[1] https://github.com/GentleZhu/EGI/tree/main/data

- Europe: 399 nodes, 11,990 edges
- USA: 1,190 nodes, 27,198 edges

We refer the reader to Zhu et al. (2021b); Liu et al. (2024b) for further details on the Airport dataset.

### C.2 Training details and network architectures

Both PLSA and Pair-Align need a source predictor (or classifier) for their implementation. In all experiments, we use a two-layer MLP with ReLU activations and 32 hidden units. The model is trained on the source data (without the source graph) for 60 epochs using the Adam optimizer with a learning rate of $10^{-3}$. The batch size is set to 256 for CSBM and 64 for Airport. After training, we apply BCTS calibration on a held-out calibration set to improve the calibration of predictor.

In the Airport experiment, once the importance weights are estimated, we reweight the source graph following the method in Section 4.3 and train GNNs on the adjusted graph to obtain models for target label prediction. For the GNN architecture, we use a two-layer Graph Convolutional Network (GCN) followed by a linear classifier, with 32 hidden units. The GNN is trained for 300 epochs using the Adam optimizer with a learning rate of $5 \times 10^{-3}$.

## D Additional experimental results

This section gives additional results for CSBM by varying the target connection probabilities under imbalanced class priors. The setting is identical to the first experiment of CSBM in Section 6, except that the class prior distribution is changed from uniform to imbalanced, with $p^{(0)}(y=1) = p^{(1)}(y=1) = 0.2$ and $p^{(0)}(y=2) = p^{(1)}(y=2) = 0.8$. Figure 3 shows the results. The overall trend of the estimation error is similar to that in Figure 1, that is, PLSA consistently outperforms Pair-Align, and the latter performs poorly especially when the within-class connection probability is small (i.e., when the graph is sparse). In contrast, PLSA maintains strong performance across different within-class probabilities.

Compared to Figure 1, a notable difference is that under imbalanced class priors, the performance generally degrades with higher standard deviations. In particular, in panels (a) and (b), the one standard deviation error bands of PLSA and Pair-Align overlap in most regions. These findings suggest that balanced class priors yield more stable and reliable performance compared to imbalanced ones.

## E Concentration inequalities for U-statistics under CSBM

We present matrix concentration inequalities for bounded U-statistics under CSBM data generating process. Throughout we work on a probability space that supports an infinite sequence of labels $(y_u)_{u\in\mathbb{N}}$, features $(x_u)_{u\in\mathbb{N}}$, and an infinite upper-triangular array of edges $(a_{uv})_{u<v}$ with the CSBM structure

$$
\begin{aligned}
&y_u \overset{\text{i.i.d.}}{\sim} p(y),\\
&x_u \mid y_u = y \sim p(x \mid y), \qquad x_u \perp\!\!\!\perp \{x_v, y_v, a_{vw}\}_{v\neq u,w} \;\big|\; y_u,\\
&a_{uv} \mid (y_u = y, y_v = y') \sim \mathrm{Ber}\big(q(y,y')\big), \qquad a_{uv} \perp\!\!\!\perp \{a_{u'v'} : (u',v') \neq (u,v)\} \;\big|\; (y_u, y_v),
\end{aligned}
$$

with $a_{vu} = a_{uv}$ and $a_{uu} = 0$. Existence of such probability space follows from Kolmogorov's extension theorem.

Given this setup, for each $n \in \mathbb{N}$, the data we observe is the subset with first $n$ nodes,

$$\{(x_u, y_u)_{1\le u\le n},\ (a_{uv})_{1\le u<v\le n}\},$$

which follows the distribution specified by the CSBM with $n$ nodes.

Let $P$ denote the joint distribution of $(x_u, x_v, a_{uv})$ under CSBM, and let $P_n$ be the empirical measure based on all node pairs $(x_u, x_v, a_{uv})_{1\le u<v\le n}$. For a matrix-valued measurable function $H : \mathcal{X}\times\mathcal{X}\times\{0,1\} \to \mathbb{S}^d$ (where $\mathbb{S}^d := \{B \in \mathbb{R}^{d\times d} : B^\top = B\}$ is the set of symmetric matrices),

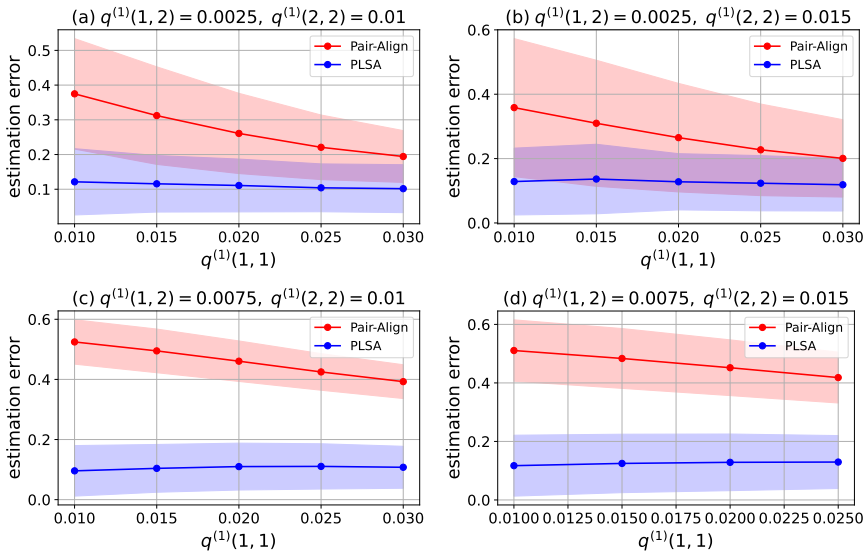


Figure 3: Estimation error of the importance weights as the target connection probability varies under CSBM with binary classes and an imbalanced class prior ($p^{(0)}(y=1) = p^{(1)}(y=1) = 0.2$, $p^{(0)}(y=2) = p^{(1)}(y=2) = 0.8$). The results are averaged over 10 trials.

define

$$P(H) := \int H(x,x',a)\, dP(x,x',a), \text{ and}$$

$$P_n(H) := \int H(x,x',a)\, dP_n(x,x',a) = \frac{1}{\binom{n}{2}} \sum_{1\le u<v\le n} H(x_u,x_v,a_{uv}).$$

The following lemma controls the deviation of $P_n(H)$ from $P(H)$ when $H$ is $P$-a.e. bounded.

**Lemma E.1** *Let $H : \mathcal{X}\times\mathcal{X}\times\{0,1\} \to \mathbb{S}^d$ be symmetric in its first two arguments and $P$-a.e. bounded, i.e., $H(x,x',a) = H(x',x,a)$ and $\|H(x,x',a)\| \le M$ for $P$-a.e. $(x,x',a)$. Then for $\delta \in (0,1)$, if $n \ge \max\{\log(3d), \log(8/\delta)\}$, with probability at least $1-\delta$,*

$$\big\|P_n(H) - P(H)\big\| \le c\, M \left( \sqrt{\frac{\log(8d/\delta)}{n}} + \frac{\log(3d)\ \log(8/\delta)}{n} \right),$$

*where $c > 0$ is a universal constant.*

The proof of Lemma E.1 is given in Appendix H.1. The standard U-statistics setting takes $x_u, x_v$ i.i.d., whereas $P_n(H)$ couples $(x_u, x_v)$ with $a_{uv}$, so existing matrix-valued concentration bounds, e.g., Minsker & Wei (2019), do not directly apply. We therefore use Hoeffding's decomposition and decouple the dependence between $(x_u, x_v)$ and $a_{uv}$ by conditioning on the labels.

A useful specialization of the lemma is when $H$ has the block form

$$H(x,x',a) = \begin{bmatrix} 0 & G(x,x',a)^\top \\ G(x,x',a) & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)},$$

for some vector-valued measurable $G : \mathcal{X}\times\mathcal{X}\times\{0,1\} \to \mathbb{R}^d$ that is symmetric and $P$-a.e. bounded, i.e., $G(x,x',a) = G(x',x,a)$ and $\|G(x,x',a)\|_2 \le M$ for $P$-a.e. $(x,x',a)$. In this case $\|H\| = \|G\|_2$, and applying Lemma E.1 to $H$ gives the following corollary (with $d+1$ in place of $d$).

**Corollary E.1** *Under the conditions stated above on $G$, for $\delta \in (0,1)$, if $n \ge \max\{\log(3(d+1)), \log(8/\delta)\}$, then with probability at least $1-\delta$,*

$$\big\|P_n(G) - P(G)\big\|_2 \le c\, M \left( \sqrt{\frac{\log(8(d+1)/\delta)}{n}} + \frac{\log(3(d+1))\ \log(8/\delta)}{n} \right),$$

| Notation | Definition |
|---|---|
| $[n]$ | $\{1, \dots, n\}$ for $n \in \mathbb{N}$ |
| $\|x\|_1$ | $\ell_1$ norm of vector $x$ |
| $\|x\|_2$ | $\ell_2$ norm of vector $x$ |
| $\|x\|_\infty$ | $\ell_\infty$ norm of vector $x$ |
| $\mathbb{S}^d$ | set of real $d \times d$ symmetric matrices |
| $\lambda_{\min}(A)$ | minimum eigenvalue of symmetric matrix $A$ |
| $\lambda_{\max}(A)$ | maximum eigenvalue of symmetric matrix $A$ |
| $\|A\|$ | spectral norm/operator norm (=maximum singular value of $A$) |
| $\mathrm{diag}(A)$ | diagonal entries of matrix $A$ |
| $\mathrm{tr}(A)$ | trace of matrix $A$ |
| $\mathrm{vec}(A)$ | vectorization of matrix $A$ |
| $\mathrm{vech}(A)$ | half-vectorization of symmetric matrix $A$ |
| $A \succeq B$ | the matrix $A - B$ is positive semidefinite |
| $A \preceq B$ | the matrix $B - A$ is positive semidefinite |
| $A \otimes B$ | kronecker product between matrix $A$ and $B$ |
| $c, c', c'', \dots$ or $c_1, c_2, \dots$ | universal constants (whose definitions may change from one result to another) |

Table 2: Notation used throughout the proofs.

