# FAIRPROP: FAIR NODE REPRESENTATION LEARNING VIA DIFFERENTIABLE PROPAGATION LAYERS

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

## ABSTRACT

Graph neural networks (GNNs) are the standard for node representation learning and are increasingly deployed in high-stakes domains. However, the message-passing backbone of GNNs can amplify topological bias, raising concerns about the fairness of their predictions. We study group fairness at the level of downstream predictions, spanning node classification, link prediction, and node regression, and derive bounds on the demographic parity gap for any number of sensitive groups. For node classification, our bound is provably no looser than the closest prior bound; for link prediction, we are the first to bound the parity gap of the deployed, sigmoid-activated prediction rather than a pre-activation proxy; and for node regression, we establish the first such bound. Common to all three tasks, our analysis reveals two distinct sources of bias: the separation between group means and the within-group covariance of the final representations. Guided by this analysis, we embed fairness directly into the propagation mechanism by augmenting the convex feature-smoothing problem solved by APPNP with a convex constraint on the group-mean gap and a within-group covariance regularizer. Unfolding projected gradient descent on this problem yields FairProp, whose layers consist of a propagation step followed by a closed-form projection, and which provably converges linearly to the unique fair optimum. Across three tasks, we empirically demonstrate that FairProp, even with exact group-mean equalization, provides a strong inductive bias that attains excellent fairness-utility trade-offs relative to prevailing baselines on benchmark datasets.

## 1 INTRODUCTION

Graph neural networks (GNNs) (Kipf & Welling, 2016; Hamilton et al., 2017; Veličković et al., 2018) exhibit state-of-the-art predictive performance across high-stakes graph machine learning (ML) tasks including credit scoring (Yuan et al., 2025), recidivism prediction (Jordan & Freiburger, 2015; Agarwal et al., 2021) and recommender systems (Wu et al., 2022). This stems from the ability of GNNs to effectively leverage graph topology and node features to learn discriminative node representations via recursive feature aggregation across local neighborhoods - a mechanism known as message passing (Gilmer et al., 2017). However, the increasing deployment of GNNs in such high-stakes tasks has raised concerns that their predictions disproportionately harm certain demographic subgroups (Rahman et al., 2019; Dai & Wang, 2021).

Such concerns are not unfounded: real-world graphs often exhibit homophily (Hofstra et al., 2017) with respect to sensitive node features (e.g., race, gender, age), where nodes sharing such an attribute preferentially connect to each other. As a result, node representations learned via message passing can become highly correlated with sensitive features, and studies have shown that message passing amplifies such topology bias, yielding disparate outcomes in downstream node classification (Kose & Shen, 2024a; Dai & Wang, 2021), link prediction (Li et al., 2021; Kose & Shen, 2024b) and node regression (Park & Lim, 2025) tasks.

While recent efforts have sought to mitigate such bias through fairness-aware GNNs (see surveys (Dong et al., 2023; Chen et al., 2024) references therein), existing approaches suffer from key limitations. Pre-processing approaches rewire the link structure alone (Spinelli et al., 2021; Loveland et al., 2022) or jointly with node features (Dong et al., 2022), independently of the downstream task. This requires solving an expensive combinatorial edge-selection problem. Moreover, since edges encode both task-relevant information and sensitive attribute correlation, correcting the latter per-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

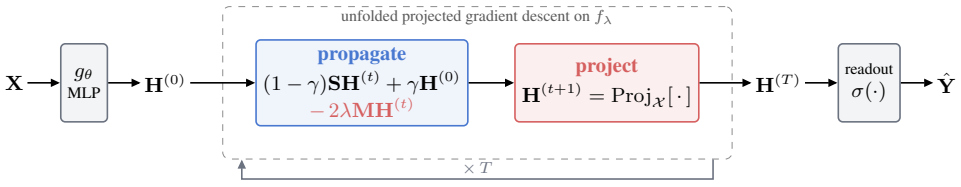



Figure 1: FairProp Architecture

turbs the former without explicit control of the trade-off. In-processing approaches intervene during training. Some target fairness of the node representations (Bose & Hamilton, 2019; Buyl & De Bie, 2020; Agarwal et al., 2021), but do not specify how it transfers to a downstream task. Others train the model weights with an additional fairness-promoting term via techniques such as adversarial training (Dai & Wang, 2021; Wang et al., 2022), invariant learning (Zhu et al., 2024a) or disentanglement (Zhu et al., 2024b; Lee et al., 2025). The regularization weight of this term then requires extensive tuning, and adversarial variants add min-max training instability. A third class reweights edges during message passing, via attention coefficients (Kose & Shen, 2024a), transition probabilities (Rahman et al., 2019; Khajehnejad et al., 2022) or the adjacency matrix (Kang et al., 2022). In contrast, FMP (Jiang et al., 2024) is the only prior method that embeds the sensitive attribute directly into the aggregation step of GNNs, but it only considers node classification.

In this paper, we consider group fairness, which seeks equitable treatment across demographic groups, and specifically adopt Demographic Parity (Dwork et al., 2012), a popular notion studied in graph ML (Dong et al., 2023). Departing from the majority of the prior art, our goal is to design a unified in-processing approach to learning fair node representations via message passing that provably mitigates bias across multiple downstream tasks, including node classification, link prediction, and node regression. Our contributions can be listed as follows:

$(i)$ *Task-level bias analysis:* We first conduct a bias analysis for the aforementioned downstream tasks, which upper bounds the demographic parity gap in terms of the separation of the group means and the within-group variances of the final node representations (before the task-level prediction layer). For node classification, our analysis yields a tighter bound than the one derived in (Kose & Shen, 2024a), and for link prediction, our result directly bounds the prediction-level parity gap, in contrast to (Li et al., 2021; Kose & Shen, 2024b), which only bound the pre-activation parity gap. Finally, to the best of our knowledge, we establish the first such bound on the parity gap for node regression. The key insight from our analysis is that bias mitigation across these tasks can be accomplished in a unified manner by controlling the separation of the group means and the within-group variances of the node representations.
$(ii)$ *The FairProp architecture:* We employ the fairness-agnostic APPNP (Gasteiger et al., 2018) as a backbone and describe how its propagation layers can be modified in a simple yet principled manner to perform effective bias mitigation, resulting in our **FairProp** architecture (Figure 1). Unlike conventional message-passing GNNs, APPNP decouples prediction from propagation, which provides a clean setting to isolate the effect of bias mitigation at the propagation level. We consider the unconstrained, strongly convex feature propagation problem that the APPNP propagation layers implicitly solve (Ma et al., 2021), and augment it to enact bias mitigation via: (a) a convex constraint bounding the separation of the group means, and (b) a quadratic regularizer minimizing within-group variance. We show that iterations of projected gradient descent applied on the modified convex problem can be unfolded into layers comprising an APPNP-style propagation step followed by a projection onto a convex set, and that this projection admits a closed-form solution in the forward and backward pass, enabling efficient end-to-end training for any downstream task.
$(iii)$ *Empirical validation:* We evaluate **FairProp** on three tasks against 18 baselines across benchmark datasets. Our empirical results show that FairProp with exact group mean equalization provides strong inductive bias which substantially reduces the parity gap of the APPNP backbone at negligible cost in utility. Compared to competing baselines, many of which achieve strong utility only at high cost of fairness, FairProp achieves the lowest parity gap on node classification, a near-zero gap on link prediction, and exactly zero on node regression, while maintaining strong performance.

### 1.1 Related Work

Most fairness-aware GNNs target node classification. One line trains fair models directly, through adversarial masking (FairGNN (Dai & Wang, 2021), FairVGNN (Wang et al., 2022)), counterfactual perturbation (NIFTY (Agarwal et al., 2021)), normalization (Fast&Fair (Kose & Shen, 2023)), invariant learning (FairINV (Zhu et al., 2024a)), disentanglement (FairSAD (Zhu et al., 2024b), DAB-

GNN (Lee et al., 2025)) or graph partitioning in graph transformers (FairGP (Luo et al., 2025)). A second family modifies how information travels along edges, by reweighting transition probabilities (FairWalk (Rahman et al., 2019), CrossWalk (Khajehnejad et al., 2022)), attention coefficients (FairGAT (Kose & Shen, 2024a)), dropout (FairDrop (Spinelli et al., 2021)) or the normalized adjacency (RawlsGCN (Kang et al., 2022)). Finally, pre-processing methods edit the input graph (EDITS (Dong et al., 2022), FairEdit (Loveland et al., 2022)). Node regression is addressed only by FnRGNN (Park & Lim, 2025). For link prediction, FairAdj (Li et al., 2021) formalizes *dyadic* fairness, later regularized structurally by FairWire (Kose & Shen, 2024b), which also provides a bias analysis for link prediction. The idea of projecting onto a fairness constraint set at every iteration has also been applied to PageRank by FairRARI (Kariotakis & Konar, 2026); we bring it into end-to-end GNN training, where the projection must also be differentiable.

Closest to our work, FMP (Jiang et al., 2024) considers demographic parity in node classification and also applies algorithm unrolling to solve a fairness-regularized optimization problem over node representations, enacting fair message passing. However, since the fairness term is non-convex, it is unclear whether the resulting algorithm converges to the underlying optimum, and it also entails complicated hyperparameter tuning for trading off accuracy and fairness. Moreover, its bias-mitigation mechanism is derived from the softmax Jacobian specific to node classification outputs, making its extension to link prediction or node regression non-trivial. In contrast, our framework unrolls a convex problem that directly constrains the separation of group means in the node representations, provably achieves linear convergence to the constrained optimum via simple projected gradient descent, and can be applied across many tasks. A full discussion is provided in Appendix A.

## 2 PRELIMINARIES

Let $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ denote an attributed, undirected graph with $n := |\mathcal{V}|$ nodes and $m := |\mathcal{E}|$ edges, $\mathbf{A} \in \mathbb{R}^{n\times n}$ the adjacency matrix, $\mathbf{D} \in \mathbb{R}^{n\times n}$ the diagonal degree matrix, and $\mathbf{X} \in \mathbb{R}^{n\times d}$ denote a matrix whose rows correspond to node features. Each node $v$ is annotated with a categorical sensitive feature $s_v \in [K]$, where $[K] := \{0, \dots, K-1\}$. The sensitive feature partitions the nodes of $\mathcal{G}$ into $K$ subgroups, with the $k^{th}$ subgroup $\mathcal{S}_k := \{v \in \mathcal{V} : s_v = k\}$, $k \in [K]$. In node-level tasks, predictions are made on nodes. Each $v \in \mathcal{V}$ has a ground truth value $Y(v)$ and a prediction $\hat{Y}(v)$, taking values in $\{0, 1\}$ for node classification and in $\mathbb{R}$ for node regression. In link prediction, predictions are made on a subset of node pairs comprising the existing edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ together with a sample of non-edges. Since each pair involves *two* nodes, subgroup membership is defined dyadically, i.e., with intra ($s_v = s_u$) and inter-group ($s_v \neq s_u$) pairs.

**Group-Fairness:** Demographic parity (DP) requires that the model predictions be independent of the sensitive attribute (Dwork et al., 2012). For binary *node classification*, independence is equivalent to equal positive-prediction rates across demographic groups. Hence, we quantify its violation by $\Delta_{\mathrm{DP}}^{\mathrm{NC}} := \max_{k,\ell\in[K]} \left|Pr(\hat{Y}(v) = 1|v \in \mathcal{S}_k) - Pr(\hat{Y}(v) = 1|v \in \mathcal{S}_\ell)\right|$. For *node regression*, where predictions are continuous, we quantify the violation by the mean parity gap $\Delta_{\mathrm{MG}} := \max_{k,\ell\in[K]} \left|\mathbb{E}[\hat{Y}(v)|v \in \mathcal{S}_k] - \mathbb{E}[\hat{Y}(v)|v \in \mathcal{S}_\ell]\right|$ (Wei et al., 2023; Park & Lim, 2025). Finally, for *link prediction*, we use the dyadic demographic parity gap (Li et al., 2021) $\Delta_{\mathrm{DP}}^{\mathrm{LP}} := \left|\mathbb{E}[f(v,u)|s_v = s_u] - \mathbb{E}[f(v,u)|s_v \neq s_u]\right|$, where $f(v,u)$ denotes the predicted probability of an edge between the nodes $v$ and $u$.

**Remark 1.** *For node classification, we evaluate $\Delta_{\mathrm{DP}}^{\mathrm{NC}}$ using the hard prediction $\hat{Y}(v)$, since the task itself naturally produces a discrete label. For link prediction, however, performance is conventionally assessed via ranking (e.g., top-k retrieval), so there is no canonical threshold for converting the score $f(v,u)$ into a hard decision; we therefore evaluate $\Delta_{\mathrm{DP}}^{\mathrm{LP}}$ directly on the predicted scores.*

## 3 TASK-LEVEL BIAS ANALYSIS

In this section, we analyze and bound the demographic and mean parity gaps in terms of the final learned representations (prior to the task-specific output layer) across three core tasks: (i) node classification, (ii) link prediction, and (iii) node regression. Prior work has derived analogous task-level bounds for (i) (Kose & Shen, 2024a) and (ii) (Li et al., 2021; Kose & Shen, 2024b), but each is limited in a distinct way. For node classification, (Kose & Shen, 2024a) bounds a metric related to the demographic parity gap. However, this metric equals the true, hard-thresholded gap under a stochastic labeling rule applied on the softmax outputs rather than the deterministic "arg-max" decision rule standard in practice. We show that this metric in fact corresponds to a soft relaxation

of the DP gap under the standard decision rule, and we provide a tighter bound in this setting. For link prediction, (Li et al., 2021; Kose & Shen, 2024b) bound the parity gap of the pre-activation link score rather than the sigmoid-activated predicted probability that is actually deployed; we extend their result to the latter. Finally, to the best of our knowledge, we establish the first such bound for node regression. Together, these results reveal how the task-specific output transformation (e.g., softmax, inner product) can itself amplify or attenuate bias, offering theoretical insight into how node embeddings must be structured to effectively constrain the fairness gap for each task.

Let $\mathbf{h}_v$ denote the final representation of node $v \in \mathcal{V}$ before the prediction layer. Throughout, $U$ denotes the discrete uniform distribution over $\mathcal{V}$, and expectations are taken with respect to it. We write $\mathbb{E}[\,\cdot\,]$ for expectations over uniformly sampled vertices $v \sim U$ or vertex pairs $(v,u) \sim U \times U$. Define $\boldsymbol{\mu}_k := \mathbb{E}[\mathbf{h}_v \,|\, v \in \mathcal{S}_k]$ and $\mathbf{C}_k := \mathrm{cov}(\mathbf{h}_v \,|\, v \in \mathcal{S}_k)$ as the *group-conditional mean* and the *within-group covariance* of the node representations corresponding to group $\mathcal{S}_k$, respectively. All proofs are deferred to Appendix B.

**(i) Node Classification:** For a node classification task, the goal is to use the model to generate a class label for a given node $v$. This is done by first applying an affine composition of the node embedding vector $\mathbf{h}_v$ with a non-linear, $L$-Lipschitz activation function $\sigma(\cdot)$; typically the softmax operator (with $L = 1/2$). Let $g(v) := \mathbf{W}\mathbf{h}_v + \mathbf{b}$ denote the affine function of $\mathbf{h}_v$, and on composing it with $\sigma(\cdot)$, we obtain the vector of "soft" class labels $f(v) := \sigma(\mathbf{W}\mathbf{h}_v + \mathbf{b})$. The final "hard" label is obtained by returning the index of the largest entry of $f(v)$.