*where $c > 0$ is a universal constant.*

The proof of Corollary E.1 is immediate from Lemma E.1 and is therefore omitted.

# F Proof of theorems

## F.1 Gradients and Hessians of the pairwise log-likelihood

Before proving the main theorems, we provide a brief derivation of the gradient and Hessian of the pairwise log-likelihood. Recall

$$S_{\mathrm{f}}(W; x, x', a) := f(x)^\top ((1-a)(1-W) + aW) f(x'),$$

so that

$$W_{\mathrm{f}} = \arg\max_{W \in \mathcal{W}} \mathbb{E}\Big[ \log S_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)}) \Big] =: \ell_{\mathrm{f}}(W),$$

$$\widehat{W}_{\mathrm{f}} = \arg\max_{W \in \mathcal{W}} \frac{1}{\binom{n^{(1)}}{2}} \sum_{u<v} \log S_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)}) =: \widehat{\ell}_{\mathrm{f}}(W).$$

For any $W \in \mathbb{S}^L$, let $\mathrm{vec}(W) \in \mathbb{R}^{L^2}$ denote its vectorization that stacks columns of $W$. Writing

$$Z_{uv} := \mathrm{vec}\big( f(x_u^{(1)}) f(x_v^{(1)})^\top \big),$$

the gradient and the Hessian of $\ell_{\mathrm{f}}(W)$ are calculated as

$$\nabla_{\mathrm{vec}(W)} \ell_{\mathrm{f}}(W) = \mathbb{E}\left[ \frac{(2a_{uv}^{(1)} - 1)\, Z_{uv}}{S_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})} \right],$$

$$\nabla^2_{\mathrm{vec}(W)} \ell_{\mathrm{f}}(W) = -\mathbb{E}\left[ \frac{Z_{uv} Z_{uv}^\top}{S_{\mathrm{f}}^2(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})} \right],$$

where we use $(2a-1)^2 = 1$ for $a \in \{0, 1\}$. Similarly, for the empirical loss,

$$\nabla_{\mathrm{vec}(W)} \widehat{\ell}_{\mathrm{f}}(W) = \frac{1}{\binom{n^{(1)}}{2}} \sum_{u<v} \frac{(2a_{uv}^{(1)} - 1)\, Z_{uv}}{S_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})},$$

$$\nabla^2_{\mathrm{vec}(W)} \widehat{\ell}_{\mathrm{f}}(W) = -\frac{1}{\binom{n^{(1)}}{2}} \sum_{u<v} \frac{Z_{uv} Z_{uv}^\top}{S_{\mathrm{f}}^2(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})}.$$

Since $W$ is symmetric, it is more natural to parameterize it with the half-vectorization operator (Magnus & Neudecker, 2019, Chapter 3.8). Let $\mathrm{vech}(W) \in \mathbb{R}^{L(L+1)/2}$ denote the vector by stacking the upper-triangular entries of $W$. Let $D_L \in \mathbb{R}^{L^2 \times L(L+1)/2}$ be the duplication matrix (Magnus & Neudecker, 2019, Equation (22)) so that

$$\mathrm{vec}(W) = D_L \mathrm{vech}(W), \text{ and } \mathrm{vech}(W) = D_L^\dagger \mathrm{vec}(W), \tag{7}$$

where $D_L^\dagger = (D_L^\top D_L)^{-1} D_L^\top$ is the Moore-Penrose inverse of $D_L$. Using the relation (7) and the chain rule, the derivatives with respect to $\mathrm{vech}(W)$ become

$$\nabla_{\mathrm{vech}(W)} \ell_{\mathrm{f}}(W) = D_L^\top \nabla_{\mathrm{vec}(W)} \ell_{\mathrm{f}}(W), \quad \nabla^2_{\mathrm{vech}(W)} \ell_{\mathrm{f}}(W) = D_L^\top \nabla^2_{\mathrm{vec}(W)} \ell_{\mathrm{f}}(W) D_L,$$

and

$$\nabla_{\mathrm{vech}(W)} \widehat{\ell}_{\mathrm{f}}(W) = D_L^\top \nabla_{\mathrm{vec}(W)} \widehat{\ell}_{\mathrm{f}}(W), \quad \nabla^2_{\mathrm{vech}(W)} \widehat{\ell}_{\mathrm{f}}(W) = D_L^\top \nabla^2_{\mathrm{vec}(W)} \widehat{\ell}_{\mathrm{f}}(W) D_L.$$

Combining with the expressions above for the derivatives with respect to $\mathrm{vec}(W)$, we obtain closed form expressions for the gradient and Hessian with respect to $\mathrm{vech}(W)$.

### F.2 Proof of Theorem 5.2

Our proof outline closely follows Garg et al. (2020, Lemma 3), where the main difference is from the concentration inequalities we apply. Classical tools such as Hoeffding's inequality or results from standard random matrix theory assume i.i.d. samples and thus do not directly apply to our setting. Instead we use the concentration bounds specific to CSBM that follow from Lemma E.1 and Corollary E.1.

For notational convenience, write $w_{\mathrm{f}} = \mathrm{vech}(W_{\mathrm{f}})$ and $\widehat{w}_{\mathrm{f}} = \mathrm{vech}(\widehat{W}_{\mathrm{f}})$. We also abuse notation and write

$$\begin{aligned} \ell(w_{\mathrm{f}}) &:= \ell_{\mathrm{f}}(W_{\mathrm{f}}), \quad \ell(\widehat{w}_{\mathrm{f}}) := \ell_{\mathrm{f}}(\widehat{W}_{\mathrm{f}}), \text{ and} \\ \widehat{\ell}(w_{\mathrm{f}}) &:= \widehat{\ell}_{\mathrm{f}}(W_{\mathrm{f}}), \quad \widehat{\ell}(\widehat{w}_{\mathrm{f}}) := \widehat{\ell}_{\mathrm{f}}(\widehat{W}_{\mathrm{f}}), \end{aligned}$$

where we suppress the explicit dependence of $\ell$ and $\widehat{\ell}$ on $f$.

Since $\widehat{w}_{\mathrm{f}}$ maximizes $\widehat{\ell}$ over $\mathcal{W}$, a Taylor expansion gives, for some $t \in (0,1)$,

$$\begin{aligned} 0 &\le \widehat{\ell}(\widehat{w}_{\mathrm{f}}) - \widehat{\ell}(w_{\mathrm{f}}) \\ &= \langle \nabla \widehat{\ell}(w_{\mathrm{f}}), \widehat{w}_{\mathrm{f}} - w_{\mathrm{f}} \rangle + \frac{1}{2} (\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}})^\top \nabla^2 \widehat{\ell}((1-t)\widehat{w}_{\mathrm{f}} + t w_{\mathrm{f}})(\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}). \end{aligned} \tag{8}$$

Observe that for any $W \in \mathcal{W}$, we have

$$\begin{aligned} S_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)}) &= f(x_u^{(1)})^\top ((1 - a_{uv}^{(1)})(1 - W) + a_{uv}^{(1)} W) f(x_v^{(1)}) \\ &= \mathrm{tr}\big(((1 - a_{uv}^{(1)})(1 - W) + a_{uv}^{(1)} W) f(x_v^{(1)}) f(x_u^{(1)})^\top\big) \\ &= \langle \mathrm{vec}((1 - a_{uv}^{(1)})(1 - W) + a_{uv}^{(1)} W), \mathrm{vec}(f(x_u^{(1)}) f(x_v^{(1)})^\top) \rangle. \end{aligned} \tag{9}$$

Applying Hölder's inequality, it follows

$$\begin{aligned} S^2_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)}) &\le \Big\| \mathrm{vec}((1 - a_{uv}^{(1)})(1 - W) + a_{uv}^{(1)} W) \Big\|_\infty^2 \Big\| \mathrm{vec}(f(x_u^{(1)}) f(x_v^{(1)})^\top) \Big\|_1^2 \\ &\le \max\{ \|\mathrm{vec}(W)\|_\infty^2, \|1 - \mathrm{vec}(W)\|_\infty^2 \} \cdot \Big\| f(x_u^{(1)}) \Big\|_1^2 \Big\| f(x_v^{(1)}) \Big\|_1^2 \\ &\le \Big\| f(x_u^{(1)}) \Big\|_1^2 \Big\| f(x_v^{(1)}) \Big\|_1^2 = 1, \end{aligned} \tag{10}$$

where in the second step we use $\|\mathrm{vec}(ab^\top)\|_1 = \|a\|_1 \|b\|_1$; in the third step, $0 \le W_{yy'} \le 1$ for $W \in \mathcal{W}$; and in the last step, we use $f(x) \in \Delta^{L-1}$ for all $x$, so $\|f(x)\|_1 = 1$.