**Proposition 2.** *If* $\max_{k,\ell\in[K]} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell\|_2 \le \delta$*, then it holds that*

$$\max_{k,\ell\in[K]} \big\|\mathbb{E}[f(v)|v\in\mathcal{S}_k] - \mathbb{E}[f(v)|v\in\mathcal{S}_\ell]\big\|_2 \le L\left(\|\mathbf{W}\|_2\,\delta + 2\max_{k\in[K]}\sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_k\mathbf{W}^\top)}\right). \quad (1)$$

**Corollary 3.** *For binary classification with softmax output, let* $\Pr(\hat{Y}(v)=1)$ *denote the predicted positive class probability of node* $v$*, and* $\bar{p}_k := \frac{1}{n_k}\sum_{v\in\mathcal{S}_k}\Pr(\hat{Y}(v)=1)$ *denote the group-mean of predicted positive scores of group* $\mathcal{S}_k$*. Under the conditions of Proposition 2, it holds that the "soft" demographic parity gap:* $\max_{k,\ell\in[K]}|\bar{p}_k - \bar{p}_\ell| \le \frac{L}{\sqrt{2}}\left(\|\mathbf{W}\|_2\,\delta + 2\max_{k\in[K]}\sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_k\mathbf{W}^\top)}\right)$.

**Comparison with the FairGAT bound.** The result closest to Proposition 2 is that of FairGAT (Kose & Shen, 2024a), which is derived for $K = 2$ sensitive groups. In this setting, their result also upper bounds the same quantity by the separation of the group means $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2$ and the within-group spread. The main difference lies in bounding the second term via a uniform deviation $\ell_\infty$ bound $\Delta_z$ on the pre-activations $\mathbf{z}_v := \mathbf{W}\mathbf{h}_v$, whereas our $\ell_2$-style analysis bounds the spread in terms of the within-group covariances $\mathbf{C}_k$. Under the same hypothesis $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 \le \delta$, the two bounds can be expressed as $\mathcal{B}_{\mathrm{ours}} := L(\|\mathbf{W}\|_2\delta + \sum_{k=0}^{1}\sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_k\mathbf{W}^\top)})$ and $\mathcal{B}_{\mathrm{FairGAT}} := L(\|\mathbf{W}\|_2\delta + 2\sqrt{n}\Delta_z)$. Hence, the difference between the two bounds boils down to the expressions corresponding to the within-group spread. Note that the bound of Kose & Shen (2024a) becomes looser as the number of nodes $n$ grows while ours exhibits no such dependence. This observation suffices to show that our bound is guaranteed to be the tighter of the two under a mild condition on $n$, which holds in essentially every practical setting. Please refer to Appendix B.2 for the proof.

**Proposition 4.** *Let* $n$ *denote the number of nodes and* $c$ *be the number of classes in a node classification task). If* $n \ge c$*, then* $\mathcal{B}_{\mathrm{ours}} \le \mathcal{B}_{\mathrm{FairGAT}}$.

**(ii) Link Prediction:** For the link prediction task, given a vertex pair $\{v,u\}$, the model generates a soft prediction that indicates the likelihood of $u$ and $v$ forming a link. This is typically accomplished by using a final layer which accepts the nodal representations $\mathbf{h}_u, \mathbf{h}_v$, and computes their inner product, followed by a non-linear, $L$-Lipschitz function $\sigma(\cdot)$; e.g. sigmoid, with $L = 1/4$. Let's define the inner product function $g(v,u) := \mathbf{h}_v^\top\boldsymbol{\Sigma}\mathbf{h}_u$, where $\boldsymbol{\Sigma} \in \mathbb{S}^d_{++}$ is a positive-definite matrix. Composing the result with the non-linearity $\sigma(\cdot)$ we obtain $f(v,u) := \sigma(\mathbf{h}_v^\top\boldsymbol{\Sigma}\mathbf{h}_u)$. We make the standard assumption, following (Li et al., 2021), that for each $v \in \mathcal{V}$, there is a scalar $Q > 0$, such that $\|\mathbf{h}_v\|_2 \le Q$.

**Proposition 5.** *If* $\max_{k,\ell\in[K]} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell\|_2 \le \delta$*, then it holds that*

$$\big|\mathbb{E}[f(v,u)|s_u = s_v] - \mathbb{E}[f(v,u)|s_u \ne s_v]\big| \le LQ\|\boldsymbol{\Sigma}\|_2\left(2\delta + \sum\nolimits_{k\in[K]}(2\alpha_k + w_k)\sqrt{\mathrm{tr}(\mathbf{C}_k)}\right) \quad (2)$$

*where* $\alpha_k := \frac{n_k^2}{\sum_{j=0}^{K-1} n_j^2}$, $\beta_{kl} := \frac{n_k n_l}{n^2 - \sum_{j=0}^{K-1} n_j^2} = \frac{n_k n_l}{\sum_{i\ne j} n_i n_j}$, *with* $w_k := \sum_{l\ne k}(\beta_{kl} + \beta_{lk})$*, and* $\mathbf{C}_k := \mathrm{cov}(\mathbf{h}_v \mid v \in \mathcal{S}_k)$*, for every* $k \in [K]$.

Unlike the existing bounds in (Li et al., 2021; Kose & Shen, 2024b), which are derived with respect to the pre-activation bilinear score $g(v, u)$, extending their result to account for the sigmoid activation and bound the actual predicted edge probability $f(v, u) = \sigma(g(v, u))$ is technically non-trivial, and does not follow as a direct corollary. Our formulation instead directly bounds $\Delta_{\text{DP}}^{\text{LP}}$ as computed on the deployed, sigmoid-activated prediction, the exact quantity reported in empirical evaluation, rather than a pre-activation proxy.

**(iii) Node Regression:** For node regression, given node $v$, the model generates a prediction via the linear transformation $g(v) = \mathbf{W}\mathbf{h}_v + \mathbf{b}$ with $\mathbf{W} \in \mathbb{R}^{1\times d}$. Here, we have the following bound.

**Proposition 6.** *If* $\max_{k,l} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2 \leq \delta$*, then it holds that*

$$\max_{k,l} \{\|\mathbb{E}[g(v)|v \in \mathcal{S}_k] - \mathbb{E}[g(v)|v \in \mathcal{S}_l]\|_2\} \leq \|\mathbf{W}\|_2 \delta. \tag{3}$$

Note that the bound simplifies to an expression that is independent of the within-group covariance, i.e., the terms stemming from the non-linearity. For a more detailed analysis refer to Appendix B.4.

# 4 FairProp

## 4.1 APPNP layers as Unfolded Diffusion

The (A)PPNP architecture proposed in Gasteiger et al. (2018) comprises two distinct stages: an MLP prediction stage that takes as input only the node features $\mathbf{X}$ (but not the adjacency $\mathbf{A}$), followed by a propagation stage which applies diffusion on the output of the MLP. Let $\mathbf{H}^{(0)} = g_\theta(\mathbf{X})$ be the embedding produced by an MLP applied to the node features $\mathbf{X}$, $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-\frac{1}{2}}$ the propagation matrix, and $\mathbf{L_S} = \mathbf{I} - \mathbf{S}$ its Laplacian. The propagation scheme of PPNP solves the strongly convex quadratic problem (Zhu et al., 2021; Ma et al., 2021)

$$\mathbf{H}^* = \arg\min_{\mathbf{H}} \left\{ f(\mathbf{H}) := \tfrac{1}{2}\gamma\|\mathbf{H} - \mathbf{H}^{(0)}\|_F^2 + \tfrac{1}{2}(1-\gamma)\text{tr}(\mathbf{H}^\top \mathbf{L_S}\mathbf{H}) \right\}, \tag{4}$$

with the parameter $\gamma \in (0, 1]$ trading off feature smoothness along the graph edges against proximity to the input features $\mathbf{H}^{(0)}$. Gradient descent on equation 4 with step size $\eta = 1$ gives exactly the APPNP iterations $\mathbf{H}^{(t+1)} = (1-\gamma)\mathbf{S}\mathbf{H}^{(t)} + \gamma\mathbf{H}^{(0)}$. Hence the APPNP propagation layers correspond to *algorithmic unrolling* (Gregor & LeCun, 2010; Monga et al., 2021) of gradient descent applied to equation 4. Stacking $T$ APPNP propagation layers implements $T$ such gradient steps, resulting in an approximation of the optimal solution $\mathbf{H}^* = \gamma[\mathbf{I} - (1-\gamma)\mathbf{S}]^{-1}\mathbf{H}^{(0)}$.

In this framework, feature propagation is decoupled from prediction. Consequently, representation-level bias can be directly addressed by modifying the propagation scheme. Since propagation itself amounts to unrolled gradient descent on equation 4, bias mitigation can be enacted directly at the optimization level - by augmenting equation 4 with fairness-promoting terms - and the resulting propagation scheme obtained via unrolling inherits this bias mitigation by construction (see Figure 1). The closest approach to ours is FMP (Jiang et al., 2024), which only considers node classification and augments the PPNP objective with an $\ell_1$ DP regularizer on the softmax outputs, unrolling a primal-dual solver into a fair message-passing scheme. However, since the fairness term is non-convex in the node representations due to the softmax nonlinearity, their unrolled forward pass lacks optimality guarantees, and its accuracy–fairness trade-off is highly sensitive to hyperparameter tuning.

In contrast to FMP, guided by the analysis of Section 3, we design *convex* fairness-inducing mechanisms that act directly on the node embeddings $\mathbf{H}^*$. In the previous section, we showed that bias across tasks arises from two distinct sources: the separation between the group means $\delta$, and the within-group covariance $\mathbf{C}_k$. The former can be controlled via a hard constraint (enforced through projection) that either drives the group-mean separation to exactly zero or bounds it within a tolerance $\delta$, while the latter is penalized via a regularization term. Propositions 2, 5, and 6 bound the prediction disparity in terms of these quantities for $K$ groups across each task. Consequently, controlling both quantities in the embedding $\mathbf{H}^*$ suffices to yield fair predictions for an arbitrary number of groups, across all three downstream tasks.

## 4.2 Fair Projections

**Exact mean-equalization projection.** For an embedding matrix $\mathbf{H} \in \mathbb{R}^{n\times d}$ ($n$ nodes, $d$ features), zeroing out the mean group separation $\delta$ requires the group-means of the embeddings to coincide, i.e., $\bar{\mathbf{h}}_{\mathcal{S}_i} = \bar{\mathbf{h}}_{\mathcal{S}_j}$, for every pair of groups $i, j$, where $\bar{\mathbf{h}}_{\mathcal{S}_i} = \frac{1}{|\mathcal{S}_i|}\mathbf{e}_{\mathcal{S}_i}^\top \mathbf{H}$, with $\mathbf{e}_{\mathcal{S}_i}$ the indicator vector

of subgroup $\mathcal{S}_i$. This is equivalent to $\bar{\mathbf{h}}_{\mathcal{S}_k} = \bar{\mathbf{h}}_{\mathcal{S}_0}$, for every $k \in [K] \setminus \{0\}$. Define the *mean-equalization* set

$$\mathcal{X}_0 := \{\mathbf{H} : \mathbf{a}_k^\top \mathbf{H} = \mathbf{0},\ k \in [K] \setminus \{0\}\}, \qquad \mathbf{a}_k := \frac{\mathbf{e}_{\mathcal{S}_0}}{n_0} - \frac{\mathbf{e}_{\mathcal{S}_k}}{n_k}, \tag{5}$$

Note that $\mathcal{X}_0 = \{\mathbf{H} : \mathbf{A}^\top \mathbf{H} = \mathbf{0}\}$, where $\mathbf{A} := [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{K-1}]$; equivalently, $\mathcal{X}_0$ consists of all matrices whose columns lie in the orthogonal complement of $\mathrm{col}(\mathbf{A})$. In particular, it admits a closed-form Euclidean projection,

$$\mathbf{H}^* = \mathrm{Proj}_{\mathcal{X}_0}(\mathbf{H}) = \mathbf{P} \cdot \mathbf{H}, \qquad \mathbf{P} = \mathbf{I} + \frac{\mathbf{e}\mathbf{e}^\top}{n} - \sum_{k=0}^{K-1} \frac{\mathbf{e}_{\mathcal{S}_k} \mathbf{e}_{\mathcal{S}_k}^\top}{|\mathcal{S}_k|}. \tag{6}$$

At a level of a single node is $\mathbf{h}_v^* = \mathbf{h}_v + \bar{\mathbf{h}} - \bar{\mathbf{h}}_{\mathcal{S}_k}$, $\forall v \in \mathcal{S}_k$, i.e., every node embedding is shifted by the difference of the global mean $\bar{\mathbf{h}}$ and its own group mean $\bar{\mathbf{h}}_{\mathcal{S}_k}$. The operator $\mathbf{P}$ is a symmetric orthogonal projection. It is *parameter-free,* and *linear*, hence differentiable with backward pass $d\mathbf{H} = \mathbf{P} \cdot d\mathbf{H}^*$, identical to the forward pass, and equalizes all group means *exactly*, so $\delta = 0$ after each projection. In practice, $\mathbf{P}$ is never formed explicitly: computing the $K$ group means and shifting each row costs $\mathcal{O}(nd)$ time and $\mathcal{O}(Kd)$ memory for both the forward and backward pass, which is dominated by the $\mathcal{O}((m+n)d)$ cost of a single propagation step, resulting to $\mathcal{O}((m+n)dL)$, for $L$ layers. For the full derivation and the backward pass please refer to Appendix C.

**Relaxed fairness constraint.** Enforcing $\delta = 0$ exactly can be overly restrictive in practice, since it may force node representations to discard task-relevant information in order to fully eliminate group separation, resulting in a significant utility drop. Hence, we also consider the relaxed set $\mathcal{X}_\delta := \{\mathbf{H} : \|\mathbf{H}^\top \mathbf{a}\|_\infty \le \delta\}$, with $\mathbf{a} = \frac{\mathbf{e}_{\mathcal{S}_0}}{n_0} - \frac{\mathbf{e}_{\mathcal{S}_1}}{n_1}$, which bounds the group-mean gap by a *fairness tolerance* $\delta$ in every feature and recovers $\mathcal{X}_0$ at $\delta = 0$. For a binary sensitive attribute (i.e., $K = 2$), the projection onto $\mathcal{X}_\delta$ remains in closed form and decouples across feature columns: $\mathrm{Proj}_{\mathcal{X}_\delta}(\mathbf{H}) = \mathbf{H} - \frac{\mathbf{a}}{\|\mathbf{a}\|_2^2}\big(\mathrm{ReLU}(\mathbf{s} - \delta\mathbf{1}) - \mathrm{ReLU}(-\mathbf{s} - \delta\mathbf{1})\big)^\top$, with $\mathbf{s} = \mathbf{H}^\top \mathbf{a}$. The fairness layer is thus a ReLU-type layer with bias $-\delta$ acting on the group-mean gap. Like a ReLU, it is gated: columns whose gap lies within $\delta$ are inactive and pass through unchanged in both the forward and backward pass, while only violating columns are shifted back to the boundary of the slab. At $\delta = 0$ the two ReLUs sum to the identity, recovering the exact projection. For $K > 2$ groups, $\mathcal{X}_\delta$ becomes an intersection of $K - 1$ slabs, and the projection can be computed with Dykstra's algorithm (Boyle & Dykstra, 1986). For the backward pass, we can differentiate though the KKT conditions of the problem using implicit differentiation (Blondel et al., 2022). The full analysis can be found in Appendix D.