Further, by Assumption 5.1, $S_{\mathrm{f}}^2(W_{\mathrm{f}}; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)}) \geq \tau_{\min}^2$. Combining with (10) yields

$$\begin{aligned}
-\nabla^2\widehat{\ell}((1-t)\widehat{w}_{\mathrm{f}} + tw_{\mathrm{f}}) &= \frac{1}{\binom{n^{(1)}}{2}} \sum_{u<v} D_L^\top \frac{Z_{uv}Z_{uv}^\top}{S_{\mathrm{f}}^2((1-t)\widehat{W}_{\mathrm{f}} + tW_{\mathrm{f}}; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})} D_L \\
&\succeq \frac{1}{\binom{n^{(1)}}{2}} \sum_{u<v} D_L^\top Z_{uv}Z_{uv}^\top D_L \\
&\succeq \frac{\tau_{\min}^2}{\binom{n^{(1)}}{2}} \sum_{u<v} D_L^\top \frac{Z_{uv}Z_{uv}^\top}{S_{\mathrm{f}}^2(W_{\mathrm{f}}; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})} D_L \\
&= -\tau_{\min}^2 \nabla^2\widehat{\ell}(w_{\mathrm{f}}). \qquad (11)
\end{aligned}$$

Substituting (11) into (8) and rearranging,

$$\langle \nabla\widehat{\ell}(w_{\mathrm{f}}), \widehat{w}_{\mathrm{f}} - w_{\mathrm{f}} \rangle \geq -\frac{\tau_{\min}^2}{2}(\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}})^\top \nabla^2\widehat{\ell}(w_{\mathrm{f}})(\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}).$$

Since $W_{\mathrm{f}}$ maximizes $\ell$ over $\mathcal{W}$ and $\mathcal{W}$ is convex, the first-order optimality condition gives

$$\langle \widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}, \nabla\ell(w_{\mathrm{f}}) \rangle \leq 0.$$

Combining with the inequality above, we obtain

$$\langle \nabla\widehat{\ell}(w_{\mathrm{f}}) - \nabla\ell(w_{\mathrm{f}}), \widehat{w}_{\mathrm{f}} - w_{\mathrm{f}} \rangle \geq -\frac{\tau_{\min}^2}{2}(\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}})^\top \nabla^2\widehat{\ell}(w_{\mathrm{f}})(\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}).$$

Applying the Cauchy–Schwarz inequality to the left-hand side,

$$\|\nabla\widehat{\ell}(w_{\mathrm{f}}) - \nabla\ell(w_{\mathrm{f}})\|_2 \|\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}\|_2 \geq -\frac{\tau_{\min}^2}{2}(\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}})^\top \nabla^2\widehat{\ell}(w_{\mathrm{f}})(\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}). \qquad (12)$$

We use the following concentration bounds whose proofs are deferred to Appendix H.

**Lemma F.1** *Suppose that $f$ satisfies Assumption 5.1. There exists a universal constant $c > 0$ such that for $\delta \in (0,1)$, if $n^{(1)} \geq (\log(3L^2))^2 \log(8/\delta)$, with probability at least $1-\delta$,*

$$\lambda_{\min}(-\nabla^2\widehat{\ell}(w_{\mathrm{f}})) \geq \lambda_{\min}(-\nabla^2\ell(w_{\mathrm{f}})) - c\tau_{\min}^{-2}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}}.$$

**Lemma F.2** *Suppose that $f$ satisfies Assumption 5.1. There exists a universal constant $c > 0$ such that for $\delta \in (0,1)$, if $n^{(1)} \geq (\log(3L^2))^2 \log(8/\delta)$, with probability at least $1-\delta$,*

$$\left\|\nabla\widehat{\ell}(w_{\mathrm{f}}) - \nabla\ell(w_{\mathrm{f}})\right\|_2 \leq c\tau_{\min}^{-1}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}}.$$

Applying Lemma F.1, Lemma F.2, and a union bound, with probability at least $1-2\delta$,

$$c_2\tau_{\min}^{-1}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}}\,\|\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}\|_2 \geq \frac{\tau_{\min}^2}{2}\left(\lambda_{\min,\mathrm{f}} - c_1\tau_{\min}^{-2}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}}\right)\|\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}\|_2^2,$$

for some universal constants $c_1, c_2 > 0$. Since $n^{(1)} \geq \frac{4c_1^2\tau_{\min}^{-4}\log(8L^2/\delta)}{(\lambda_{\min,\mathrm{f}})^2}$ by assumption of the theorem, rearranging and simplifying the above inequality yields

$$\|\widehat{w}_{\mathrm{f}} - w_{\mathrm{f}}\|_2 \leq \frac{4c_2\tau_{\min}^{-3}}{\lambda_{\min,\mathrm{f}}}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}},$$

which completes the proof of the theorem.

### F.3 Proof of Theorem 5.3

We adapt the proof of Garg et al. (2020, Lemma 4) to our setting. For notational convenience, write $w_{\mathrm{f}} = \mathrm{vech}(W_{\mathrm{f}})$ and $w^\star = \mathrm{vech}(W^\star)$. We also abuse notation and write

$$\ell(w_{\mathrm{f}}) := \ell_{\mathrm{f}}(W_{\mathrm{f}}), \quad \ell(w^\star) := \ell_{\mathrm{f}}(W^\star), \text{ and } \ell^\star(w^\star) := \ell_{\mathrm{f}^\star}(W^\star).$$

Taylor expansion gives, for some $t \in (0,1)$,

$$\ell(w_{\mathrm{f}}) = \ell(w^\star) + \langle \nabla \ell(w^\star), w_{\mathrm{f}} - w^\star \rangle + \frac{1}{2}(w_{\mathrm{f}} - w^\star)^\top \nabla^2 \ell((1-t)w_{\mathrm{f}} + t w^\star)(w_{\mathrm{f}} - w^\star).$$

Observe that under Assumption 5.1, the argument used in (11) can be applied to yield

$$-\nabla^2 \ell((1-t)w_{\mathrm{f}} + t w^\star) \succeq -\tau_{\min}^2 \nabla^2 \ell(w_{\mathrm{f}}).$$

Plugging this into the above inequality, we have

$$\ell(w_{\mathrm{f}}) \le \ell(w^\star) + \langle \nabla \ell(w^\star), w_{\mathrm{f}} - w^\star \rangle + \frac{\tau_{\min}^2}{2}(w_{\mathrm{f}} - w^\star)^\top \nabla^2 \ell(w_{\mathrm{f}})(w_{\mathrm{f}} - w^\star).$$

Since $\ell(w_{\mathrm{f}}) \ge \ell(w^\star)$ by optimality and $w^\top \nabla^2 \ell(w_{\mathrm{f}}) w \le -\lambda_{\min,\mathrm{f}} \|w\|_2^2$ for all vectors $w$, it follows

$$0 \le \langle \nabla \ell(w^\star), w_{\mathrm{f}} - w^\star \rangle - \frac{\tau_{\min}^2 \lambda_{\min,\mathrm{f}}}{2} \|w_{\mathrm{f}} - w^\star\|_2^2.$$

Because $W^\star$ maximizes $\ell^\star$ over convex $\mathcal{W}$, the first-order optimality condition gives

$$\langle w_{\mathrm{f}} - w^\star, \nabla \ell^\star(w^\star) \rangle \le 0.$$

Combining with the inequality above and rearranging,

$$\langle \nabla \ell(w^\star) - \nabla \ell^\star(w^\star), w_{\mathrm{f}} - w^\star \rangle \ge \frac{\tau_{\min}^2 \lambda_{\min,\mathrm{f}}}{2} \|w_{\mathrm{f}} - w^\star\|_2^2.$$

Applying Cauchy-Schwarz inequality and simplifying, we get

$$\|w_{\mathrm{f}} - w^\star\|_2 \le \frac{2\tau_{\min}^{-2}}{\lambda_{\min,\mathrm{f}}} \left\| \nabla \ell(w^\star) - \nabla \ell^\star(w^\star) \right\|_2. \tag{13}$$

It remains to bound $\left\| \nabla \ell(w^\star) - \nabla \ell^\star(w^\star) \right\|_2$. Using the closed form gradient from Section F.1,

$$\begin{aligned}
&\left\| \nabla \ell(w^\star) - \nabla \ell^\star(w^\star) \right\|_2 \\
&= \left\| D_L^\top \mathbb{E}\left[ \frac{(2a_{uv}^{(1)} - 1)\mathrm{vec}(f(x_u^{(1)}) f(x_v^{(1)})^\top)}{S_{\mathrm{f}}(W^\star; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})} - \frac{(2a_{uv}^{(1)} - 1)\mathrm{vec}(f^\star(x_u^{(1)}) f^\star(x_v^{(1)})^\top)}{S_{\mathrm{f}^\star}(W^\star; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})} \right] \right\|_2 \\
&\le 2 \left\| \mathbb{E}\left[ \frac{\mathrm{vec}(f(x_u^{(1)}) f(x_v^{(1)})^\top)}{S_{\mathrm{f}}(W^\star; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})} - \frac{\mathrm{vec}(f^\star(x_u^{(1)}) f^\star(x_v^{(1)})^\top)}{S_{\mathrm{f}^\star}(W^\star; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})} \right] \right\|_2,
\end{aligned} \tag{14}$$

where in the last step we use the fact that $\|D_L\| \le 2$ and $|2a - 1| \le 1$ for $a \in \{0, 1\}$.

For shorthand, write $S_{\mathrm{f}} := S_{\mathrm{f}}(W^\star; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})$ and $S_{\mathrm{f}^\star} := S_{\mathrm{f}^\star}(W^\star; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})$. By Assumption 5.1, both satisfy $S_{\mathrm{f}} \ge \tau_{\min}$ and $S_{\mathrm{f}^\star} \ge \tau_{\min}$, hence

$$\begin{aligned}
&\left\| \mathbb{E}\left[ \frac{\mathrm{vec}(f(x_u^{(1)}) f(x_v^{(1)})^\top)}{S_{\mathrm{f}}} - \frac{\mathrm{vec}(f^\star(x_u^{(1)}) f^\star(x_v^{(1)})^\top)}{S_{\mathrm{f}^\star}} \right] \right\|_2 \\
&= \left\| \mathbb{E}\left[ \frac{\mathrm{vec}(f(x_u^{(1)}) f(x_v^{(1)})^\top) S_{\mathrm{f}^\star} - \mathrm{vec}(f^\star(x_u^{(1)}) f^\star(x_v^{(1)})^\top) S_{\mathrm{f}}}{S_{\mathrm{f}} S_{\mathrm{f}^\star}} \right] \right\|_2 \\
&\le \tau_{\min}^{-2} \left\| \mathbb{E}\left[ \mathrm{vec}(f(x_u^{(1)}) f(x_v^{(1)})^\top)(S_{\mathrm{f}^\star} - S_{\mathrm{f}}) + (\mathrm{vec}(f(x_u^{(1)}) f(x_v^{(1)})^\top) - \mathrm{vec}(f^\star(x_u^{(1)}) f^\star(x_v^{(1)})^\top)) S_{\mathrm{f}} \right] \right\|_2.
\end{aligned} \tag{15}$$