**Interleaving Fair Projections with Propagation.** In order to enforce the exact mean-equalization constraint of the set $\mathcal{X}_0$, or its relaxed slab alternative $\mathcal{X}_\delta$, we can solve the constrained version of the problem of equation 4, which is the following constrained strongly convex quadratic problem

$$\min \{f(\mathbf{H}) \ \text{ s.t. } \ \mathbf{H} \in \mathcal{X}_\delta\}. \tag{7}$$

With $\delta = 0$, the solution of the problem satisfies exact mean equalization, while $\delta > 0$ refers to the relaxed mean separation. Applying projected gradient descent on this problem, we get $\mathbf{H}^{(t+1)} = \mathrm{Proj}_{\mathcal{X}_\delta}[\mathbf{H}^{(t)} - \eta \nabla f(\mathbf{H}^{(t)})]$. Setting $\eta = 1$, results in $\mathbf{H}^{(t+1)} = \mathrm{Proj}_{\mathcal{X}_\delta}[(1-\gamma)\mathbf{S}\mathbf{H}^{(t)} + \gamma \mathbf{H}^{(0)}]$, which is equivalent to interleaving the fair projection layer after each propagation step of APPNP.

### 4.3 Variance Control

**Within-Group Variance Regularization.** In order to fully suppress the parity gap bounds, we have to not only minimize the mean gap, but also suppress the within-group variance terms. For each demographic group $\mathcal{S}_k$, let $\mathbf{H}_{\mathcal{S}_k} = \mathbf{B}_k \mathbf{H}$, where $\mathbf{B}_k \in \{0,1\}^{n_k \times n}$ is a binary selection matrix. Its mean-centered representation is given by $\tilde{\mathbf{H}}_{\mathcal{S}_k} = \mathbf{M}_k \mathbf{H}_{\mathcal{S}_k}$, with centering matrix $\mathbf{M}_k = \mathbf{I}_{n_k} - \frac{1}{n_k}\mathbf{1}\mathbf{1}^\top$. The empirical within-group covariance is then $\hat{\mathbf{C}}_k = \frac{1}{n_k}\tilde{\mathbf{H}}_{\mathcal{S}_k}^\top \tilde{\mathbf{H}}_{\mathcal{S}_k}$, leading to the regularizer, $\mathcal{L}_{\mathrm{cov}}(\mathbf{H}) := \sum_{k\in[K]} \mathrm{tr}(\hat{\mathbf{C}}_k) = \sum_{k\in[K]} \frac{1}{n_k}\|\tilde{\mathbf{H}}_{\mathcal{S}_k}\|_F^2$. Note that the centering of the embeddings is necessary, since without it we would penalize $\frac{1}{n_k}\|\mathbf{H}_{\mathcal{S}_k}\|_F^2 = \mathrm{tr}(\hat{\mathbf{C}}_k) + \|\bar{\mathbf{h}}_{\mathcal{S}_k}\|_2^2$, collapsing the group means $\bar{\mathbf{h}}_{\mathcal{S}_k}$ to zero. Defining $\mathbf{M} := \sum_{k\in[K]} \frac{1}{n_k}\mathbf{B}_k^\top \mathbf{M}_k \mathbf{B}_k$ gives $\mathrm{tr}(\mathbf{H}^\top \mathbf{M} \mathbf{H}) = \sum_{k\in[K]} \frac{1}{n_k}\|\tilde{\mathbf{H}}_{\mathcal{S}_k}\|_F^2$. Finally, we can express the regularizer compactly in matrix form, $\mathcal{L}_{\mathrm{cov}}(\mathbf{H}) = \mathrm{tr}(\mathbf{H}^\top \mathbf{M} \mathbf{H})$.

**Adding Variance Control.** The regularizer of Section 4.3 can be placed inside the diffusion objective itself, resulting in a task-agnostic regularization that controls the variance at every propagation step. Define the *variance-regularized* objective

$$f_\lambda(\mathbf{H}) := \tfrac{1}{2}\gamma\|\mathbf{H}-\mathbf{H}^{(0)}\|_F^2 + \tfrac{1}{2}(1-\gamma)\,\mathrm{tr}(\mathbf{H}^\top \mathbf{L_S}\mathbf{H}) \;+\; \lambda\,\mathrm{tr}(\mathbf{H}^\top \mathbf{M}\mathbf{H}), \tag{8}$$

with $\mathbf{M}$ as in Section 4.3. Since $\mathbf{L_S} \succeq 0$, and $\mathbf{M} \succeq 0$ as a sum of PSD terms, this remains a convex quadratic with Hessian $\mathbf{Q}_\lambda := \nabla^2 f_\lambda = \gamma\mathbf{I} + (1-\gamma)\mathbf{L_S} + 2\lambda\mathbf{M}$. Projected gradient descent on $\min\{f_\lambda(\mathbf{H}) \text{ s.t. } \mathbf{H} \in \mathcal{X}_0\}$ with step size $\eta$ gives

$$\mathbf{H}^{(t+1)} = \mathbf{P}[\mathbf{T}_{\eta,\lambda}\mathbf{H}^{(t)} + \eta\gamma\mathbf{H}^{(0)}], \qquad \mathbf{T}_{\eta,\lambda} := \mathbf{I} - \eta\,\mathbf{Q}_\lambda, \tag{9}$$

which for $\eta = 1$ becomes the FairProp propagation step with an added variance-control term, $\mathbf{H}^{(t+1)} = \mathbf{P}[(1-\gamma)\mathbf{S}\mathbf{H}^{(t)} + \gamma\mathbf{H}^{(0)} - 2\lambda\mathbf{M}\mathbf{H}^{(t)}]$, and fixed-point $\mathbf{H}^* = \eta\gamma\big(\mathbf{I} - \mathbf{P}\mathbf{T}_{\eta,\lambda}\big)^{-1}\mathbf{P}\,\mathbf{H}^{(0)}$. The following proposition provides convergence guarantees for the proposed fair propagation scheme. For the proof please refer to Appendix E.

**Proposition 7.** *Let $\gamma \in (0,1]$, $\lambda \geq 0$, and $n_{\min} := \min_k n_k$. If*

$$0 \;<\; \eta \;<\; \frac{2}{2-\gamma+\frac{2\lambda}{n_{\min}}} \Longleftrightarrow \lambda < n_{\min}\left(\frac{1}{\eta} - 1 + \frac{\gamma}{2}\right) \tag{10}$$

*then $\rho := \|\mathbf{T}_{\eta,\lambda}\|_2 < 1$ and the iterations in equation 9 converge linearly, $\|\mathbf{H}^{(t)} - \mathbf{H}^*\|_F \leq \rho^t\|\mathbf{H}^{(0)} - \mathbf{H}^*\|_F$, to the unique minimizer $\mathbf{H}^*$ of $\min\{f_\lambda(\mathbf{H}) \text{ s.t. } \mathbf{H} \in \mathcal{X}\}$. In particular, the standard APPNP step size $\eta = 1$ satisfies equation 10 if $\lambda < \gamma\, n_{\min}/2$, placing an upper bound on how aggressively within-group variance can be penalized while guaranteeing convergence.*

Note that for $\lambda < \infty$, the term $\lambda\,\mathrm{tr}(\mathbf{H}^\top\mathbf{M}\mathbf{H})$ *reduces* but does not zero the within-group variance, avoiding collapsing each group to a single point. On the other hand, the projection to the mean-equalization constraint admits a closed-form, is non-expansive, and removes the mean gap exactly. Replacing $\mathbf{P}$ by the projection operator of the slab projection $\mathrm{Proj}_{\mathcal{X}_\delta}$ constrains the group-mean gap to $\delta$ rather than to zero; since $\mathcal{X}_\delta$ is closed and convex, the projection is non-expansive and Proposition 7 continues to hold. This yields a second fairness-utility trade-off mechanism alongside $\lambda$, expressed as a tolerance on the mean gap.

### 4.4 Why Project at Every Step

Since the bounds in Section 3 rely exclusively on the final-layer representations, a single final projection layer is sufficient to yield $\delta = 0$. However, this alters the underlying optimization problem: (per-step) $\arg\min_{\mathbf{H}\in\mathcal{X}_0} f_\lambda(\mathbf{H}) \neq \mathbf{P}(\arg\min_{\mathbf{H}} f_\lambda(\mathbf{H}))$ (end-only), with distinct fixed points $\mathbf{H}^*_{\mathrm{per-step}} = \eta\gamma(\mathbf{I} - \mathbf{P}\mathbf{T}_{\eta,\lambda})^{-1}\mathbf{P}\mathbf{H}^{(0)}$ and $\mathbf{H}^*_{\mathrm{end-only}} = \eta\gamma\mathbf{P}(\mathbf{I} - \mathbf{T}_{\eta,\lambda})^{-1}\mathbf{H}^{(0)}$. While end-only projection just subtracts group means from the unconstrained optimum, per-step projection solves for the *optimal fair representation* - identifying the smoothest representation subject to equal group means. In Appendix F (Theorem 14), we make this contrast explicit via KKT conditions. Although both methods modify the original APPNP fixed point via optimal dual variables, end-only projection imposes a direct, graph-agnostic shift. By contrast, FairProp applies the graph-diffusion term $[\mathbf{I} - (1-\gamma)\mathbf{S} + 2\lambda\mathbf{M}]^{-1} = \sum_{j\geq 0}[(1-\gamma)\mathbf{S} - 2\lambda\mathbf{M}]^j$, ensuring a smoother correction. As illustrated, for the one dimensional case, in Proposition 15 (Appendix F.1), projecting only once imposes a coarse correction that fails to account for graph structure, resulting in utility being sacrificed almost entirely on the minority side. Hence, a criterion imposed for fairness reasons is itself satisfied unfairly (Remark 16).

## 5 Experimental Evaluation

**Baselines & Datasets.** In order to demonstrate the impact of introducing differentiable fairness-promoting layers in the standard APPNP architecture (Gasteiger et al., 2018), we add it as a fairness-agnostic baseline. We compare our proposed approach against several relevant baselines in fairness aware GNNs - spanning general purpose methods to more task-specialized ones. For all baselines, we use the authors' public implementations and reported hyperparameter settings where available; otherwise, we implement the models ourselves following the descriptions in their respective papers. All results are averaged over 20 independent experiments. Note that for link prediction, we also use methods originally proposed for node classification. We discard the classification head and replace it with an inner-product decoder applied to the learned node embeddings, optimizing

Table 1: **Node classification** results. Best and second best per column in **bold** and underlined, respectively, excluding APPNP. Dashed results indicate out-of-memory (OOM) or time-consuming experiments. All results are presented by $\times 100$. *FairProp is always the best in terms of fairness and among the best in terms of accuracy.*

| | Credit | | Recidivism | | Pokec-z | | Pokec-n | |
|---|---|---|---|---|---|---|---|---|
| | Acc ↑ | $\Delta_{\text{DP}}^{\text{NC}}$ ↓ | Acc ↑ | $\Delta_{\text{DP}}^{\text{NC}}$ ↓ | Acc ↑ | $\Delta_{\text{DP}}^{\text{NC}}$ ↓ | Acc ↑ | $\Delta_{\text{DP}}^{\text{NC}}$ ↓ |
| APPNP | 80.9 ±0.2 | 3.5 ±0.7 | 67.1 ±0.2 | 18.4 ±0.3† | 67.6 ±0.6 | 5.2 ±0.6 | 67.7 ±0.6 | 6.6 ±0.9 |
| FairProp (proj) | 80.8 ±0.2 | **0.4 ±0.3** | 66.8 ±0.3 | **3.5 ±0.4**† | 67.2 ±0.7 | **0.7 ±0.3** | 67.3 ±0.8 | **0.2 ±0.2** |
| FairGAT | 79.4 ±0.5 | 1.2 ±0.3 | 67.8 ±1.1 | 15.1 ±0.5 | **69.4 ±0.9** | 2.0 ±1.2 | **68.6 ±0.8** | 2.3 ±3.0 |
| FairVGNN | 80.0 ±0.3 | 0.8 ±0.1 | 65.5 ±2.3 | 19.5 ±0.4 | — | — | — | — |
| FMP | 81.2 ±0.1 | 3.5 ±0.3 | **69.8 ±0.3** | 19.1 ±0.3 | 65.8 ±0.4 | 3.9 ±0.3 | 65.3 ±0.5 | 1.7 ±0.5 |
| FairSAD | 78.1 ±0.5 | **0.4 ±0.5** | 63.4 ±3.2 | 10.0 ±3.8 | — | — | — | — |
| DAB-GNN | 79.6 ±1.1 | 4.6 ±2.0 | 63.3 ±6.3 | 8.7 ±6.3 | 66.5 ±2.0 | 3.9 ±3.3 | 65.1 ±1.3 | 4.6 ±3.8 |
| FairGP | **81.6 ±0.3** | 3.3 ±0.8 | 68.0 ±0.9 | 6.4 ±2.5 | — | — | — | — |

† Recidivism results for APPNP / FairProp (proj) use $\gamma = 0.3$; all other datasets use $\gamma = 0.1$.

Table 2: **Link prediction** results. Best and second best per column in **bold** and underlined, respectively, excluding APPNP. *FairProp yields the best utility-fairness trade-off, with near 0* $\Delta_{\text{DP}}^{\text{LP}}$.

| | Cora | | Citeseer | | Pubmed | |
|---|---|---|---|---|---|---|
| | AUC ↑ | $\Delta_{\text{DP}}^{\text{LP}}$ ↓ | AUC ↑ | $\Delta_{\text{DP}}^{\text{LP}}$ ↓ | AUC ↑ | $\Delta_{\text{DP}}^{\text{LP}}$ ↓ |
| APPNP | 0.908 ±0.007 | 0.031 ±0.004 | 0.885 ±0.009 | 0.014 ±0.001 | 0.957 ±0.002 | 0.081 ±0.002 |
| FairProp (proj) | 0.897 ±0.009 | 0.003 ±0.001 | 0.875 ±0.009 | 0.002 ±0.000 | 0.942 ±0.003 | 0.003 ±0.001 |
| EDITS | 0.905 ±0.008 | 0.009 ±0.002 | 0.879 ±0.007 | 0.008 ±0.001 | 0.966 ±0.001 | 0.028 ±0.005 |
| FairGAT | 0.915 ±0.006 | 0.050 ±0.017 | 0.919 ±0.007 | 0.027 ±0.016 | 0.845 ±0.017 | 0.066 ±0.034 |
| FairAdj | 0.915 ±0.006 | 0.019 ±0.001 | 0.888 ±0.006 | 0.013 ±0.001 | **0.969 ±0.001** | 0.050 ±0.001 |
| FairDrop | 0.922 ±0.006 | 0.022 ±0.003 | 0.900 ±0.007 | 0.015 ±0.002 | 0.967 ±0.001 | 0.060 ±0.002 |
| FairWire | 0.811 ±0.013 | **0.000 ±0.000** | 0.817 ±0.013 | **0.000 ±0.000** | 0.917 ±0.002 | **0.001 ±0.001** |
| FairINV | **0.928 ±0.006** | 0.027 ±0.001 | **0.924 ±0.007** | 0.014 ±0.001 | 0.953 ±0.001 | 0.080 ±0.001 |
| FairSAD | 0.848 ±0.019 | 0.001 ±0.000 | 0.817 ±0.028 | 0.001 ±0.000 | 0.743 ±0.012 | **0.001 ±0.001** |
| FairGP | 0.746 ±0.021 | **0.000 ±0.000** | 0.865 ±0.010 | 0.001 ±0.000 | 0.835 ±0.009 | **0.001 ±0.001** |

binary cross-entropy loss. For node regression, since there is only one dedicated baseline, we adapt methods originally proposed for node classification, by replacing the final softmax prediction layer with a linear output layer and optimizing the mean-squared error loss. Owing to space limitations, we report only the best-performing baselines for each task; full experimental comparisons are provided in Appendix G. Details regarding dataset statistics, and our experimental setup are available in Appendices G.2 and G.1. We further study the effect of relaxing the exact projection to a slab of tolerance $\delta$ in Appendix G.3. A runtime comparison of FairProp against all baselines is provided in Appendix G.5. The implementation of our method is provided in the supplementary material.

### 5.1 Main Results

**Node Classification.** Table 1 reports test accuracy and $\Delta_{\text{DP}}^{\text{NC}}$ on four benchmarks. It is evident that the standard APPNP pipeline offers good predictive accuracy, but violates demographic parity to a large extent. In contrast, FairProp with exact subgroup mean-centering via projection, markedly reduces the demographic parity gap ($5-33\times$) while yielding a very small decrease in mean accuracy (at most $0.4\%$). Moreover, FairProp attains a lower demographic parity gap than *every* competing method on *every* dataset (tied with FairSAD on Credit), and is second best in terms of accuracy in 2 out of 4 datasets. It is worth pointing out that the baselines that surpass it in accuracy are significantly worse-off in terms of the fairness gap. Overall, FairProp exhibits the best utility-fairness trade-off.

**Link Prediction.** Table 2 reports AUC and $\Delta_{\text{DP}}^{\text{LP}}$ on three benchmarks. Relative to the APPNP backbone, FairProp with exact mean centering reduces the parity gap by $7-27\times$, to at most $0.003$, while retaining over $98\%$ of the backbone AUC on every dataset. No other method does better w.r.t both metrics. Those that improve upon AUC (FairGAT, FairAdj, FairDrop, FairINV) do so at the cost of a $3-27\times$ larger parity gap. Conversely, every method attaining a smaller gap than FairProp (FairWire, FairSAD, FairGP) pays for it in terms of AUC, losing up to $22\%$ of the backbone AUC. Restricting attention to methods that retain at least $98\%$ of the backbone AUC, FairProp attains the smallest gap on every dataset. FairProp best handles this trade-off because the mean-gap compo-

Table 3: **Node regression** results. Best and second best per column in **bold** and underlined, respectively, excluding APPNP. *FairProp is almost always the best in terms of both MAE and fairness.*

| | GermanCredit | | Pokec-z | | Pokec-n | |
|---|---|---|---|---|---|---|
| | MAE ↓ | $\Delta_{\text{MG}}$ ↓ | MAE ↓ | $\Delta_{\text{MG}}$ ↓ | MAE ↓ | $\Delta_{\text{MG}}$ ↓ |
| APPNP | 0.404 ±0.035 | 0.089 ±0.015 | 0.199 ±0.001 | 0.091 ±0.001† | 0.199 ±0.001 | 0.152 ±0.001† |
| FairProp (proj) | 0.481 ±0.027 | **0.000 ±0.000** | **0.200 ±0.001** | **0.000 ±0.000**† | **0.199 ±0.001** | **0.000 ±0.000**† |
| FairGAT | 0.442 ±0.012 | 0.017 ±0.012 | 0.652 ±0.052 | 0.061 ±0.007 | 0.654 ±0.042 | 0.094 ±0.008 |
| EDITS | 0.414 ±0.013 | 0.024 ±0.019 | 0.587 ±0.004 | 0.006 ±0.002 | 0.586 ±0.005 | 0.020 ±0.001 |
| FairVGNN | **0.407 ±0.024** | 0.049 ±0.021 | — | — | — | — |
| FairWalk | 0.445 ±0.008 | 0.017 ±0.012 | 0.478 ±0.004 | 0.055 ±0.001 | 0.472 ±0.004 | 0.100 ±0.002 |
| FMP | 0.558 ±0.006 | 0.074 ±0.008 | 0.204 ±0.002 | 0.096 ±0.001 | 0.201 ±0.002 | 0.154 ±0.001 |
| FnRGNN | 0.521 ±0.011 | 0.022 ±0.018 | 0.249 ±0.021 | 0.004 ±0.003 | 0.239 ±0.005 | 0.020 ±0.008 |

† Pokec-z/-n results for APPNP / FairProp (proj) use $\gamma = 0.5$; GermanCredit use $\gamma = 0.1$.

nent is removed exactly, resulting in almost zero parity gap, at almost no cost in expressiveness. This trade-off is visually illustrated in Figure 9 in Appendix G.6.2, where FairProp consistently sits closest to the ideal top-left corner across all datasets.

**Node Regression.** Table 3 reports Mean Absolute Error (MAE) and $\Delta_{\text{MG}}$ on three benchmarks. As established in Section 3, for a linear prediction layer (without activations), the mean-gap bound contains no variance term, so the mean-equalizing projection of FairProp eliminates $\Delta_{\text{MG}}$ *exactly*. This result is confirmed empirically, as FairProp achieves a mean gap of precisely 0 across all datasets. On Pokec-z and Pokec-n it matches the MAE of APPNP. On GermanCredit, the MAE increases from 0.404 to 0.481, which the slab relaxation partially recovers (Figure 6, Appendix G.3). FnRGNN, the only baseline specifically designed for fair node regression, does not attain 0 mean gap, while yielding a higher MAE on every dataset. Other baselines, which are adapted from node classification, struggle to reduce $\Delta_{\text{MG}}$ in the regression setting.

### 5.2 Ablation Study

**Effect of the within-group variance regularizer.** The projection alone already suffices for node classification and node regression (Section 5.1). In link prediction it attains a near-zero parity gap, but a few baselines reach smaller gaps, albeit at a large cost in AUC. We therefore test whether also increasing the variance term of Proposition 5, via the regularizer of Section 4.3, closes this gap. Figure 2 sweeps $\lambda$, normalized by $\lambda^* := \gamma\, n_{\min}/2$, the largest value for which Proposition 7 guarantees contraction at $\eta = 1$; the shaded region $\lambda > \lambda^*$ lies outside this regime. The disparity does fall to zero, but only under strong regularization, by which point AUC has collapsed. This is expected: once the projection has equalized the group means, shrinking the within-group variance to zero drives all embeddings toward a single point.

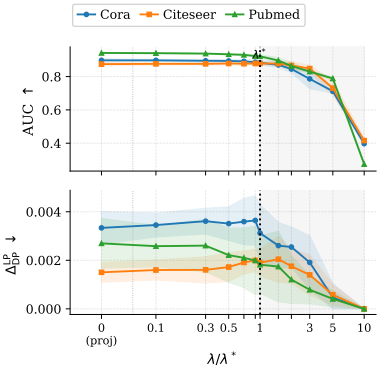



Figure 2: The effect of variance regularization in link prediction.

## 6 Conclusion

We studied fairness in GNNs at the level of the downstream prediction. We analyzed and bounded the demographic and mean parity gaps via the group-mean separation and the within-group covariance of the final node representations, tightening the prior bound for node classification, extending it to the deployed, sigmoid-activated prediction for link prediction, and providing the first such bound for node regression. Guided by this analysis, we proposed FairProp, which unfolds projected gradient descent on a fairness-constrained version of the APPNP objective. Each propagation layer of FairProp comprises a standard propagation step, regularized by within-group variance, followed by a closed-form projection that bounds group-mean separation, and the iteration provably converges linearly to the unique fair optimum. Empirically, using the projection alone to exactly equalize group means already achieves the lowest parity gap on node classification, a near-zero gap on link prediction, and a zero mean gap on node regression, while largely preserving utility against state-of-the-art baselines. Extending the framework to other unfolded GNN architectures and fairness notions, such as equal opportunity or individual fairness, are further directions for future work.

### AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing of the manuscript, research execution (implementation of experiments) and fact-checking proofs of mathematical claims. We have not used generative AI tools for retrieval and discovery, draft sections of the paper, and generating synthetic datasets are not applicable to this work. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

GNNs propagate information along edges that can perpetuate or amplify structural biases across vertex subgroups. This work advances the field of algorithmic fairness on graphs by analyzing where such bias enters the learned representations and by developing mechanisms that constrain it during propagation, for node classification, link prediction and node regression. Our aim is to reduce disparity in predictions while preserving the utility of the underlying model. While this work addresses group fairness, and in particular demographic parity, we acknowledge that it does not explicitly treat other relevant notions, such as individual, counterfactual or degree-based fairness. Hence, our results should be interpreted within this scope: the objective of this work is to improve the trade-off between group-fairness constraints and predictive utility, and our findings should not be directly extrapolated to other notions of fairness. We further note that enforcing a fairness constraint is not without cost when the sensitive attribute is correlated with the target, and that a small measured disparity does not by itself certify that a deployed system is equitable. We view our work as a step in this direction and anticipate future research extending the framework to additional fairness notions. We hope this contribution supports the development of more equitable graph-learning systems and motivates further research in algorithmic fairness for graph machine learning.

## REFERENCES

Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in artificial intelligence*, pp. 2114–2124. PMLR, 2021.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Heinz H Bauschke and Patrick L Combettes. Correction to: convex analysis and monotone operator theory in hilbert spaces. In *Convex analysis and monotone operator theory in Hilbert spaces*, pp. C1–C4. Springer, 2020.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *Advances in neural information processing systems*, 35:5230–5242, 2022.

Avishek Bose and William Hamilton. Compositional fairness constraints for graph embeddings. In *International conference on machine learning*, pp. 715–724. PMLR, 2019.

James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in Order Restricted Statistical Inference: Proceedings of the Symposium on Order Restricted Statistical Inference held in Iowa City, Iowa, September 11–13, 1985*, pp. 28–47. Springer, 1986.

Maarten Buyl and Tijl De Bie. Debayes: a bayesian method for debiasing network embeddings. In *International Conference on Machine Learning*, pp. 1220–1229. PMLR, 2020.

April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–23, 2024.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 680–688, 2021.

Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM web conference 2022*, pp. 1259–1269, 2022.

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602, 2023.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. *arXiv preprint arXiv:1810.05997*, 2018.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pp. 399–406, 2010.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *CoRR*, abs/1706.02216, 2017. URL http://arxiv.org/abs/1706.02216.

Bas Hofstra, Rense Corten, Frank Van Tubergen, and Nicole B Ellison. Sources of segregation in social networks: A novel approach using facebook. *American Sociological Review*, 82(3): 625–656, 2017.

Zhimeng Jiang, Xiaotian Han, Chao Fan, Zirui Liu, Na Zou, Ali Mostafavi, and Xia Hu. Chasing fairness in graphs: A gnn architecture perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21214–21222, 2024.

Kareem L Jordan and Tina L Freiburger. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3): 179–196, 2015.

Jian Kang, Yan Zhu, Yinglong Xia, Jiebo Luo, and Hanghang Tong. Rawlsgcn: Towards rawlsian difference principle on graph convolutional network. In *Proceedings of the ACM Web Conference 2022*, pp. 1214–1225, 2022.

Emmanouil Kariotakis and Aritra Konar. Fairrari: A plug and play framework for fairness-aware pagerank. *arXiv preprint arXiv:2602.08589*, 2026.

Ahmad Khajehnejad, Moein Khajehnejad, Mahmoudreza Babaei, Krishna P Gummadi, Adrian Weller, and Baharan Mirzasoleiman. Crosswalk: Fairness-enhanced node representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11963–11970, 2022.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. URL http://arxiv.org/abs/1609.02907.

O Deniz Kose and Yanning Shen. Fairgat: Fairness-aware graph attention networks. *ACM Transactions on Knowledge Discovery from Data*, 18(7):1–20, 2024a.

Oyku Deniz Kose and Yanning Shen. Fast&fair: Training acceleration and bias mitigation for gnns. *Transactions on Machine Learning Research*, 2023.

Oyku Deniz Kose and Yanning Shen. Fairwire: Fair graph generation. *Advances in neural information processing systems*, 37:124451–124478, 2024b.

Yeon-Chang Lee, Hojung Shin, and Sang-Wook Kim. Disentangling, amplifying, and debiasing: Learning disentangled representations for fair graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12013–12021, 2025.

Peizhao Li, Yifei Wang, Han Zhao, Pengyu Hong, and Hongfu Liu. On dyadic fairness: Exploring and mitigating bias in graph connections. In *International conference on learning representations*, 2021.

Donald Loveland, Jiayi Pan, Aaresh Farrokh Bhathena, and Yiyang Lu. Fairedit: Preserving fairness in graph neural networks through greedy graph editing. *arXiv preprint arXiv:2201.03681*, 2022.

Renqiang Luo, Huafei Huang, Ivan Lee, Chengpei Xu, Jianzhong Qi, and Feng Xia. Fairgp: A scalable and fair graph transformer using graph partitioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12319–12327, 2025.

Yao Ma, Xiaorui Liu, Tong Zhao, Yozen Liu, Jiliang Tang, and Neil Shah. A unified view on graph neural networks as graph signal denoising. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 1202–1211, 2021.

Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

SoYoung Park and Sungsu Lim. Fnrgnn: Distribution-aware fairness in graph neural network. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5110–5114, 2025.

Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. Fairwalk: Towards fair graph embedding. 2019.

Indro Spinelli, Simone Scardapane, Amir Hussain, and Aurelio Uncini. Fairdrop: Biased edge dropout for enhancing fairness in graph representation learning. *IEEE Transactions on Artificial Intelligence*, 3(3):344–354, 2021.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. URL https://arxiv.org/abs/1710.10903.

Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1938–1948, 2022.

Shaokui Wei, Jiayin Liu, Bing Li, and Hongyuan Zha. Mean parity fair regression in rkhs. In *International conference on artificial intelligence and statistics*, pp. 4602–4628. PMLR, 2023.

Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: a survey. *ACM computing surveys*, 55(5):1–37, 2022.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

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

Qi Yuan, Yang Liu, Yateng Tang, Xinhuan Chen, Xuehao Zheng, Qing He, and Xiang Ao. Dynamic graph learning with static relations for credit risk assessment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 13133–13141, 2025.

Meiqi Zhu, Xiao Wang, Chuan Shi, Houye Ji, and Peng Cui. Interpreting and unifying graph neural networks with an optimization framework. In *Proceedings of the web conference 2021*, pp. 1215–1226, 2021.