From equation (10), $\|\text{vec}(f(x)f(x)^\top)\|_2 \le 1$, and so

$$\begin{aligned}
&\left\|\mathbb{E}\left[\text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)(S_{\mathrm{f}^\star} - S_{\mathrm{f}})\right]\right\|_2 \le \mathbb{E}\left[|S_{\mathrm{f}^\star} - S_{\mathrm{f}}|\right] \\
&= \mathbb{E}\left[\left|\langle \text{vec}((1-a_{uv}^{(1)})(1-W^\star) + a_{uv}^{(1)}W^\star), \text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top) - \text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)\rangle\right|\right] \\
&\le \mathbb{E}\left[\left\|\text{vec}((1-a_{uv}^{(1)})(1-W^\star) + a_{uv}^{(1)}W^\star)\right\|_\infty \left\|\text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top) - \text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)\right\|_1\right] \\
&\le \mathbb{E}\left[\left\|\text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top) - \text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)\right\|_1\right],
\end{aligned}$$

where the first step applies Jensen's inequality, the second step follows from (9), and the third step applies Hölder's inequality. Additionally, since $S_{\mathrm{f}}, S_{\mathrm{f}^\star} \le 1$ by equation (10) and Jensen's inequality,

$$\begin{aligned}
&\left\|\mathbb{E}\left[(\text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top) - \text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top))S_{\mathrm{f}}\right]\right\|_2 \\
&\le \mathbb{E}\left[|S_{\mathrm{f}}| \cdot \left\|\text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top) - \text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top)\right\|_2\right] \\
&\le \mathbb{E}\left[\left\|\text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top) - \text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top)\right\|_1\right].
\end{aligned}$$

Plugging these into (15) and using triangle inequality,

$$\begin{aligned}
&\left\|\mathbb{E}\left[\frac{\text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)}{S_{\mathrm{f}}} - \frac{\text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top)}{S_{\mathrm{f}^\star}}\right]\right\|_2 \\
&\qquad \le 2\tau_{\min}^{-2}\mathbb{E}\left[\left\|\text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top) - \text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top)\right\|_1\right]. \quad (16)
\end{aligned}$$

To control the right-hand side term, we invoke the followng lemma.

**Lemma F.3** *For any vectors $w_1, w_2, w_1', w_2' \in \mathbb{R}^n$, we have*

$$\|\text{vec}(w_1 w_2^\top) - \text{vec}(w_1' {w_2'}^\top)\|_1 \le \|w_1\|_1\|w_2 - w_2'\|_1 + \|w_2'\|_1\|w_1 - w_1'\|_1.$$

The proof is given in Appendix H.4.

Applying Lemma F.3 and noting that $\|f(x)\|_1 = 1$ and $\|f^\star(x)\|_1 = 1$, it follows

$$\left\|\mathbb{E}\left[\frac{\text{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)}{S_{\mathrm{f}}} - \frac{\text{vec}(f^\star(x_u^{(1)})f^\star(x_v^{(1)})^\top)}{S_{\mathrm{f}^\star}}\right]\right\|_2 \le 4\tau_{\min}^{-2}\mathbb{E}\left[\|f(x_u^{(1)}) - f^\star(x_u^{(1)})\|_1\right].$$

Substituting this into (14), we get

$$\|\nabla\ell(w^\star) - \nabla\ell^\star(w^\star)\|_2 \le 8\tau_{\min}^{-2}\mathbb{E}\left[\|f(x_u^{(1)}) - f^\star(x_u^{(1)})\|_1\right].$$

Finally combining with (13) and using the fact that the node feature $x_u$ has the same distribution across source and target domains under CSS, we conclude the proof.

# G Proof of propositions

## G.1 Proof of Proposition 4.1

First, we assume $\{p(z \mid y), y = 1, \ldots, L\}$ is linearly independent. Suppose that two solutions exist, $\widetilde{W}, W^\star \in \mathcal{W}$, which both satisfy equation (2), i.e., for $W \in \{\widetilde{W}, W^\star\}$,

$$p^{(1)}(z_u, z_v, a_{uv}) = p_W(z_u, z_v, a_{uv}) \text{ for all } (z_u, z_v, a_{uv}) \in \mathcal{Z} \times \mathcal{Z} \times \{0, 1\}.$$

Setting the two expressions for $p_{\widetilde{W}}$ and $p_{W^\star}$ to be equal, we get

$$\begin{aligned}
0 &= \sum_{y_u, y_v = 1}^{L} p(z_u, z_v, y_u, y_v)[(2a_{uv} - 1)(\widetilde{W}_{y_u y_v} - W^\star_{y_u y_v})] \\
&= \sum_{y_u, y_v = 1}^{L} p(z_u \mid y_u)p(z_v \mid y_v)p(y_u)p(y_v)[(2a_{uv} - 1)(\widetilde{W}_{y_u y_v} - W^\star_{y_u y_v})],
\end{aligned}$$

where the second step follows since the pairs $(z_u, y_u), (z_v, y_v)$ are independent. Since $\{p(z \mid y), y = 1, \ldots, L\}$ is linearly independent, the set of product densities $\{p(z_u \mid y_u)p(z_v \mid y_v), y_u, y_v = 1, \ldots, L\}$ is also linearly independent. This implies

$$p(y_u)p(y_v)[(2a_{uv} - 1)(\widetilde{W}_{y_u y_v} - W^\star_{y_u y_v})] = 0 \text{ for all } a_{uv} \in \{0, 1\}, y_u, y_v \in \mathcal{Y}.$$

Since $p(y) > 0$ for all $y \in \mathcal{Y}$ by assumption, it follows that $\widetilde{W}_{y_u y_v} = W^\star_{y_u y_v}$ for all $(y_u, y_v)$, which concludes $\widetilde{W} = W^\star$.

Next, we show that if $\{p(z \mid y), y = 1, \ldots, L\}$ is linearly dependent, there exists a solution $\widetilde{W} \neq W^\star \in \mathcal{W}$ that satisfies equation (2) and therefore the solution is not unique. To see this, the linear dependence of $\{p(z \mid y), y = 1, \ldots, L\}$ implies that there exists a nonzero vector $v = (v_1, \ldots, v_L)$ such that $\sum_{y=1}^{L} v_y p(z \mid y) = 0$ for all $z \in \mathcal{Z}$. Then we construct $\widetilde{W} = W^\star + \epsilon\Delta$ where $\epsilon > 0$ is a small constant, and $\Delta$ is a nonzero symmetric matrix defined as

$$\Delta_{y_u y_v} = \frac{v_{y_u} v_{y_v}}{p(y_u)p(y_v)} \text{ for all } y_u, y_v \in \mathcal{Y}.$$

To verify that $\widetilde{W}$ also satisfies equation (2), we need to show

$$\sum_{y_u, y_v = 1}^{L} p(z_u \mid y_u)p(z_v \mid y_v)p(y_u)p(y_v)\big(\widetilde{W}_{y_u y_v} - W^\star_{y_u y_v}\big) = 0.$$

Substituting our definitions for $\widetilde{W}$ and $\Delta$, the left-hand side becomes

$$\begin{aligned}
&\sum_{y_u, y_v = 1}^{L} p(z_u \mid y_u)p(z_v \mid y_v)p(y_u)p(y_v)\big(\widetilde{W}_{y_u y_v} - W^\star_{y_u y_v}\big) \\
&= \epsilon \sum_{y_u, y_v = 1}^{L} p(z_u \mid y_u)p(z_v \mid y_v)p(y_u)p(y_v)\Delta_{y_u y_v} \\
&= \epsilon \sum_{y_u, y_v = 1}^{L} p(z_u \mid y_u)p(z_v \mid y_v)v_{y_u} v_{y_v} = 0.
\end{aligned}$$

This proves that $p_{\widetilde{W}}(z_u, z_v, a_{uv}) = p_{W^\star}(z_u, z_v, a_{uv}) = p^{(1)}(z_u, z_v, a_{uv})$. Furthermore, by assumption of the proposition, each element of $W^\star$ is strictly between 0 and 1, i.e., $0 < W^\star_{yy'} < 1$. So we can always choose a sufficiently small non-zero $\epsilon$ such that the entries of $\widetilde{W} = W^\star + \epsilon\Delta$ remain in the interval $(0, 1)$, which ensures $\widetilde{W} \in \mathcal{W}$. This complete the proof of the proposition.

### G.2 Proof of Proposition 4.2

The proposition follows directly from the definition of canonical calibration. Since $f$ is calibrated, $f_y(x) = p(y \mid f(x))$. By substituting this into the objective (5) and applying a change of variables from $x$ to $z = f(x)$, the objective function for $W_{\mathrm{f}}$ is exactly identical to the objective for $W^\star$ given in (4), i.e.,

$$\begin{aligned}
&\mathbb{E}\Big[\log \sum_{y, y' = 1}^{L} f_y(x_u^{(1)}) f_{y'}(x_v^{(1)})[(1 - a_{uv}^{(1)})(1 - W_{yy'}) + a_{uv}^{(1)} W_{yy'}]\Big] \\
&= \mathbb{E}\Big[\log \sum_{y, y' = 1}^{L} p(y \mid z_u^{(1)})p(y' \mid z_v^{(1)})[(1 - a_{uv}^{(1)})(1 - W_{yy'}) + a_{uv}^{(1)} W_{yy'}]\Big].
\end{aligned}$$

Since Proposition 4.1 guarantees that $W^\star$ is the unique maximizer, it follows that $W_{\mathrm{f}} = W^\star$.