Yuchang Zhu, Jintang Li, Yatao Bian, Zibin Zheng, and Liang Chen. One fits all: Learning fair graph neural networks for various sensitive attributes. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4688–4699, 2024a.

Yuchang Zhu, Jintang Li, Zibin Zheng, and Liang Chen. Fair graph representation learning via sensitive attribute disentanglement. In *Proceedings of the ACM Web Conference 2024*, pp. 1182–1192, 2024b.

# A Extended Related Work

## A.1 Task-Agnostic Learning

A separate line of work aims to make the node embeddings themselves independent of the sensitive attribute. Bose & Hamilton (2019) train adversarial filters that remove sensitive information from graph embeddings. Their filters are compositional, so a single model can enforce invariance to any combination of sensitive attributes. DeBayes (Buyl & De Bie, 2020) builds on conditional network embeddings: during training, it encodes sensitive information in a biased structural prior, so that the embeddings only capture structure not explained by it. NIFTY (Agarwal et al., 2021) similarly regularizes representations to be stable under perturbations of the sensitive attribute. These methods target fairness at the representation level. However, they do not formally specify how it transfers to the predictions of a particular downstream task, which is precisely the gap our task-level analysis in Section 3 addresses.

## A.2 Node Classification

**In-processing.** A number of methods enforce group fairness during training. FairGNN (Dai & Wang, 2021) uses adversarial training to mask sensitive attributes under partial observability, while FairVGNN (Wang et al., 2022) adversarially masks sensitive-correlated feature channels. NIFTY (Agarwal et al., 2021) promotes counterfactual fairness via graph perturbations, and Fast&Fair (Kose & Shen, 2023) regularizes group-wise normalization operators for provably faster convergence. FairINV (Zhu et al., 2024a) recasts group fairness as invariant learning, inferring sensitive groups so that one training run is fair to several sensitive attributes without accessing them. These methods augment the task loss with a fairness term whose weight governs the trade-off, requiring per-dataset tuning without a principled selection procedure; adversarial variants additionally suffer from min-max instability.

**Interventions on the propagation operator.** A large family of methods mitigates bias by modifying how information travels along edges. FairWalk (Rahman et al., 2019) and CrossWalk (Khajehnejad et al., 2022) rebalance random-walk transition probabilities across sensitive groups, FairGAT (Kose & Shen, 2024a) down-weights attention on edges contributing to disparity, and FairDrop (Spinelli et al., 2021) biases edge dropout against homophilous connections. RawlsGCN (Kang et al., 2022) renormalizes the adjacency in the same spirit but targets a different notion, degree-based Rawlsian fairness, in which the sensitive attribute plays no role.

**Fairness inside the propagation operator.** Closest to our work, FMP (Jiang et al., 2024) also unfolds a fairness-regularized smoothness objective into message passing. However, it softly penalizes disparity in post-softmax probability space and is limited to node classification. In contrast, we constrain the representations themselves, drive the mean gap to exactly zero at every step, and prove linear convergence to the constrained optimum.

**Disentanglement.** Arguing that deleting sensitive-attribute information also deletes task-related information correlated with it, FairSAD (Zhu et al., 2024b) isolates the sensitive attribute into an independent representation channel and decorrelates it by channel masking, while DAB-GNN (Lee et al., 2025) disentangles attribute, structure and interaction biases into separate embeddings and aligns their subgroup distributions by Wasserstein distance.

**Pre-processing and graph editing.** EDITS (Dong et al., 2022) and FairEdit (Loveland et al., 2022) debias input graphs via feature and edge editing independently of downstream tasks. These require solving a combinatorial edge-selection problem, and since edges encode both task-relevant information and sensitive correlation, correcting the latter perturbs the former without explicit control over the trade-off.

## A.3 Link Prediction

Li et al. (2021) formalize *dyadic* fairness and propose FairAdj, which learns a fair adjacency matrix under a dyadic constraint; their representation-level analysis is the one our bounds extend. FairDrop (Spinelli et al., 2021) biases edge dropout against homophilous connections, and FairWire (Kose & Shen, 2024b) identifies structural sources of dyadic disparity to construct a regularizer

for link prediction and graph generation. Note that previously mentioned FairGAT (Kose & Shen, 2024a), is proposed to tackle also the link prediction task.

### A.4 Node Regression

Node regression is addressed only by FnRGNN (Park & Lim, 2025), which combines edge reweighting, representation alignment and prediction-level normalization to match the full group-conditional distributions.

### A.5 Representation-level bias analysis

Beyond its attention mechanism, FairGAT (Kose & Shen, 2024a) bounds the demographic parity gap of a GAT layer and of the fully connected readout typically used for node classification; our analysis tightens this bound. FairGP (Luo et al., 2025) bounds a related but distinct quantity for graph transformers, namely the distortion of the sensitive-attribute distribution under global attention. In (Kose & Shen, 2024b) they provide a bias analysis for link prediction. A closely related idea to ours, projecting onto a fairness constraint set at each iteration of a graph algorithm, has been applied to PageRank in the FairRARI framework (Kariotakis & Konar, 2026); we instead apply per-step projections inside end-to-end GNN training.

## B Bias Analysis Proofs

Adapting the definitions from the main text, we have that

$$\boldsymbol{\mu}_k := \mathbb{E}[\mathbf{h}_v \mid v \in \mathcal{S}_k] \quad \text{and} \quad \mathbf{C}_k := \mathrm{cov}(\mathbf{h}_v \mid v \in \mathcal{S}_k).$$

### B.1 Node Classification

For a node classification model, typically the final layer consists of a linear projection and a non-linear, $L$-Lipschitz function $\sigma(\cdot)$ (e.g. softmax, $L = 1/2$). Let's define the linear projection function $g(\cdot)$ as $g(v) := \mathbf{W}\mathbf{h}_v + \mathbf{b}$, and combining it with the non-linearity we obtain $f(v) := \sigma(\mathbf{W}\mathbf{h}_v + \mathbf{b})$.

**Proposition 8** (Linear readout)**.** *If the difference of the mean embeddings of the two subgroups $\mathcal{S}_0$ and $\mathcal{S}_1$ is small, i.e., $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 \leq \delta$, then*

$$\|\mathbb{E}[g(v)|v \in \mathcal{S}_0] - \mathbb{E}[g(v)|v \in \mathcal{S}_1]\|_2 \leq \|\mathbf{W}\|_2 \cdot \delta. \tag{11}$$

*Proof.* For $\|\mathbb{E}_{v\sim U}[\mathbf{h}_v|v \in S_0] - \mathbb{E}_{v\sim U}[\mathbf{h}_v|v \in S_1]\|_2 \leq \delta$, we have that

$$\begin{aligned} \|\mathbb{E}_{v\sim U}[g(v)|v \in S_0] - \mathbb{E}_{v\sim U}[g(v)|v \in S_1]\|_2 &= \|\mathbb{E}[\mathbf{W}\mathbf{h}_v + \mathbf{b}|v \in S_0] - \mathbb{E}[\mathbf{W}\mathbf{h}_v + \mathbf{b}|v \in S_1]\|_2 \\ &= \|\mathbf{W}(\mathbb{E}[\mathbf{h}_v|v \in S_0] - \mathbb{E}[\mathbf{h}_v|v \in S_1])\|_2 \\ &\leq \|\mathbf{W}\|_2 \|\mathbb{E}[\mathbf{h}_v|v \in S_0] - \mathbb{E}[\mathbf{h}_v|v \in S_1]\|_2 \\ &\leq \|\mathbf{W}\|_2 \cdot \delta \end{aligned} \tag{12}$$

□

**Proposition 9** (Linear readout - Multiple Groups)**.** *If the maximum difference of the mean embeddings of all the different subgroup pairs, i.e., $\max_{k,l}\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2 \leq \delta$, then*

$$\max_{k,l} \{\|\mathbb{E}[g(v)|v \in \mathcal{S}_k] - \mathbb{E}[g(v)|v \in \mathcal{S}_l]\|_2\} \leq \|\mathbf{W}\|_2 \cdot \delta. \tag{13}$$

*Proof.* The proof trivially follows from taking the max of each side in Proposition 8. □

**Proposition 10** (Non-linear readout)**.** *If $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 \leq \delta$, then*

$$\left\|\mathbb{E}[f(v)|v \in \mathcal{S}_0] - \mathbb{E}[f(v)|v \in \mathcal{S}_1]\right\|_2 \leq L\left(\|\mathbf{W}\|_2\,\delta + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_0\mathbf{W}^\top)} + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_1\mathbf{W}^\top)}\right). \tag{14}$$

*Proof.* For any $\mathbf{h}_v, \mathbf{h}_u \in \mathbb{R}^d$, the $L$-Lipschitzness of $\sigma$ and the definition of the spectral norm give

$$\|f(v)-f(u)\|_2 = \|\sigma(\mathbf{W}\mathbf{h}_v+\mathbf{b})-\sigma(\mathbf{W}\mathbf{h}_u+\mathbf{b})\|_2 \le L\|\mathbf{W}(\mathbf{h}_v-\mathbf{h}_u)\|_2 \le L\|\mathbf{W}\|_2\|\mathbf{h}_v-\mathbf{h}_u\|_2, \quad (15)$$

so $f$ is $L\|\mathbf{W}\|_2$-Lipschitz.

In order to introduce the expectations we have to do the following analysis

$$\begin{aligned}\|\mathbb{E}[f(v)|S_0]-\mathbb{E}[f(u)|S_1]\|_2 &= \|\mathbb{E}[f(v)|S_0]-f(\boldsymbol{\mu}_0)+f(\boldsymbol{\mu}_0)-f(\boldsymbol{\mu}_1)+f(\boldsymbol{\mu}_1)-\mathbb{E}[f(u)|S_1]\|_2 \\ &\le \underbrace{\|\mathbb{E}[f(v)|S_0]-f(\boldsymbol{\mu}_0)\|_2}_{(i)} + \underbrace{\|f(\boldsymbol{\mu}_0)-f(\boldsymbol{\mu}_1)\|_2}_{(ii)} + \underbrace{\|f(\boldsymbol{\mu}_1)-\mathbb{E}[f(u)|S_1]\|_2}_{(iii)}\end{aligned} \quad (16)$$

*Term $(ii)$:* By the $L$-Lipschitzness of $\sigma$ and the definition of the spectral norm,

$$\|f(\boldsymbol{\mu}_0)-f(\boldsymbol{\mu}_1)\|_2 \le L\|\mathbf{W}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\|_2 \le L\|\mathbf{W}\|_2\,\delta. \quad (17)$$

*Terms $(i)$ and $(iii)$:* For every $k \in \{0,1\}$, by Jensen's inequality for the convex map $\mathbf{y} \mapsto \|\mathbf{y}\|_2$, followed by the $L$-Lipschitzness of $\sigma$,

$$\begin{aligned}\big\|\mathbb{E}[f(v) \mid \mathcal{S}_k] - f(\boldsymbol{\mu}_k)\big\|_2 &= \big\|\mathbb{E}[\,f(v)-f(\boldsymbol{\mu}_k) \mid \mathcal{S}_k\,]\big\|_2 \le \mathbb{E}\big[\,\|f(v)-f(\boldsymbol{\mu}_k)\|_2 \,\big|\, \mathcal{S}_k\,\big] \\ &= \mathbb{E}\big[\,\|\sigma(\mathbf{W}\mathbf{h}_v+\mathbf{b})-\sigma(\mathbf{W}\boldsymbol{\mu}_k+\mathbf{b})\|_2 \,\big|\, \mathcal{S}_k\,\big] \\ &\le L\,\mathbb{E}\big[\,\|\mathbf{W}(\mathbf{h}_v-\boldsymbol{\mu}_k)\|_2 \,\big|\, \mathcal{S}_k\,\big].\end{aligned} \quad (18)$$

Applying Jensen's inequality once more, now for the concave map $t \mapsto \sqrt{t}$,

$$L\,\mathbb{E}\big[\,\|\mathbf{W}(\mathbf{h}_v-\boldsymbol{\mu}_k)\|_2 \,\big|\, \mathcal{S}_k\,\big] \le L\sqrt{\mathbb{E}\big[\,\|\mathbf{W}(\mathbf{h}_v-\boldsymbol{\mu}_k)\|_2^2 \,\big|\, \mathcal{S}_k\,\big]} = L\sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_k\mathbf{W}^\top)}, \quad (19)$$

where the equality holds because, by linearity of the trace and of the expectation,

$$\mathbb{E}\big[\|\mathbf{W}(\mathbf{h}_v-\boldsymbol{\mu}_k)\|_2^2 \,\big|\, \mathcal{S}_k\big] = \mathbb{E}\big[\mathrm{tr}\big(\mathbf{W}(\mathbf{h}_v-\boldsymbol{\mu}_k)(\mathbf{h}_v-\boldsymbol{\mu}_k)^\top\mathbf{W}^\top\big) \,\big|\, \mathcal{S}_k\big] = \mathrm{tr}(\mathbf{W}\mathbf{C}_k\mathbf{W}^\top). \quad (20)$$

Substituting equation 17, equation 18 and equation 19 into equation 16 yields equation 14. □

### B.1.1 Extension to Multiple Sensitive Groups (Proof of Proposition 2)

Proposition 10 trivially extends to any pair of subgroups $i, j \in \mathcal{K}$. Therefore, the following bound applies

$$\big\|\mathbb{E}[f(v) \mid \mathcal{S}_i] - \mathbb{E}[f(v) \mid \mathcal{S}_j]\big\|_2 \le L\Big(\|\mathbf{W}\|_2\,\delta_{ij} + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_i\mathbf{W}^\top)} + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_j\mathbf{W}^\top)}\Big), \quad (21)$$

with $\delta_{ij} := \|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2$. Since $\Delta_{\mathrm{DP}}$ is a maximum over pairs, taking the maximum of the right-hand side yields

$$\max_{i,j\in[K]} \big\|\mathbb{E}[f(v)|\mathcal{S}_i] - \mathbb{E}[f(v)\mathcal{S}_j]\big\|_2 \le L\Big(\|\mathbf{W}\|_2\,\delta + 2\max_{i\in[K]}\sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_i\mathbf{W}^\top)}\Big), \quad (22)$$

with $\delta := \max_{i,j\in[K]}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2$.

### B.1.2 Binary Classification (Proof of Corollary 3)

*Proof.* Since $\sigma(\cdot)$ is the softmax and the task is binary, each $f(v)$ is a probability vector over the two classes:

$$f(v) = \begin{bmatrix}\Pr(\hat{Y}(v)=0)\\ \Pr(\hat{Y}(v)=1)\end{bmatrix} = \begin{bmatrix}1-p_v\\ p_v\end{bmatrix}. \quad (23)$$

Taking the expectation uniformly over the nodes of group $\mathcal{S}_k$ gives

$$\mathbb{E}[f(v) \mid v \in \mathcal{S}_k] = \frac{1}{n_k}\sum_{v\in\mathcal{S}_k} f(v) = \begin{bmatrix}1-\bar{p}_k\\ \bar{p}_k\end{bmatrix}. \quad (24)$$

Therefore, for any $k, \ell \in [K]$,

$$\begin{aligned}\big\|\mathbb{E}[f(v) \mid v\in\mathcal{S}_k]-\mathbb{E}[f(v) \mid v\in\mathcal{S}_\ell]\big\|_2 &= \left\|\begin{bmatrix}(1-\bar{p}_k)-(1-\bar{p}_\ell)\\ \bar{p}_k-\bar{p}_\ell\end{bmatrix}\right\|_2 = \left\|(\bar{p}_k-\bar{p}_\ell)\begin{bmatrix}-1\\1\end{bmatrix}\right\|_2 \\ &= \sqrt{2}\,|\bar{p}_k-\bar{p}_\ell|.\end{aligned} \quad (25)$$

Taking the maximum over $k, \ell \in [K]$ on both sides yields

$$\max_{k,\ell\in[K]} \left\|\mathbb{E}[f(v) \mid v \in \mathcal{S}_k] - \mathbb{E}[f(v) \mid v \in \mathcal{S}_\ell]\right\|_2 = \sqrt{2} \max_{k,\ell\in[K]} |\bar{p}_k - \bar{p}_\ell|. \tag{26}$$

Substituting this identity into the left-hand side of Proposition 2 and dividing by $\sqrt{2}$ gives

$$\max_{k,\ell\in[K]} |\bar{p}_k - \bar{p}_\ell| \leq \frac{L}{\sqrt{2}} \Big( \|\mathbf{W}\|_2 \, \delta + 2 \max_{k\in[K]} \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_k\mathbf{W}^\top)} \Big), \tag{27}$$

which completes the proof. □

### B.2 Comparison of the Bounds (Proof of Proposition 4)

We adopt the setting of Section **??** with $\mathbf{b} = \mathbf{0}$, which cancels in the difference of the two conditional expectations, and write $\mathbf{z} := \mathbf{W}\mathbf{h}_v \in \mathbb{R}^c$ for the pre-activation output on a graph with $n$ nodes. In place of the covariances $\mathbf{C}_k$, FairGAT (Kose & Shen, 2024a) assumes a uniform bound on the deviation of each node from its group mean, $\|\mathbf{z}_j - \bar{\mathbf{z}}_s\|_\infty \leq \Delta_z^{(s)}$ for $s = 0, 1$, and sets $\Delta_z := \max_s \Delta_z^{(s)}$. Under the same hypothesis $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 \leq \delta$, the two upper bounds on the mean gap $\|\mathbb{E}[f(v) \,|\, v \in \mathcal{S}_0] - \mathbb{E}[f(v) \,|\, v \in \mathcal{S}_1]\|_2$ are

$$\begin{aligned} \mathcal{B}_{\text{ours}} &:= L \left( \|\mathbf{W}\|_2 \, \delta + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_0\mathbf{W}^\top)} + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_1\mathbf{W}^\top)} \right), \\ \mathcal{B}_{\text{FairGAT}} &:= L \left( \sigma_{\max}(\mathbf{W}) \, \delta + 2\sqrt{n}\Delta_z \right) = L \left( \|\mathbf{W}\|_2 \, \delta + 2\sqrt{n}\Delta_z \right). \end{aligned} \tag{28}$$

The two share an identical mean-gap term and differ only in how they account for the within-group spread, so the comparison of Proposition 4 reduces to the variance parts alone. Note that $\mathcal{B}_{\text{ours}}$ is based on Proposition 10 below, the special case of Proposition 2 for $K = 2$.

For an experimental comparison, please refer to Figure 10 in Appendix G.7.

*Proof.* Given a graph $\mathcal{G}$ with $n$ nodes, let's define $f(v) = \sigma(\mathbf{W}\mathbf{h}_v)$, where the function $\sigma(\cdot)$ is $L$-Lipschitz. Suppose that we have two groups of nodes $\mathcal{S}_0$ and $\mathcal{S}_1$, then $\boldsymbol{\mu}_0 = \mathbb{E}[\mathbf{h}_v \,|\, v \in \mathcal{S}_0]$, $\boldsymbol{\mu}_1 = \mathbb{E}[\mathbf{h}_v \,|\, v \in \mathcal{S}_1]$, $\mathbf{C}_0 = \mathrm{cov}(\mathbf{h}_v \,|\, v \in \mathcal{S}_0)$ and $\mathbf{C}_1 = \mathrm{cov}(\mathbf{h}_v \,|\, v \in \mathcal{S}_1)$, and define $\delta := \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$.
Also, define $\mathbf{z} = \mathbf{W}\mathbf{h}_v$, $\Delta_z = \max(\Delta_z^{(0)}, \Delta_z^{(1)})$, $\|\mathbf{z}_j - \bar{\mathbf{z}}_s\|_\infty \leq \Delta_z^{(s)}$, $s = 0, 1$.
Then, we have the following upper bounds, for

$$\|\mathbb{E}[f(v) \,|\, v \in \mathcal{S}_0] - \mathbb{E}[f(v) \,|\, v \in \mathcal{S}_1]\|_2 \tag{29}$$

• Our bound:

$$\mathcal{B}_{\text{ours}} := L\|\mathbf{W}\|_2\delta + \underbrace{L\left(\sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_0\mathbf{W}^\top)} + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_1\mathbf{W}^\top)}\right)}_{\mathcal{D}_{\text{ours}}} \tag{30}$$

• FairGAT bound:

$$\mathcal{B}_{\text{FairGAT}} := L\left(\sigma_{\max}(\mathbf{W}) \cdot \delta + 2\sqrt{n}\Delta_z\right) = L\|\mathbf{W}\|_2\delta + \underbrace{2L\sqrt{n}\Delta_z}_{\mathcal{D}_{\text{FairGAT}}} \tag{31}$$

We notice that the first parts of the bounds are the same, so we only have to compare $\mathcal{D}_{\text{ours}}$ with $\mathcal{D}_{\text{FairGAT}}$.

The inequality $\mathcal{D}_{\text{ours}} \leq \mathcal{D}_{\text{FairGAT}}$ is equivalent to

$$\left(\sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_0\mathbf{W}^\top)} + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_1\mathbf{W}^\top)}\right) \;\leq\; 2\sqrt{n}\,\Delta_z, \tag{32}$$

where we canceled the common factor $L > 0$. Solving for $n$ (both sides are nonnegative, so we may square), yields

$$n \;\geq\; n^* := \frac{\left(\sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_0\mathbf{W}^\top)} + \sqrt{\mathrm{tr}(\mathbf{W}\mathbf{C}_1\mathbf{W}^\top)}\right)^2}{4\,\Delta_z^2}. \tag{33}$$

Let $c$ denote the output dimension, i.e. $\mathbf{z} \in \mathbb{R}^c$. The trace of the $z$-space covariance is the expected squared $\ell_2$ deviation around the group mean $\bar{\mathbf{z}}_s = \mathbf{W}\boldsymbol{\mu}_s$:

$$\operatorname{tr}(\mathbf{W}\mathbf{C}_s\mathbf{W}^\top) = \mathbb{E}\big[\|\mathbf{z} - \bar{\mathbf{z}}_s\|_2^2 \,\big|\, s\big], \qquad s = 0, 1. \tag{34}$$

The assumption $\|\mathbf{z}_j - \bar{\mathbf{z}}_s\|_\infty \le \Delta_z^{(s)} \le \Delta_z$ bounds each of the $c$ coordinates, hence

$$\|\mathbf{z}_j - \bar{\mathbf{z}}_s\|_2^2 = \sum_{k=1}^{c} \big(z_{j,k} - \bar{z}_{s,k}\big)^2 \;\le\; \sum_{k=1}^{c} \Delta_z^2 = c\,\Delta_z^2. \tag{35}$$

Taking the conditional expectation gives, for $s = 0, 1$,

$$\mathbb{E}\big[\|\mathbf{z} - \bar{\mathbf{z}}_s\|_2^2 \,\big|\, s\big] \le c\,\Delta_z^2 \Rightarrow \operatorname{tr}(\mathbf{W}\mathbf{C}_s\mathbf{W}^\top) \;\le\; c\,\Delta_z^2 \Rightarrow \sqrt{\operatorname{tr}(\mathbf{W}\mathbf{C}_s\mathbf{W}^\top)} \;\le\; \sqrt{c}\,\Delta_z. \tag{36}$$

Substituting this into $n^*$,

$$n^* \;\le\; \frac{\big(\sqrt{c}\,\Delta_z + \sqrt{c}\,\Delta_z\big)^2}{4\,\Delta_z^2} = \frac{\big(2\sqrt{c}\,\Delta_z\big)^2}{4\,\Delta_z^2} = c. \tag{37}$$

Therefore $c \ge n^*$, and the condition $n \ge n^*$ is guaranteed whenever the number of nodes exceeds the output dimension, $n \ge c$. In summary, for $n \ge c$ we obtain

$$\mathcal{B}_{\text{ours}} \;\le\; \mathcal{B}_{\text{FairGAT}}. \tag{38}$$

□

### B.3 Link Prediction

For a link prediction model, typically the final layer consists of an inner product and a non-linear, $L$-Lipschitz function $\sigma(\cdot)$ (e.g. sigmoid, $L = 1/4$). Let's define the inner product function $g(\cdot, \cdot)$ as $g(v, u) = \mathbf{h}_v^\top \Sigma \mathbf{h}_u$, where $\Sigma \in \mathbb{S}_{++}^d$ is a positive-definite matrix, and combining it with the non-linearity we obtain $f(v, u) := \sigma(\mathbf{h}_v^\top \Sigma \mathbf{h}_u)$. Also, assume that for each $v \in \mathcal{V}$, there is a scalar $Q > 0$, such that $\|\mathbf{h}_v\|_2 \le Q$.

**Proposition 11** (Linear readout - Proposition 4.1. of (Li et al., 2021)). *If $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 \le \delta$, then*

$$|\mathbb{E}[g(v, u)|\mathcal{S}(v) = \mathcal{S}(u)] - \mathbb{E}[g(v, u)|\mathcal{S}(v) \ne \mathcal{S}(u)]| \le Q\|\Sigma\|_2\delta. \tag{39}$$

The extension of Proposition 11 for multiple sensitive subgroups ($K > 2$), can be found in (Kose & Shen, 2024b, Proposition 1).

**Proposition 12** (Non-linear readout). *If $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 \le \delta$, then*

$$\begin{aligned} &\|\mathbb{E}[f(v, u)|\mathcal{S}(v) = \mathcal{S}(u)] - \mathbb{E}[f(v, u)|\mathcal{S}(v) \ne \mathcal{S}(u)]\| \\ &\le LQ\|\Sigma\|_2 \left(\delta + (2\alpha_0 + 1)\sqrt{\operatorname{tr}(\mathbf{C}_0)} + (2\alpha_1 + 1)\sqrt{\operatorname{tr}(\mathbf{C}_1)}\right) \end{aligned} \tag{40}$$

*where $\alpha_k = \frac{n_k^2}{n_0^2 + n_1^2}$.*

*Proof.* To simplify the notations, let $\boldsymbol{\mu}_0 := \mathbb{E}_{v\sim U}[v|v \in S_0]$ and $\boldsymbol{\mu}_1 := \mathbb{E}_{u\sim U}[u|u \in S_1]$ denote the mean representations. Thus, $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 \le \delta$.
Let

$$\begin{aligned} \mathbb{E}_{\text{intra}} &:= \mathbb{E}[f(v, u)|S(v) = S(u)] = \mathbb{E}[f(v, u)|v \in S_0, u \in S_0 \vee v \in S_1, u \in S_1]| \\ &= \frac{|S_0|^2}{|S_0|^2 + |S_1|^2}\mathbb{E}[f(v, u)|v, u \in S_0] + \frac{|S_1|^2}{|S_0|^2 + |S_1|^2}\mathbb{E}[f(v, u)|v, u \in S_1] \\ &= \alpha_0 \cdot \mathbb{E}[f(v, u)|v, u \in S_0] + \alpha_1 \cdot \mathbb{E}[f(v, u)|v, u \in S_1] \end{aligned} \tag{41}$$

and

$$\mathbb{E}_{\text{inter}} := \mathbb{E}[f(v, u)|S(v) \ne S(u)] = \mathbb{E}[f(v, u)|v \in S_0, u \in S_1]. \tag{42}$$

To bridge the expectations with the mean representations, we define three mean-based variables:

$$M_{00} := \sigma(\boldsymbol{\mu}_0^\top \Sigma \boldsymbol{\mu}_0) \quad M_{11} := \sigma(\boldsymbol{\mu}_1^\top \Sigma \boldsymbol{\mu}_1) \quad M_{01} := \sigma(\boldsymbol{\mu}_0^\top \Sigma \boldsymbol{\mu}_1) \tag{43}$$

Using the triangle inequality, we can bound $\Delta_{\mathrm{DP}}$ by separating the variance within the groups from the gap between the group means:

$$\begin{aligned}\Delta_{\mathrm{DP}} &= |\mathbb{E}_{\mathrm{intra}} - \mathbb{E}_{\mathrm{inter}}| \\ &= |\alpha_0 \cdot \mathbb{E}[f(v,u)|v,u \in S_0] + \alpha_1 \cdot \mathbb{E}[f(v,u)|v,u \in S_1] - \mathbb{E}[f(v,u)|v \in S_0, u \in S_1]| \\ &= |\alpha_0 \mathbb{E}[f(v,u)|S_0,S_0] - \alpha_0 M_{00} + \alpha_0 M_{00} \\ &\quad + \alpha_1 \mathbb{E}[f(v,u)|S_1,S_1] - \alpha_1 M_{11} + \alpha_1 M_{11} \\ &\quad - \mathbb{E}[f(v,u)|S_0,S_1] + M_{01} - M_{01}| \\ &\leq \alpha_0 |\mathbb{E}[f(v,u)|S_0,S_0] - M_{00}| + \alpha_1 |\mathbb{E}[f(v,u)|S_1,S_1] - M_{11}| \\ &\quad + |\mathbb{E}[f(v,u)|S_0,S_1] - M_{01}| + |\alpha_0 M_{00} + \alpha_1 M_{11} - M_{01}|\end{aligned} \tag{44}$$

**1. Bounding the Mean Group Gap** (Term 4):
Since $\alpha_0 + \alpha_1 = 1$, we can rewrite the last term:

$$\alpha_0 M_{00} + \alpha_1 M_{11} - M_{01} = \alpha_0(M_{00} - M_{01}) + \alpha_1(M_{11} - M_{01}). \tag{45}$$

Using the Lipschitz property of the non-linear $\sigma()$ and Cauchy-Schwarz:

$$\begin{aligned}|M_{00} - M_{01}| &= |\sigma(p^\top \Sigma p) - \sigma(p^\top \Sigma q)| \\ &\leq L|p^\top \Sigma (p-q)| \leq L\|p\|_2 \|\Sigma\|_2 \|p-q\|_2 \leq LQ\|\Sigma\|_2 \delta.\end{aligned} \tag{46}$$

Similarly, $|M_{11} - M_{01}| \leq LQ\|\Sigma\|_2\delta$. Thus, the total mean gap is bounded by:

$$\begin{aligned}|\alpha_0 M_{00} + \alpha_1 M_{11} - M_{01}| &= |\alpha_0(M_{00} - M_{01}) + \alpha_1(M_{11} - M_{01})| \\ &\leq LQ\|\Sigma\|_2 \delta(\alpha_0 + \alpha_1) = LQ\|\Sigma\|_2\delta.\end{aligned} \tag{47}$$