### G.3 Proof of Proposition 5.4

Here we prove a more general statement: if $\mathbb{E}[f(x_u^{(1)})f(x_u^{(1)})^\top] \succeq \overline{\lambda}_{\min}\mathbb{I}_L$, then for any $W \in \mathcal{W}$, we have

$$-\nabla^2_{\mathrm{vech}(W)}\ell_{\mathrm{f}}(W) \succeq \overline{\lambda}^2_{\min}\mathbb{I}_{L(L+1)/2}.$$

Recall from Section F.1 that

$$\nabla^2_{\mathrm{vec}(W)}\ell_{\mathrm{f}}(W) = -\mathbb{E}\left[\frac{Z_{uv}Z_{uv}^\top}{S^2_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)})}\right],$$

where $Z_{uv} = \mathrm{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)$. By equation (10), $S^2_{\mathrm{f}}(W; x_u^{(1)}, x_v^{(1)}, a_{uv}^{(1)}) \le 1$ for any $W \in \mathcal{W}$, so we trivially have

$$-\nabla^2_{\mathrm{vec}(W)}\ell_{\mathrm{f}}(W) \succeq \mathbb{E}\left[Z_{uv}Z_{uv}^\top\right]. \tag{17}$$

Next, for any vectors $a, b$, $\mathrm{vec}(ab^\top) = b \otimes a$. Then

$$\begin{aligned}
\mathbb{E}[Z_{uv}Z_{uv}^\top] &= \mathbb{E}[\mathrm{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)\mathrm{vec}(f(x_u^{(1)})f(x_v^{(1)})^\top)^\top] \\
&= \mathbb{E}[(f(x_v^{(1)}) \otimes f(x_u^{(1)}))(f(x_v^{(1)}) \otimes f(x_u^{(1)}))^\top] \\
&= \mathbb{E}[(f(x_v^{(1)}) \otimes f(x_u^{(1)}))(f(x_v^{(1)})^\top \otimes f(x_u^{(1)})^\top)] \\
&= \mathbb{E}[(f(x_v^{(1)})f(x_v^{(1)})^\top) \otimes (f(x_u^{(1)})f(x_u^{(1)})^\top)],
\end{aligned}$$

where we use the identity $(a \otimes b)(a \otimes b)^\top = (a \otimes b)(a^\top \otimes b^\top) = (aa^\top) \otimes (bb^\top)$. Since $x_u^{(1)}$ and $x_v^{(1)}$ are independent given labels $y_u, y_v$, the tower property gives

$$\begin{aligned}
\mathbb{E}[Z_{uv}Z_{uv}^\top] &= \sum_{y,y'=1}^{L} \mathbb{E}\left[(f(x_v^{(1)})f(x_v^{(1)})^\top) \otimes (f(x_u^{(1)})f(x_u^{(1)})^\top) \;\middle|\; y_u^{(1)} = y, y_v^{(1)} = y'\right] p(y, y') \\
&= \sum_{y,y'=1}^{L} \mathbb{E}\left[f(x_v^{(1)})f(x_v^{(1)})^\top \;\middle|\; y_v^{(1)} = y'\right] \otimes \mathbb{E}\left[f(x_u^{(1)})f(x_u^{(1)})^\top \;\middle|\; y_u^{(1)} = y\right] p(y)p(y') \\
&= \left(\sum_{y'=1}^{L} \mathbb{E}\left[f(x_v^{(1)})f(x_v^{(1)})^\top \;\middle|\; y_v^{(1)} = y'\right] p(y')\right) \otimes \left(\sum_{y=1}^{L} \mathbb{E}\left[f(x_u^{(1)})f(x_u^{(1)})^\top \;\middle|\; y_u^{(1)} = y\right] p(y)\right) \\
&= \mathbb{E}[f(x_v^{(1)})f(x_v^{(1)})^\top] \otimes \mathbb{E}[f(x_u^{(1)})f(x_u^{(1)})^\top].
\end{aligned}$$

By assumption of the lemma, $\mathbb{E}[f(x_u^{(1)})f(x_u^{(1)})^\top] \succeq \overline{\lambda}_{\min}\mathbb{I}_L$, so it follows

$$\mathbb{E}[Z_{uv}Z_{uv}^\top] \succeq \overline{\lambda}^2_{\min}\mathbb{I}_{L^2},$$

where we use the fact that $\lambda_{\min}(A \otimes A) = \lambda_{\min}(A)^2$ for any positive semidefinite matrix $A$. Combining with (17), we get

$$-\nabla^2_{\mathrm{vec}(W)}\ell_{\mathrm{f}}(W) \succeq \overline{\lambda}^2_{\min}\mathbb{I}_{L^2}.$$

Finally, we have

$$-\nabla^2_{\mathrm{vech}(W)}\ell_{\mathrm{f}}(W) = -D_L^\top\nabla^2_{\mathrm{vec}(W)}\ell_{\mathrm{f}}(W)D_L \succeq \overline{\lambda}^2_{\min}D_L^\top D_L \succeq \overline{\lambda}^2_{\min}\mathbb{I}_{L(L+1)/2},$$

where the last step uses that $D_L^\top D_L$ is a diagonal matrix with entries either 1 or 2 (see the proof of Lemma H.3). This completes the proof.

# H Proof of lemmas

## H.1 Proof of Lemma E.1

We first introduce some notation. For each $a_{uv} \in \{0, 1\}$, we write $H_{uv}(x, x') = H(x, x', a_{uv})$ so that $H_{uv}$ depends on $a_{uv}$. By the symmetry of $H$, it follows $H_{uv}(x, x') = H_{uv}(x', x)$. We also define conditional expectations with respect to $x$, or $(x, x')$, where we write

$$P_y(H_{uv})(x') := \int H_{uv}(x, x')p(x \mid y)dx = \int H_{uv}(x', x)p(x \mid y)dx,$$

and

$$P_{y,y'}(H_{uv}) := \int H_{uv}(x,x')p(x \mid y)p(x' \mid y')dxdx'.$$

Let $\mathcal{F}_y := \sigma\left((y_u)_{u\in\mathbb{N}}\right)$ denote the $\sigma$-field generated by the sequence of node labels, and let $\mathcal{F}_{y,a} := \sigma\left((y_u)_{u\in\mathbb{N}}, (a_{uv})_{u<v}\right)$ denote the $\sigma$-field generated by both the node labels and edges.

With this notation, a Hoeffding-type decomposition (Lee, 2019) yields

$$\sum_{u<v} H_{uv}(x_u,x_v) = \underbrace{\sum_{u<v}\left(H_{uv}(x_u,x_v) - P_{y_v}(H_{uv})(x_u) - P_{y_u}(H_{uv})(x_v) + P_{y_u,y_v}(H_{uv})\right)}_{=:(A)}$$
$$+ \underbrace{\sum_{u<v}\left(P_{y_v}(H_{uv})(x_u) - P_{y_u,y_v}(H_{uv})\right)}_{=:(B)} + \underbrace{\sum_{u<v}\left(P_{y_u}(H_{uv})(x_v) - P_{y_u,y_v}(H_{uv})\right)}_{=:(C)} + \underbrace{\sum_{u<v} P_{y_u,y_v}(H_{uv})}_{=:(D)}. \tag{18}$$

Our strategy is to derive concentration bounds for each term.

**Term (A)** Starting with $(A)$, we condition on $\mathcal{F}_{y,A}$. For $u<v$, define

$$G^{uv}_{y_u,y_v}(x,x') := H_{uv}(x,x') - P_{y_v}(H_{uv})(x) - P_{y_u}(H_{uv})(x') + P_{y_u,y_v}(H_{uv}).$$

Then

$$\begin{aligned}
&\int G^{uv}_{y_u,y_v}(x,x')p(x \mid y_u)dx \\
&= \int \left(H_{uv}(x,x') - P_{y_v}(H_{uv})(x) - P_{y_u}(H_{uv})(x') + P_{y_u,y_v}(H_{uv})\right)p(x \mid y_u)dx \\
&= P_{y_u}(H_{uv})(x') - P_{y_u,y_v}(H_{uv}) - P_{y_u}(H_{uv})(x') + P_{y_u,y_v}(H_{uv}) = 0,
\end{aligned}$$

and similarly,

$$\int G^{uv}_{y_u,y_v}(x,x')p(x' \mid y_v)dx = 0.$$

So, conditioned on $\mathcal{F}_{y,a}$, the sum

$$(A) = \sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v),$$

is a canonical (matrix-valued) U-statistic of order 2 (Giné & Nickl, 2021, Section 3.4.3). Using concentration results from Minsker & Wei (2019, Section 3.3 and Section 4), we obtain the following Bernstein inequality, whose proof is deferred to the end.

**Lemma H.1** *For all $t \geq 2$, we have*

$$\mathbb{P}\left\{\left\|\sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v)\right\| \geq c_0 M \cdot B(t)\middle|\mathcal{F}_{y,a}\right\} \leq e^{-t},$$

*where $c_0 > 0$ is a universal constant, and*

$$B(t) := \log(3d)\cdot n(1+\sqrt{t}) + nt + \sqrt{n}\left((\log d)^{3/2} + t^{3/2}\right) + t^2.$$

By the tower property, taking expectations over $\mathcal{F}_{y,a}$ then yields the unconditional bound, i.e., for $t \geq 2$,

$$\mathbb{P}\left\{\left\|\sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v)\right\| \geq c_0 M \cdot B(t)\right\} \leq e^{-t}.$$

Setting $t = \log(8/\delta)$, since $n \geq \max\{\log(3d), t\}$, it follows that $\sqrt{n}\left((\log d)^{3/2} + t^{3/2}\right) + t^2 \leq 2\log(3d)\cdot n(1+\sqrt{t}) + 3nt$ and the last two terms in $B(t)$ can be absorbed. Since $\log(8/\delta) \geq 2$ for $\delta \in (0,1)$, this yields

$$\mathbb{P}\left\{\left\|\sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v)\right\| \geq c_1 M\left(\log(3d)\cdot n(1+\sqrt{\log(8/\delta)}) + n\log(8/\delta)\right)\right\} \leq \frac{\delta}{8}, \tag{19}$$

for a universal constant $c_1 > 0$.

**Terms (B), (C)** Combining the sums for $(B)$ and $(C)$, we have

$$(B)+(C)=\sum_{u=1}^{n}\underbrace{\sum_{v\neq u}(P_{y_v}(H_{uv})(x_u)-P_{y_u,y_v}(H_{uv}))}_{=:G_n^u(x_u)}=\sum_{u=1}^{n}G_n^u(x_u).$$

Conditional on $\mathcal{F}_{y,a}$, the terms $G_n^u(x_u)$ are independent (but not identically distributed) random matrices. Since $\|H\|\leq M$ $P$-a.e., the triangle inequality gives $\|G_n^u(x_u)\|\leq 2nM$ for $P$-a.e. $x_u$. Thus $\sum_{u=1}^n G_n^u(x_u)$ is a sum of independent, (conditional) mean-zero random matrices, so the matrix Bernstein inequality (Tropp et al., 2015, Theorem 6.1.1) applies, i.e, for all $t\geq 0$,

$$\mathbb{P}\left\{\left\|\sum_{u=1}^{n}G_n^u(x_u)\right\|\geq t\;\middle|\;\mathcal{F}_{y,a}\right\}\leq 2d\exp\left(-\frac{t^2/2}{\nu+2nMt/3}\right),\tag{20}$$

where

$$\nu=\left\|\sum_{u=1}^{n}\mathbb{E}\left[G_n^u(x_u)^2\mid\mathcal{F}_{y,a}\right]\right\|.$$

By convexity of the operator norm and Jensen's inequality, we can further bound

$$v\leq\sum_{u=1}^{n}\left\|\mathbb{E}\left[G_n^u(x_u)^2\mid\mathcal{F}_{y,a}\right]\right\|\leq\sum_{u=1}^{n}\mathbb{E}\left[\left\|G_n^u(x_u)^2\right\|\mid\mathcal{F}_{y,a}\right]\leq\sum_{u=1}^{n}\mathbb{E}\left[\left\|G_n^u(x_u)\right\|^2\;\middle|\;\mathcal{F}_{y,a}\right]$$
$$\leq 4M^2n^3,$$

where the first step uses the triangle inequality, and the third step applies the submultiplicativity of the operator norm. Setting $t=2\sqrt{\nu\log(8d/\delta)}+\frac{8Mn}{3}\log(8d/\delta)$ in (20) and plugging the above bound, we obtain the conditional bound

$$\mathbb{P}\left\{\left\|\sum_{u=1}^{n}G_n^u(x_u)\right\|\geq 4Mn^{3/2}\sqrt{\log(8d/\delta)}+\frac{8Mn}{3}\log(8d/\delta)\;\middle|\;\mathcal{F}_{y,a}\right\}\leq\frac{\delta}{4}.$$

By the tower property, taking expectation over $\mathcal{F}_{y,a}$ gives the unconditional bound

$$\mathbb{P}\left\{\left\|\sum_{u=1}^{n}G_n^u(x_u)\right\|\geq 4Mn^{3/2}\sqrt{\log(8d/\delta)}+\frac{8Mn}{3}\log(8d/\delta)\right\}\leq\frac{\delta}{4}.\tag{21}$$

**Term (D)** Define the conditional expectation of $P_{y,y'}(H_{uv})=P_{y,y'}(H(\cdot,\cdot,a_{uv}))$ over $a_{uv}$ as

$$P_{y,y'}(H):=\int H(x,x',1)p(x\mid y)p(x'\mid y')q(y,y')dxdx'$$
$$+\int H(x,x',0)p(x\mid y)p(x'\mid y')(1-q(y,y'))dxdx'.$$

Because $H$ is symmetric and $q(y,y')=q(y',y)$, it follows that $P_{y,y'}(H)=P_{y',y}(H)$. Then we decompose term $(D)$ as

$$\sum_{u<v}P_{y_u,y_v}(H_{uv})=\sum_{u<v}P_{y_u,y_v}(H(\cdot,\cdot,a_{uv}))=\sum_{u<v}(P_{y_u,y_v}(H(\cdot,\cdot,a_{uv}))-P_{y_u,y_v}(H))$$
$$+\sum_{u<v}P_{y_u,y_v}(H).\tag{22}$$

Conditional on $\mathcal{F}_y$, the $a_{uv}$ are independent Bernoulli with parameter $q(y_u,y_v)$. Therefore the first sum on the right-hand side above is a sum of independent (conditional) mean-zero random matrices where each summand has operator norm at most $2M$ by the triangle inequality. So applying matrix Bernstein inequality (Tropp et al., 2015, Theorem 6.1.1) yields for all $t\geq 0$,

$$\mathbb{P}\left\{\left\|\sum_{u<v}(P_{y_u,y_v}(H(\cdot,\cdot,A_{uv}))-P_{y_u,y_v}(H))\right\|\geq t\;\middle|\;\mathcal{F}_y\right\}\leq 2d\exp\left(-\frac{t^2/2}{\nu'+2Mt/3}\right),\tag{23}$$

where

$$\nu' = \left\| \sum_{u<v} \mathbb{E}\left[ \left(P_{y_u,y_v}(H(\cdot,\cdot,a_{uv})) - P_{y_u,y_v}(H)\right)^2 \,\middle|\, \mathcal{F}_y \right] \right\|.$$

Taking $t = 2\sqrt{\nu' \log(8d/\delta)} + \frac{8M}{3}\log(8d/\delta)$ in (23) and then averaging over $\mathcal{F}_y$ gives

$$\mathbb{P}\left\{ \left\| \sum_{u<v} \left(P_{y_u,y_v}(H(\cdot,\cdot,a_{uv})) - P_{y_u,y_v}(H)\right) \right\| \geq 2\sqrt{\nu' \log(8d/\delta)} + \frac{8M}{3}\log(8d/\delta) \right\} \leq \frac{\delta}{4}.$$

It can be easily checked that $\nu' \leq 2M^2n^2$ using the triangle inequality, Jensen's inequality, and the submultiplicativity as in the earlier argument for $\nu$, which then yields

$$\mathbb{P}\left\{ \left\| \sum_{u<v} \left(P_{y_u,y_v}(H(\cdot,\cdot,a_{uv})) - P_{y_u,y_v}(H)\right) \right\| \geq 2\sqrt{2}Mn\sqrt{\log(8d/\delta)} + \frac{8M}{3}\log(8d/\delta) \right\} \leq \frac{\delta}{4}. \tag{24}$$

For the second sum in (22), note that $\{y_u\}$ are i.i.d., so $\sum_{u<v} P_{y_u,y_v}(H)$ is a matrix-valued U-statistic of order 2 in $(y_u)_{u=1}^n$. Since $\mathbb{E}[P_{y_u,y_v}(H)] = P(H)$, Hoeffding's decomposition Lee (2019) can be applied to $P_{y_u,y_v}(H) - P(H)$, and concentration bounds for each term of the decomposition, as in the bound of terms (A)-(C), give the following lemma, whose proof is deferred to the end.

**Lemma H.2** *For any $0 < \delta < 1$, if $n \geq \max\{\log(3d), \log(8/\delta)\}$, there is a universal constant $c_2 > 0$ such that*

$$\mathbb{P}\Bigg\{ \|P_{y_u,y_v}(H) - P(H)\| \geq c_2 M \Bigg( \log(3d)\cdot n(1+\sqrt{\log(8/\delta)}) + n\log(8/\delta) + n^{3/2}\sqrt{\log(8d/\delta)} + n\log(8d/\delta) \Bigg) \Bigg\} \leq \frac{3\delta}{8}. \tag{25}$$

**Putting everything together** Returning to (18), and combining (19), (21), (24), (25), and simplifying, it follows that with probability at least $1-\delta$, there exists a universal constant $c_3 > 0$ such that

$$\left\| \sum_{u<v} (H_{uv}(x_u, x_v) - P(H)) \right\| \leq c_3 M \left( n^{3/2}\sqrt{\log(8d/\delta)} + n\log(3d)\log(8/\delta) \right),$$

where we use the fact that for $\delta \in (0,1)$ and $d > 1$, $\log(3d)\log(8/\delta) \geq \log(8d/\delta)$. Dividing both sides by $\binom{n}{2}$ yields the desired result, therefore completing the proof.

**Proof of Lemma H.1** Conditional on $\mathcal{F}_{y,a}$, $\{x_u\}_{1\leq u\leq n}$ is independent sequence and $H_{uv}$ is a deterministic function defined as

$$H_{uv}(x,x') = \begin{cases} H(x,x',0), \text{ if } a_{uv} = 0, \\ H(x,x',1), \text{ if } a_{uv} = 1. \end{cases}$$

Throughout the proof we work conditionally on $\mathcal{F}_{y,a}$, so all probabilities and expectations are taken with respect to the conditional distribution given $\mathcal{F}_{y,a}$.

Since $\sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v)$ is a canonical U-statistic of order 2, we can apply the results from Minsker & Wei (2019) (see Equation (14), Theorem 4.1, and the subsequent inequalities) to obtain that, for all $q \geq 1$ and $t \geq 2$,

$$\mathbb{P}\left\{ \left\| \sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v) \right\| \geq c\left( \mathbb{E}\left\| \sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v) \right\| + A\sqrt{t} + Bt + Ct^{3/2} + Dt^2 \right) \right\} \leq e^{-t}, \tag{26}$$

where

$$A = 2\log(de)\left(\sum_u \mathbb{E}_{x_u}\left\|\sum_{v:v>u}\mathbb{E}_{x_v}\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\| + \left\|\sum_{u<v}\mathbb{E}_{x_u,x_v}\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|\right)^{1/2},$$

$$B = 2\left(\left\|\sum_{u<v}\mathbb{E}_{x_u,x_v}\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|\right)^{1/2},$$

$$C = 2\sqrt{1+\frac{\log d}{q}}\left(\sum_u \mathbb{E}_{x_u}\left\|\sum_{v:v>u}\mathbb{E}_{x_v}\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|^q\right)^{1/(2q)},$$

$$D = 2\left(\sum_{u<v}\mathbb{E}_{x_u,x_v}\left\|\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|^q\right)^{1/(2q)} + \left(1+\frac{\log d}{q}\right)\left(\sum_u \mathbb{E}_{x_u,x_v}\max_{v:v>u}\left\|\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|^q\right)^{1/(2q)}.$$

Here $\mathbb{E}_{x_u}[\cdot]$ and $\mathbb{E}_{x_v}[\cdot]$ denote the conditional expectations with respect to $x_u$ and $x_v$, respectively, while $\mathbb{E}_{x_u,x_v}[\cdot]$ denotes the conditional expectation with respect to both $x_u$ and $x_v$.