**2. Bounding the Representation Spread** (Terms 1, 2, and 3):
We now evaluate the distance between the expectation of the function and the function of the expectation. For $S_0$ (Jensen's and then Lipschitz for each term of the expectations - thus also for the expectation):

$$\begin{aligned}|\mathbb{E}[\sigma(v^\top \Sigma u)|S_0,S_0] - M_{00}| &\leq \mathbb{E}[|\sigma(v^\top \Sigma u) - \sigma(p^\top \Sigma p)||S_0,S_0] \\ &\leq L\mathbb{E}[|v^\top \Sigma u - p^\top \Sigma p||S_0,S_0]\end{aligned} \tag{48}$$

We decompose the inner product difference:

$$|v^\top \Sigma u - p^\top \Sigma p| = |(v-p)^\top \Sigma u + p^\top \Sigma (u-p)| \leq Q\|\Sigma\|_2(\|v-p\|_2 + \|u-p\|_2) \tag{49}$$

Taking the expectation over independent variables $v, u \in S_0$,
and using Jensen's inequality $\mathbb{E}[\|v-p\|_2|S_0] \leq \sqrt{\mathrm{tr}(\mathbf{C}_0)}$, we get:

$$\alpha_0 |\mathbb{E}[f(v,u)|S_0,S_0] - M_{00}| \leq \alpha_0 LQ\|\Sigma\|_2(2\sqrt{\mathrm{tr}(\mathbf{C}_0)}) = \alpha_0 2LQ\|\Sigma\|_2\sqrt{\mathrm{tr}(\mathbf{C}_0)} \tag{50}$$

Similarly, for $S_1$:

$$\alpha_1 |\mathbb{E}[f(v,u)|S_1,S_1] - M_{11}| \leq \alpha_1 2LQ\|\Sigma\|_2\sqrt{\mathrm{tr}(\mathbf{C}_1)} \tag{51}$$

For the inter-group variance ($v \in S_0, u \in S_1$):

$$\begin{aligned}|\mathbb{E}[f(v,u)|S_0,S_1] - M_{01}| &\leq L\mathbb{E}[|v^\top \Sigma u - p^\top \Sigma q||S_0,S_1] \\ &= L\mathbb{E}[|v^\top \Sigma u - p^\top \Sigma u + p^\top \Sigma u - p^\top \Sigma q||S_0,S_1] \\ &= L\mathbb{E}[|(v-p)^\top \Sigma u + p^\top \Sigma (u-q)||S_0,S_1] \\ &\leq LQ\|\Sigma\|_2 \left(\sqrt{\mathrm{tr}(\mathbf{C}_0)} + \sqrt{\mathrm{tr}(\mathbf{C}_1)}\right)\end{aligned} \tag{52}$$

**3. Final Combination**
Substituting all the bounded terms back into the triangle inequality for $\Delta_{\mathrm{DP}}$ and factoring out $LQ\|\Sigma\|_2$ yields the final bound:

$$\Delta_{\mathrm{DP}} \leq LQ\|\Sigma\|_2 \left(\delta + (2\alpha_0 + 1)\sqrt{\mathrm{tr}(\mathbf{C}_0)} + (2\alpha_1 + 1)\sqrt{\mathrm{tr}(\mathbf{C}_1)}\right) \tag{53}$$

completing the proof. □

### B.3.1 Extension to Multiple Sensitive Groups (Proof of Proposition 5)

*Proof.* To simplify the notations, let $\boldsymbol{\mu}_k := \mathbb{E}_{v\sim U}[v \mid v \in S_k]$ denote the mean representation of group $S_k$, for $k = 0, \ldots, K-1$. Thus, $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2 \le \delta$ for all $k, l$.
Let

$$\begin{aligned}\mathbb{E}_{\text{intra}} := \mathbb{E}[f(v,u) \mid S(v) = S(u)] &= \sum_{k=0}^{K-1} \frac{|S_k|^2}{\sum_j |S_j|^2}\, \mathbb{E}[f(v,u) \mid v,u \in S_k] \\ &= \sum_{k=0}^{K-1} \alpha_k \cdot \mathbb{E}[f(v,u) \mid v,u \in S_k]\end{aligned} \tag{54}$$

and

$$\begin{aligned}\mathbb{E}_{\text{inter}} := \mathbb{E}[f(v,u)\,|\,S(v) \neq S(u)] &= \sum_{k\neq l} \frac{|S_k||S_l|}{(\sum_j |S_j|)^2 - \sum_j |S_j|^2}\, \mathbb{E}[f(v,u)\,|\,v \in S_k, u \in S_l] \\ = \sum_{k\neq l} \beta_{kl} \cdot \mathbb{E}[f(v,u) \mid v \in S_k, u \in S_l],&\end{aligned} \tag{55}$$

where $\sum_k \alpha_k = 1$ and $\sum_{k\neq l} \beta_{kl} = 1$. To bridge the expectations with the mean representations, we define the mean-based variables

$$M_{kl} := \sigma(\boldsymbol{\mu}_k^\top \Sigma \boldsymbol{\mu}_l), \qquad 0 \le k, l \le K-1. \tag{56}$$

Using the triangle inequality, we can bound $\Delta_{\text{DP}}$ by separating the variance within the groups from the gap between the group means:

$$\begin{aligned}\Delta_{\text{DP}} &= |\mathbb{E}_{\text{intra}} - \mathbb{E}_{\text{inter}}| \\ &= \Big| \sum_k \alpha_k \mathbb{E}[f(v,u) \mid v,u \in S_k] - \sum_{k\neq l} \beta_{kl} \mathbb{E}[f(v,u) \mid v \in S_k, u \in S_l] \Big| \\ &= \Big| \sum_k \alpha_k \mathbb{E}[f(v,u) \mid S_k, S_k] - \sum_k \alpha_k M_{kk} + \sum_k \alpha_k M_{kk} \\ &\quad - \sum_{k\neq l} \beta_{kl} \mathbb{E}[f(v,u) \mid S_k, S_l] + \sum_{k\neq l} \beta_{kl} M_{kl} - \sum_{k\neq l} \beta_{kl} M_{kl} \Big| \\ &\le \sum_k \alpha_k \big|\mathbb{E}[f(v,u) \mid S_k, S_k] - M_{kk}\big| + \sum_{k\neq l} \beta_{kl} \big|\mathbb{E}[f(v,u) \mid S_k, S_l] - M_{kl}\big| \\ &\quad + \Big| \sum_k \alpha_k M_{kk} - \sum_{k\neq l} \beta_{kl} M_{kl} \Big|\end{aligned} \tag{57}$$

**1. Bounding the Mean Group Gap** (last term):
Since $\sum_k \alpha_k = 1$ and $\sum_{l\neq m} \beta_{lm} = 1$, we can rewrite the last term as a doubly-indexed average of differences:

$$\sum_k \alpha_k M_{kk} - \sum_{l\neq m} \beta_{lm} M_{lm} = \sum_k \sum_{l\neq m} \alpha_k \beta_{lm} \left(M_{kk} - M_{lm}\right). \tag{58}$$

Using the Lipschitz property of the non-linear $\sigma(\cdot)$ and Cauchy–Schwarz:

$$\begin{aligned}|M_{kk} - M_{lm}| = |\sigma(\boldsymbol{\mu}_k^\top \Sigma \boldsymbol{\mu}_k) - \sigma(\boldsymbol{\mu}_l^\top \Sigma \boldsymbol{\mu}_m)| &\le L \left|(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)^\top \Sigma \boldsymbol{\mu}_k + \boldsymbol{\mu}_l^\top \Sigma (\boldsymbol{\mu}_k - \boldsymbol{\mu}_m)\right| \\ &\le LQ\|\Sigma\|_2 \left(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2 + \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_m\|_2\right) \le 2LQ\|\Sigma\|_2 \delta.\end{aligned} \tag{59}$$

Note that when $k \in \{l, m\}$ one of the two differences vanishes and a single $\delta$ is paid; this is what happens for every pair when $K = 2$. Thus, the total mean gap is bounded by:

$$\Big| \sum_k \alpha_k M_{kk} - \sum_{l\neq m} \beta_{lm} M_{lm} \Big| \le 2LQ\|\Sigma\|_2 \delta \sum_k \sum_{l\neq m} \alpha_k \beta_{lm} = 2LQ\|\Sigma\|_2 \delta. \tag{60}$$

**2. Bounding the Representation Spread** (first two terms):
We now evaluate the distance between the expectation of the function and the function of the expectation. For a generic pair of groups $S_k, S_l$ (Jensen's and then Lipschitz for each term of the

expectations - thus also for the expectation):

$$\begin{aligned}|\mathbb{E}[\sigma(v^\top \Sigma u) \mid S_k, S_l] - M_{kl}| &\le \mathbb{E}[|\sigma(v^\top \Sigma u) - \sigma(\boldsymbol{\mu}_k^\top \Sigma \boldsymbol{\mu}_l)| \mid S_k, S_l] \\ &\le L\,\mathbb{E}[|v^\top \Sigma u - \boldsymbol{\mu}_k^\top \Sigma \boldsymbol{\mu}_l| \mid S_k, S_l]\end{aligned} \tag{61}$$

We decompose the inner product difference:

$$|v^\top \Sigma u - \boldsymbol{\mu}_k^\top \Sigma \boldsymbol{\mu}_l| = |(v-\boldsymbol{\mu}_k)^\top \Sigma u + \boldsymbol{\mu}_k^\top \Sigma (u - \boldsymbol{\mu}_l)| \le Q\|\Sigma\|_2 \left(\|v-\boldsymbol{\mu}_k\|_2 + \|u-\boldsymbol{\mu}_l\|_2\right) \tag{62}$$

Taking the expectation over independent variables $v \in S_k$, $u \in S_l$,
and using Jensen's inequality $\mathbb{E}[\|v-\boldsymbol{\mu}_k\|_2 \mid S_k] \le \sqrt{\mathrm{tr}(\mathbf{C}_k)}$, we get:

$$|\mathbb{E}[f(v,u) \mid S_k, S_l] - M_{kl}| \le LQ\|\Sigma\|_2 \left(\sqrt{\mathrm{tr}(\mathbf{C}_k)} + \sqrt{\mathrm{tr}(\mathbf{C}_l)}\right) \tag{63}$$

Applying equation 63 with $l = k$ for the intra-group terms:

$$\sum_k \alpha_k |\mathbb{E}[f(v,u) \mid S_k, S_k] - M_{kk}| \le 2LQ\|\Sigma\|_2 \sum_k \alpha_k \sqrt{\mathrm{tr}(\mathbf{C}_k)} \tag{64}$$

and with $l \neq k$ for the inter-group terms:

$$\begin{aligned}\sum_{k\neq l} \beta_{kl} |\mathbb{E}[f(v,u) \mid S_k, S_l] - M_{kl}| &\le LQ\|\Sigma\|_2 \sum_{k\neq l} \beta_{kl} \left(\sqrt{\mathrm{tr}(\mathbf{C}_k)} + \sqrt{\mathrm{tr}(\mathbf{C}_l)}\right) \\ &= LQ\|\Sigma\|_2 \sum_k w_k \sqrt{\mathrm{tr}(\mathbf{C}_k)},\end{aligned} \tag{65}$$

where we collected, for each group $k$, the total inter-group weight in which it appears:

$$w_k := \sum_{l\neq k} (\beta_{kl} + \beta_{lk}), \qquad \sum_k w_k = 2. \tag{66}$$

**3. Final Combination**
Substituting all the bounded terms back into the triangle inequality for $\Delta_{\mathrm{DP}}$:

$$\Delta_{\mathrm{DP}} \le 2LQ\|\Sigma\|_2 \delta + 2LQ\|\Sigma\|_2 \sum_k \alpha_k \sqrt{\mathrm{tr}(\mathbf{C}_k)} + LQ\|\Sigma\|_2 \sum_k w_k \sqrt{\mathrm{tr}(\mathbf{C}_k)} \tag{67}$$

Factoring out $LQ\|\Sigma\|_2$ yields the final bound:

$$\Delta_{\mathrm{DP}} \le LQ\|\Sigma\|_2 \left(2\delta + \sum_{k=0}^{K-1} (2\alpha_k + w_k) \sqrt{\mathrm{tr}(\mathbf{C}_k)}\right) \tag{68}$$

completing the proof. □

### B.4 The Effect of Non-Linearity

**Remark 13** (Non-linearity). *Taking* $\mathbf{W} = \mathbf{I}$ *in Proposition 2 gives, for* any *L-Lipschitz* $\sigma(\cdot)$ *acting on* any *intermediate representation,*

$$\max_{k,\ell} \|\mathbb{E}[\sigma(\mathbf{h}_v)|\mathcal{S}_k] - \mathbb{E}[\sigma(\mathbf{h}_v)|\mathcal{S}_\ell]\|_2 \le L(\delta + 2 \max_k \sqrt{\mathrm{tr}(\mathbf{C}_k)}). \tag{69}$$

The key difference in the bound for the non-linear readouts is the inclusion of the within-group variance terms introduced by the non-linearity. By contrast, the bound for the linear readouts depends purely on the group-conditional means. The variance terms constitute a *second* and *independent* source of fairness violation. In particular, equalizing just the group means may not be sufficient, even when $\delta = 0$, a large within-group variance can still allow a classifier to recover the sensitive attribute. Figure 3 illustrates two subgroups with identical means but different spreads remain easily separable, so a classifier acting on such representations can still predict the sensitive attribute.

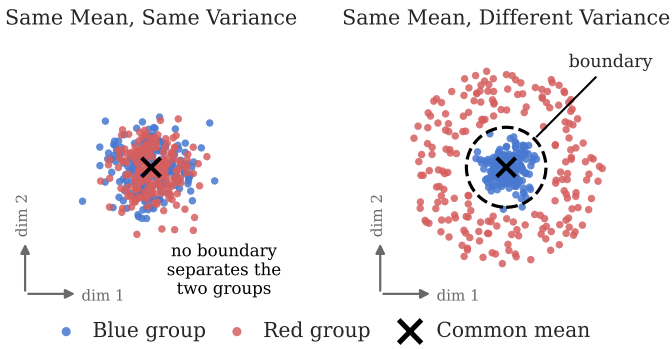



Figure 3: Illustrative example showing that mean matching is not enough, and that variance can still result in discrimination of sensitive attributes.