To bound $A$ and $B$, note that $\|H\| \le M$, $P$-a.e., implies $\left\|G^{uv}_{y_u,y_v}\right\| \le 4M$, $P$-a.e., due to the triangle inequality. By the submultiplicativity of the operator norm, it follows that $\left\|(G^{uv}_{y_u,y_v})^2\right\| \le 16M^2$ $P$-a.e.. Applying Jensen's inequality and the triangle inequality then yields the bounds

$$A \le 2\log(de)\left(16n(n-1)M^2\right)^{1/2} \le 8M\log(3d)\cdot n, \text{ and}$$
$$B \le 2\sqrt{8}Mn.$$

Furthermore, the function $\|\cdot\|^q$ is convex for $q \ge 1$, so we have

$$\begin{aligned}\sum_u \mathbb{E}_{x_u}\left\|\sum_{v:v>u}\mathbb{E}_{x_v}\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|^q &\overset{(i)}{\le} \sum_u \mathbb{E}_{x_u}(n-u)^{q-1}\sum_{v:v>u}\left\|\mathbb{E}_{x_v}\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|^q \\ &\overset{(ii)}{\le} \sum_u \mathbb{E}_{x_u}(n-u)^{q-1}\sum_{v:v>u}\mathbb{E}_{x_v}\left\|\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|^q \\ &\le \sum_{u<v} n^{q-1}16^q M^{2q} \le 16^q n^{q+1}M^{2q},\end{aligned}$$

where step $(i)$ follows from Minkowski's inequality and step $(ii)$ follows from Jensen's inequality. Using this bound, $C$ can be bounded as

$$C = 2\sqrt{1+\frac{\log d}{q}}\left(\sum_u \mathbb{E}_{x_u}\left\|\sum_{v:v>u}\mathbb{E}_{x_v}\left(G^{uv}_{y_u,y_v}(x_u,x_v)\right)^2\right\|^q\right)^{1/(2q)} \le 8M\sqrt{1+\frac{\log d}{q}}\,n^{\frac{1}{2}+\frac{1}{2q}}.$$

Since this holds for any $q \ge 1$, taking $q \to \infty$ yields

$$C \le 8M\sqrt{n}.$$

Finally, $D$ can be bounded as

$$\begin{aligned}D &\le 2\left(\frac{n(n-1)}{2}16^q M^{2q}\right)^{1/(2q)} + \left(1+\frac{\log d}{q}\right)\left(16^q M^{2q} n\right)^{1/(2q)} \\ &\le 8Mn^{1/q} + 4M\left(1+\frac{\log d}{q}\right)n^{1/(2q)},\end{aligned}$$

and letting $q \to \infty$ gives

$$D \le 12M.$$

Combining together the bounds for $A, B, C, D$, and plugging into (26), yields

$$\mathbb{P}\left\{ \left\| \sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v) \right\| \geq c' \left[ \mathbb{E} \left\| \sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v) \right\| + M\log(3d)\cdot n\sqrt{t} + Mnt \right. \right. \\ \left. \left. + M\sqrt{n}t^{3/2} + Mt^2 \right] \right\} \leq e^{-t}. \quad (27)$$

It remains to bound $\mathbb{E}\left\|\sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v)\right\|$. By (Minsker & Wei, 2019, Equation (12)), we have

$$\mathbb{E}\left\| \sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v) \right\| \leq c'' \log d \left[ \left( \sum_u \mathbb{E}_{x_u} \left\| \sum_{v:v>u} \mathbb{E}_{x_v} \left( G^{uv}_{y_u,y_v}(x_u,x_v) \right)^2 \right\| \right)^{1/2} \right. \\ \left. + \left\| \sum_{u<v} \mathbb{E}_{x_u,x_v} \left( G^{uv}_{y_u,y_v}(x_u,x_v) \right)^2 \right\|^{1/2} + \sqrt{\log d} \left( \mathbb{E}_{x_u,x_v} \max_u \left\| \sum_{v:v>u} \left( G^{uv}_{y_u,y_v}(x_u,x_v) \right)^2 \right\| \right)^{1/2} \right].$$

Using Jensen's inequality and the triangle inequality, we can bound the right-hand side similarly to the term $A$, to get

$$\mathbb{E}\left\| \sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v) \right\| \leq c'' \log d \left( 2\sqrt{8n(n-1)M^2} + \sqrt{\log d}\sqrt{16(n-1)M^2} \right) \\ \leq c'' M \left( \log d \cdot n + (\log d)^{3/2} \cdot \sqrt{n} \right).$$

Plugging into (27) and simplifying, we obtain

$$\mathbb{P}\left\{ \left\| \sum_{u<v} G^{uv}_{y_u,y_v}(x_u,x_v) \right\| \geq c''' M \left[ \log(3d)\cdot n(1+\sqrt{t}) + nt + \sqrt{n}((\log d)^{3/2} + t^{3/2}) + t^2 \right] \right\} \leq e^{-t},$$

which completes the proof of the lemma.

**Proof of Lemma H.2** For notational convenience, write $\widetilde{H}(y_u, y_v) := P_{y_u,y_v}(H)$. Let

$$U_n = \sum_{u<v} \widetilde{H}(y_u, y_v).$$

Since $\widetilde{H}$ is symmetric ($\widetilde{H}(y_u,y_v) = \widetilde{H}(y_v,y_u)$), Hoeffding's decomposition Lee (2019) gives

$$U_n - \mathbb{E}[U_n] = \underbrace{\sum_{u<v} \left( \widetilde{H}(y_u,y_v) - \mathbb{E}_{y_u}\left[\widetilde{H}(y_u,y_v)\right] - \mathbb{E}_{y_v}\left[\widetilde{H}(y_u,y_v)\right] + \mathbb{E}_{y_u,y_v}\left[\widetilde{H}(y_u,y_v)\right] \right)}_{=:T_1} \\ + \underbrace{(n-1)\sum_{u=1}^{n} \left( \mathbb{E}_{y_v}\left[\widetilde{H}(y_u,y_v)\right] - \mathbb{E}_{y_u,y_v}\left[\widetilde{H}(y_u,y_v)\right] \right)}_{=:T_2},$$

where $\mathbb{E}_{y_u}[\cdot]$ denotes expectation over $y_u$, and $\mathbb{E}_{y_u,y_v}[\cdot]$ denotes expectation over $y_u, y_v$.

The term $T_1$ is a canonical U-statistic of order 2. So applying the same argument from Lemma H.1 (now without conditioning on $\mathcal{F}_{y,a}$), we can obtain

$$\mathbb{P}\left\{\|T_1\| \geq c_0 M \cdot B(t)\right\} \leq e^{-t},$$

where $c_0 > 0$ and $B(t)$ are as defined in Lemma H.1. Setting $t = \log(8/\delta)$, as long as $n \geq \max\{\log(3d), \log(8/\delta)\}$, the same reasoning below Lemma H.1 gives

$$\mathbb{P}\left\{ \|T_1\| \geq c_1 M \left( \log(3d)\cdot n(1+\sqrt{\log(8/\delta)}) + n\log(8/\delta) \right) \right\} \leq \frac{\delta}{8}.$$

For $T_2$, we follow the same reasoning that yields (21) (again without conditioning on $\mathcal{F}_{y,a}$), to get

$$\mathbb{P}\left\{ \|T_2\| \geq 4Mn^{3/2}\sqrt{\log(8d/\delta)} + \frac{8Mn}{3}\log(8d/\delta) \right\} \leq \frac{\delta}{4}.$$

Combining the two bounds gives the desired result.

### H.2 Proof of Lemma F.1

Define

$$H(x,x',a) := D_L^\top \frac{\mathrm{vec}(f(x)f(x')^\top)\mathrm{vec}(f(x)f(x')^\top)^\top}{S_{\mathrm{f}}^2(W_{\mathrm{f}};x,x',a)} D_L.$$

We first show that $H$ is symmetric, i.e., $H(x,x',a) = H(x',x,a)$. By Magnus & Neudecker (2019, Theorem 3.14), for any matrix $B$, we have

$$D_L^\top \mathrm{vec}(B) = \mathrm{vech}(B + B^\top - \mathrm{diag}(B)).$$

Since $\mathrm{diag}(B) = \mathrm{diag}(B^\top)$, it follows that $D_L^\top \mathrm{vec}(B) = D_L^\top \mathrm{vec}((B+B^\top)/2)$. Writing $M_{x,x'} = f(x)f(x')^\top$ and $\overline{M} = (M_{x,x'} + M_{x',x})/2$, we obtain

$$\begin{aligned} D_L^\top \mathrm{vec}(f(x)f(x')^\top)\mathrm{vec}(f(x)f(x')^\top)^\top D_L &= D_L^\top \mathrm{vec}(M_{x.x'})\mathrm{vec}(M_{x.x'})^\top D_L \\ &= D_L^\top \mathrm{vec}(\overline{M})\mathrm{vec}(\overline{M})^\top D_L. \end{aligned}$$

This expression is clearly unchanged if we swap $x$ and $x'$. Moreover, $S_{\mathrm{f}}(W_{\mathrm{f}};x,x',a) = f(x)^\top((1-a)(1-W_{\mathrm{f}}) + aW_{\mathrm{f}})f(x')$ also does not change if we swap $x$ and $x'$ since $W_{\mathrm{f}}$ is symmetric. Hence $H(x,x',a) = H(x',x,a)$.

Next the following lemma states that $H$ is bounded in the support of the target distribution, whose proof is deferred to the end.

**Lemma H.3** *Under the condition of Lemma F.1, for all $(x,x',a) \in \mathcal{X}\times\mathcal{X}\times\{0,1\}$ in the support of $p^{(1)}(x_u,x_v,a_{uv})$, $\|H(x,x',a)\| \le 2\tau_{\min}^{-2}$.*

Observe that

$$\nabla^2\ell(w_{\mathrm{f}}) = -\mathbb{E}[H(x_u^{(1)},x_v^{(1)},a_{uv}^{(1)})], \text{ and } \nabla^2\widehat{\ell}(w_{\mathrm{f}}) = -\frac{1}{\binom{n^{(1)}}{2}}\sum_{u<v} H(x_u^{(1)},x_v^{(1)},a_{uv}^{(1)}).$$

Then $H$ satisfies the conditions of Lemma E.1. Since $n^{(1)} \ge (\log(3L^2))^2\log(8/\delta)$ by assumption of lemma, applying Lemma E.1 with $d = L(L+1)/2 \le L^2$ yields that with probability at least $1-\delta$, there exists a universal constant $c_1 > 0$ such that

$$\left\|\nabla^2\ell(w_{\mathrm{f}}) - \nabla^2\widehat{\ell}(w_{\mathrm{f}})\right\| \le c_1\tau_{\min}^{-2}\left(\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}} + \frac{\log(3L^2)\log(8/\delta)}{n^{(1)}}\right).$$

Using $n^{(1)} \ge (\log(3L^2))^2\log(8/\delta)$ again, the second term on the right-hand side is dominated by the first term and so we obtain

$$\left\|\nabla^2\ell(w_{\mathrm{f}}) - \nabla^2\widehat{\ell}(w_{\mathrm{f}})\right\| \le 2c_1\tau_{\min}^{-2}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}}.$$

Finally, Weyl's inequality yields

$$\begin{aligned} \lambda_{\min}(-\nabla^2\widehat{\ell}(w_{\mathrm{f}})) &\ge \lambda_{\min}(-\nabla^2\ell(w_{\mathrm{f}})) - \left\|\nabla^2\ell(w_{\mathrm{f}}) - \nabla^2\widehat{\ell}(w_{\mathrm{f}})\right\| \\ &\ge \lambda_{\min}(-\nabla^2\ell(w_{\mathrm{f}})) - 2c_1\tau_{\min}^{-2}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}}. \end{aligned}$$

This completes the proof of the lemma.

**Proof of Lemma H.3** First, by Assumption 5.1, $S_{\mathrm{f}}(W_{\mathrm{f}};x,x',a) \ge \tau_{\min}$ on the support of $p^{(1)}(x_u,x_v,a_{uv})$. Since

$$\|\mathrm{vec}(f(x)f(x')^\top)\|_2 = \|f(x')\otimes f(x)\|_2 = \|f(x)\|_2\|f(x')\|_2,$$

we have

$$\begin{aligned} \frac{1}{S_{\mathrm{f}}(W_{\mathrm{f}};x,x',a)}\left\|\mathrm{vec}(f(x)f(x')^\top)\right\|_2 &= \frac{1}{S_{\mathrm{f}}(W_{\mathrm{f}};x,x',a)}\|f(x)\|_2\|f(x')\|_2 \\ &\le \tau_{\min}^{-1}\|f(x)\|_1\|f(x')\|_1 = \tau_{\min}^{-1}, \end{aligned} \tag{28}$$

where the last two steps use the fact that $\|\cdot\|_2 \le \|\cdot\|_1$ and $\|f(x)\|_1 = 1$ for all $x \in \mathcal{X}$. Since $\|ww^\top\| = \|w\|^2$ for any vector $w$, it follows

$$\|H(x,x',a)\| \le \frac{\|D_L\|^2 \left\|\mathrm{vec}(f(x)f(x')^\top)\right\|_2^2}{S_{\mathrm{f}}^2(W_{\mathrm{f}};x,x',a)} \le \tau_{\min}^{-2}\|D_L\|^2,$$

where the first step uses the submultiplicativity of the operator norm. To bound $\|D_L\|$, note that since $D_L$ has full column rank,

$$\|D_L\| = \sqrt{\lambda_{\max}\left(D_L^\top D_L\right)}.$$

We claim that $D_L^\top D_L$ is a diagonal matrix with entries equal to either 1 or 2. If so, $\|D_L\| \le \sqrt{2}$ and by the calculation above, we conclude

$$\|H(x,x',a)\| \le 2\tau_{\min}^{-2},$$

which proves the lemma.

It remains to prove the claim. Recall that for any symmetric matrix $B$, $\mathrm{vec}(B) = D_L\mathrm{vech}(B)$ by definition of $D_L$. Note that each column of $D_L$ corresponds to one coordinate in $\mathrm{vech}(B)$: (1) If the coordinate of $\mathrm{vech}(B)$ corresponds to the diagonal entry $(i,i)$ of $B$, the corresponding column of $D_L$ contains a single nonzero entry equal to 1 located at the vectorized position of $(i,i)$. Its squared norm is therefore 1; (2) If the coordinate of $\mathrm{vech}(B)$ corresponds to an off-diagonal entry $(i,j)$ with $i<j$, the corresponding column of $D_L$ has exactly two nonzero entries, both equal to 1, located at the vectorized positions of $(i,j)$ and $(j,i)$. Its squared norm is therefore 2.

Moreover, two different columns of $D_L$ must have disjoint supports, so they are orthogonal. It follows that $D_L^\top D_L$ is diagonal with entries equal to 1 or 2. This establishes the claim and completes the proof.

### H.3 Proof of Lemma F.2

Define

$$G(x,x',a) := D_L^\top \frac{(2a-1)\mathrm{vec}(f(x)f(x')^\top)}{S_{\mathrm{f}}(W_{\mathrm{f}};x,x',a)}.$$

Following the same reasoning used in the proof of Lemma F.1, we can easily check that $G(x,x',a) = G(x',x,a)$. Furthermore, since $|2a-1| \le 1$ for $a \in \{0,1\}$, combining with (28) yields

$$\|G(x,x',a)\|_2 \le \tau_{\min}^{-1} \text{ for all } (x,x',a) \in \mathcal{X}\times\mathcal{X}\times\{0,1\} \text{ in the support of } p^{(1)}(x_u,x_v,a_{uv}).$$

Now observe that

$$\nabla\ell(w_{\mathrm{f}}) = \mathbb{E}[G(x_u^{(1)},x_v^{(1)},a_{uv}^{(1)})],$$
$$\nabla\widehat{\ell}(w_{\mathrm{f}}) = \frac{1}{\binom{n^{(1)}}{2}}\sum_{u<v} G(x_u^{(1)},x_v^{(1)},a_{uv}^{(1)}).$$

Since $n^{(1)} \ge (\log(3L^2))^2\log(8/\delta) \ge \max\{\log(3L^2),\log(8/\delta)\}$ by assumption of lemma and $L(L+1)/2+1 \le L^2$ for $L \ge 2$, the conditions of Corollary E.1 hold. Therefore, Corollary E.1 gives that with probability at least $1-\delta$,

$$\left\|\nabla\ell(w_{\mathrm{f}}) - \nabla\widehat{\ell}(w_{\mathrm{f}})\right\|_2 \le c_1\tau_{\min}^{-1}\left(\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}} + \frac{\log(3L^2)\log(8/\delta)}{n^{(1)}}\right),$$

for some universal constant $c_1 > 0$. Under the assumption $n^{(1)} \ge (\log(3L^2))^2\log(8/\delta)$, the second term on the right-hand side is dominated by the first term, so the above bound simplifies to

$$\left\|\nabla\ell(w_{\mathrm{f}}) - \nabla\widehat{\ell}(w_{\mathrm{f}})\right\|_2 \le 2c_1\tau_{\min}^{-1}\sqrt{\frac{\log(8L^2/\delta)}{n^{(1)}}}.$$

This completes the proof.

### H.4 Proof of Lemma F.3

We begin by expanding the $\ell_1$ norm

$$\|\text{vec}(w_1 w_2^\top) - \text{vec}(w_1' w_2'^\top)\|_1 = \sum_i \sum_j |(w_1)_i (w_2)_j - (w_1')_i (w_2')_j|.$$

Adding and subtracting the term $(w_1)_i (w_2')_j$ inside the absolute value, we have

$$\begin{aligned}
&\|\text{vec}(w_1 w_2^\top) - \text{vec}(w_1' w_2'^\top)\|_1 \\
&= \sum_i \sum_j |(w_1)_i (w_2)_j - (w_1)_i (w_2')_j + (w_1)_i (w_2')_j - (w_1')_i (w_2')_j| \\
&\leq \sum_i \sum_j \left( |(w_1)_i (w_2)_j - (w_1)_i (w_2')_j| + |(w_1)_i (w_2')_j - (w_1')_i (w_2')_j| \right) \\
&= \sum_i \sum_j \left( |(w_1)_i| |(w_2)_j - (w_2')_j| + |(w_2')_j| |(w_1)_i - (w_1')_i| \right),
\end{aligned}$$

where the second step follows from triangle inequality. Simplifying the right-hand side yields

$$\begin{aligned}
&\|\text{vec}(w_1 w_2^\top) - \text{vec}(w_1' w_2'^\top)\|_1 \\
&= \left( \sum_i |(w_1)_i| \right) \left( \sum_j |(w_2)_j - (w_2')_j| \right) + \left( \sum_j |(w_2')_j| \right) \left( \sum_i |(w_1)_i - (w_1')_i| \right) \\
&= \|w_1\|_1 \|w_2 - w_2'\|_1 + \|w_2'\|_1 \|w_1 - w_1'\|_1.
\end{aligned}$$

This proves the lemma.

## I Use of large language models (LLMs)

We used LLMs solely to aid in correcting typos and checking grammar.