## C Exact Mean-Equalization Projection

For an embedding matrix $\mathbf{H} \in \mathbb{R}^{n \times d}$ ($n$ nodes, $d$ features), in order to zero out the group-conditional mean $\delta$ we need $\bar{\mathbf{h}}_{\mathcal{S}_i} = \bar{\mathbf{h}}_{\mathcal{S}_j}$, for every pair of subgroups $\mathcal{S}_i, \mathcal{S}_j$, where $\bar{\mathbf{h}}_{\mathcal{S}_i} = \frac{1}{|\mathcal{S}_i|}\mathbf{e}_{\mathcal{S}_i}^\top \mathbf{H}$, with $\mathbf{e}_{\mathcal{S}_i}$ being the indicator vector of subgroup $\mathcal{S}_i$. This is equivalent to $\bar{\mathbf{h}}_{\mathcal{S}_k} = \bar{\mathbf{h}}_{\mathcal{S}_0}$, for every $k \in [K] \setminus \{0\}$. Define the fairness constraint set

$$\mathcal{X}_0 = \{\mathbf{X} : \mathbf{a}_k^\top \mathbf{X} = \mathbf{0},\ k \in [K] \setminus \{0\}\}, \qquad \mathbf{a}_k := \frac{\mathbf{e}_{\mathcal{S}_0}}{n_0} - \frac{\mathbf{e}_{\mathcal{S}_k}}{n_k}. \tag{70}$$

Then, projecting an embedding matrix $\mathbf{H}$ onto $\mathcal{X}_0$, amounts to solving:

$$\min_{\mathbf{X} \in \mathcal{X}_0} \left\{ \|\mathbf{X} - \mathbf{H}\|_2^2 = \sum_{j=0}^{d-1} \|\mathbf{X}_{:,j} - \mathbf{H}_{:,j}\|_2^2 \right\} \tag{71}$$

with $\mathbf{X}_{:,j}$ denoting the $j$-th column of $\mathbf{X}$. The above is equivalent to solving the following for each column vector $\mathbf{h}$ of $\mathbf{H}$:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x} - \mathbf{h}\|_2^2 \\ \text{s.t.} \quad & \mathbf{a}_k^\top \mathbf{x} = 0,\ k \in [K] \setminus \{0\} \end{aligned} \tag{72}$$

Each $\mathbf{a}_k^\top \mathbf{x} = 0$ constraint is $\left(\frac{\mathbf{e}_{\mathcal{S}_0}}{|\mathcal{S}_0|} - \frac{\mathbf{e}_{\mathcal{S}_k}}{|\mathcal{S}_k|}\right)^\top \mathbf{x} = 0$.

Hence, the Lagrangian is:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \|\mathbf{x} - \mathbf{h}\|_2^2 + \sum_{k=1}^{K-1} \lambda_k \left(\frac{\mathbf{e}_{\mathcal{S}_0}}{|\mathcal{S}_0|} - \frac{\mathbf{e}_{\mathcal{S}_k}}{|\mathcal{S}_k|}\right)^\top \mathbf{x} \tag{73}$$

Using the KKT conditions, we have the following.

• Stationarity:

$$\begin{aligned} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}) = 0 \Leftrightarrow 2(\mathbf{x}^* - \mathbf{h}) + \sum_{k=1}^{K-1} \lambda_k \left(\frac{\mathbf{e}_{\mathcal{S}_0}}{|\mathcal{S}_0|} - \frac{\mathbf{e}_{\mathcal{S}_k}}{|\mathcal{S}_k|}\right) = 0 \\ \mathbf{x}^* = \mathbf{h} - \frac{1}{2} \sum_{k=1}^{K-1} \lambda_k \left(\frac{\mathbf{e}_{\mathcal{S}_0}}{|\mathcal{S}_0|} - \frac{\mathbf{e}_{\mathcal{S}_k}}{|\mathcal{S}_k|}\right) \end{aligned} \tag{74}$$

Let's define $\Lambda = \sum_{k=1}^{K-1} \lambda_k$, and express the above equality entry-wise:

$$\mathbf{x}_v^* = \begin{cases} \mathbf{h}_v - \frac{1}{2n_0}\Lambda & v \in \mathcal{S}_0 \\ \mathbf{h}_v + \frac{1}{2n_i}\lambda_i & v \in \mathcal{S}_i,\ i \neq 0 \end{cases} \tag{75}$$

Taking averages over each group:

$$\begin{aligned} \bar{\mathbf{x}}_{\mathcal{S}_0}^* &= \bar{\mathbf{h}}_{\mathcal{S}_0} - \frac{1}{2n_0}\Lambda \\ \bar{\mathbf{x}}_{\mathcal{S}_i}^* &= \bar{\mathbf{h}}_{\mathcal{S}_i} + \frac{1}{2n_i}\lambda_i \end{aligned} \tag{76}$$

• Primal Feasibility:

$$\bar{\mathbf{x}}^*_{\mathcal{S}_0} = \bar{\mathbf{x}}^*_{\mathcal{S}_i} \Leftrightarrow \bar{\mathbf{h}}_{\mathcal{S}_0} - \frac{1}{2n_0}\Lambda = \bar{\mathbf{h}}_{\mathcal{S}_i} + \frac{1}{2n_i}\lambda_i \Leftarrow \lambda_i = 2n_i(\bar{\mathbf{h}}_{\mathcal{S}_0} - \bar{\mathbf{h}}_{\mathcal{S}_i}) - \frac{n_i}{n_0}\Lambda \tag{77}$$

if we take the sums of both sides we get:

$$\begin{aligned}\sum_{i=1}^{K-1}\lambda_i = 2\sum_{i=1}^{K-1} n_i(\bar{\mathbf{h}}_{\mathcal{S}_0} - \bar{\mathbf{h}}_{\mathcal{S}_i}) - \sum_{i=1}^{K-1}\frac{n_i}{n_0}\Lambda \Leftrightarrow \frac{n}{n_0}\Lambda = 2\sum_{i=1}^{K-1} n_i(\bar{\mathbf{h}}_{\mathcal{S}_0} - \bar{\mathbf{h}}_{\mathcal{S}_i}) \\ \Lambda = 2n_0(\bar{\mathbf{h}}_{\mathcal{S}_0} - \bar{\mathbf{h}})\end{aligned} \tag{78}$$

replacing this to equation 77, we get:

$$\lambda_i = 2n_i(\bar{\mathbf{h}}_{\mathcal{S}_0} - \bar{\mathbf{h}}_{\mathcal{S}_i}) - \frac{n_i}{n_0}2n_0(\bar{\mathbf{h}}_{\mathcal{S}_0} - \bar{\mathbf{h}}) \Leftrightarrow \lambda_i = 2n_i(\bar{\mathbf{h}} - \bar{\mathbf{h}}_{\mathcal{S}_i}) \tag{79}$$

and then replacing this back to equation 75, we get:

$$\mathbf{x}^*_v = \begin{cases} \mathbf{h}_v - (\bar{\mathbf{h}}_{\mathcal{S}_0} - \bar{\mathbf{h}}) \\ \mathbf{h}_v + (\bar{\mathbf{h}} - \bar{\mathbf{h}}_{\mathcal{S}_i}) \quad v \in \mathcal{S}_i,\ i \neq 0 \end{cases} \tag{80}$$

Unifying both cases, we get:

$$\mathbf{x}^*_v = \mathbf{h}_v + \bar{\mathbf{h}} - \bar{\mathbf{h}}_{\mathcal{S}_i},\ v \in \mathcal{S}_i \tag{81}$$

▷ **Matrix Form:**
Define $\mathbf{M} := \sum_{k=0}^{K-1} \frac{\mathbf{e}_{\mathcal{S}_k}\mathbf{e}_{\mathcal{S}_k}^\top}{|\mathbf{e}_{\mathcal{S}_k}|} \in \mathbb{R}^{n\times n}$, then $\mathbf{M}\mathbf{h}_v = \bar{\mathbf{h}}_{\mathcal{S}_i},\ v \in \mathcal{S}_i$. Then, for the projection matrix $\mathbf{P} := \mathbf{I} + \frac{\mathbf{e}\mathbf{e}^\top}{n} - \mathbf{M}$, it holds that $\mathbf{x}^* = \mathbf{P}\cdot\mathbf{h}$.
It is easy to see that $\mathbf{P}$ is symmetric ($\mathbf{P} = \mathbf{P}^\top$). So, for the full embedding matrices, we get:

$$\mathbf{X}^* = \mathbf{P}\cdot\mathbf{H} \Rightarrow d\mathbf{H} = \mathbf{P}\cdot d\mathbf{X}. \tag{82}$$

Hence, *the projection of the backward pass is exactly the same as that of the forward pass.*

## D Relaxed Projection onto the Fairness Slab

The projection of Section 4.2 enforces $\delta = 0$ exactly. However, this can be restrictive, a small value of $\delta$ suffices to keep the fairness gap small, and forcing it to exact 0 can displace embeddings more than fairness demands. Hence, we consider the following relaxation of a *slab* of half-width $\delta$.

Let $\mathbf{a} := \mathbf{a}_1 = \frac{\mathbf{e}_{\mathcal{S}_0}}{n_0} - \frac{\mathbf{e}_{\mathcal{S}_1}}{n_1}$ be the group-contrast vector as in equation 5, and $\mathbf{s} := \mathbf{H}^\top\mathbf{a} \in \mathbb{R}^d$. The slab requires the group-mean gap to be at most $\delta$ (*fairness tolerance*) in every feature coordinate

$$\mathcal{X}_\delta := \big\{\mathbf{H} \in \mathbb{R}^{n\times d} : \|\mathbf{H}^\top\mathbf{a}\|_\infty \leq \delta\big\}, \tag{83}$$

and recovers the exact constraint set $\mathcal{X}_0$, at $\delta = 0$. It is an intersection of two half-spaces, of which at most one can be violated at a time, so the projection stays available in closed form,

$$\mathrm{Proj}_{\mathcal{X}_\delta}(\mathbf{H}) = \mathbf{H} - \frac{\mathbf{a}}{\|\mathbf{a}\|_2^2} g(\mathbf{s},\delta)^\top, \qquad g(\mathbf{s},\delta) := \mathrm{sign}(\mathbf{s}) \odot (|\mathbf{s}| - \delta\mathbf{1})_+, \tag{84}$$

where $(\mathbf{x})_+ = \max\{\mathbf{x}, \mathbf{0}\} = \mathrm{ReLU}(\mathbf{x})$, and $g$ acts entrywise, $[g(\mathbf{s},\delta)]_j = \mathrm{sign}(s_j)(|s_j| - \delta)_+$.

Unlike the exact projection, which applies $\mathbf{P}$ to every column, this map is *gated*, i.e., a column for which $|s_j| \leq \delta$ has $g(s_j,\delta) = 0$ and is returned unchanged, while only violating columns are shifted. Its backward pass is gated in the same way, $d\mathbf{h} = d\mathbf{h}^* - \mathbf{1}\big[|s| > \delta\big]\frac{\mathbf{a}\mathbf{a}^\top}{\|\mathbf{a}\|_2^2} d\mathbf{h}^*$, so gradients pass through unchanged for columns inside the slab and get shifted otherwise. For $K > 2$ the constraint set is the intersection of $K-1$ such slabs, hence a polytope on which equation 84 no longer applies directly. Then, the projection can be computed by Dykstra's algorithm (Boyle & Dykstra, 1986). Derivations are provided in the following subsections.

### D.1 FORWARD PASS

Given a vector $\mathbf{h} \in \mathbb{R}^n$, consider the slab defined by the constraint set $\mathcal{X}_\delta := \{\mathbf{h} \in \mathbb{R}^n : |\mathbf{a}^T\mathbf{h}| \leq \delta\}$. This for a matrix $\mathbf{H} \in \mathbb{R}^{n\times d}$ results to $\mathcal{X}_\delta := \{\mathbf{H} \in \mathbb{R}^{n\times d} : \|\mathbf{H}^\top\mathbf{a}\|_\infty \leq \delta\}$. The following analysis will be on the vector level, and in the end we will generalize it for an $n \times d$ matrix.

The problem of projecting a point $\mathbf{d} \in \mathbb{R}^n$ onto $\mathcal{X}_\delta$ is given by

$$\mathbf{\Pi}_{\mathcal{X}_\delta}(\mathbf{d}) := \arg\min_{\mathbf{h}\in\mathcal{X}_\delta} \frac{1}{2}\|\mathbf{h} - \mathbf{d}\|_2^2 \tag{85}$$

Before solving the problem, we first consider the simpler problem of projecting onto the half-space defined by $\mathcal{X}_\delta^+ := \{\mathbf{h} \in \mathbb{R}^n : \mathbf{a}^T\mathbf{h} \leq \delta\}$. The solution is

$$\mathbf{\Pi}_{\mathcal{X}_\delta^+}(\mathbf{d}) = \mathbf{d} - \frac{(\mathbf{a}^T\mathbf{d} - \delta)_+}{\|\mathbf{a}\|_2^2}\mathbf{a}, \tag{86}$$

where $(\mathbf{x})_+ = \max\{\mathbf{x}, \mathbf{0}\}$. In other words, we first check whether the point is in the halfspace, and if not, we project onto the hyperplane defining the boundary of the halfspace.

Note that the slab $\mathcal{X}_\delta$ is the intersection of the halfspaces $\mathcal{X}_\delta^+ := \{\mathbf{h} \in \mathbb{R}^n : \mathbf{a}^T\mathbf{h} \leq \delta\}$ and $\mathcal{X}_\delta^- := \{\mathbf{h} \in \mathbb{R}^n : -\mathbf{a}^T\mathbf{h} \leq \delta\}$. The projection is given by

$$\mathbf{\Pi}_{\mathcal{X}_\delta}(\mathbf{d}) = \mathbf{d} - \frac{(\mathbf{a}^T\mathbf{d} - \delta)_+}{\|\mathbf{a}\|_2^2}\mathbf{a} - \frac{(-\mathbf{a}^T\mathbf{d} - \delta)_+}{\|\mathbf{a}\|_2^2}\mathbf{a} = \mathbf{d} - [(\mathbf{a}^T\mathbf{d} - \delta)_+ - (-\mathbf{a}^T\mathbf{d} - \delta)_+]\frac{\mathbf{a}}{\|\mathbf{a}\|_2^2} \tag{87}$$

This stems from the fact that a point lying outside the slab already satisfies one of the halfspace constraints $\mathcal{X}_\delta^+$ or $\mathcal{X}_\delta^-$, and it remains to project the point onto the hyperplane defining the boundary of the halfspace of the other constraint. The result can be confirmed by analyzing the KKT conditions of problem equation 86.

**KKT analysis:** Let $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ denote the Lagrange multipliers of $\mathcal{X}_\delta^+$ and $\mathcal{X}_\delta^-$ respectively. The Lagrangian of equation 86 is given by

$$L(\mathbf{h}, \lambda_1, \lambda_2) = \frac{1}{2}\|\mathbf{h} - \mathbf{d}\|_2^2 + \lambda_1(\mathbf{a}^T\mathbf{h} - \delta) + \lambda_2(-\mathbf{a}^T\mathbf{h} - \delta) \tag{88}$$

The KKT conditions are then given by the complementary slackness conditions

$$\lambda_1(\mathbf{a}^T\mathbf{h} - \delta) = 0, \lambda_2(-\mathbf{a}^T\mathbf{h} - \delta) = 0, \tag{89}$$

and setting the gradient of the Lagrangian w.r.t. $\mathbf{h}$ equal to $\mathbf{0}$, which yields

$$\mathbf{h} - \mathbf{d} + \lambda_1\mathbf{a} - \lambda_2\mathbf{a} = \mathbf{0} \tag{90}$$

When $\lambda_1 = \lambda_2 = 0$, complementary slackness mandates that both halfspace constraints $\mathcal{X}_\delta^+$ and $\mathcal{X}_\delta^-$ are satisfied; i.e., $\mathbf{d} \in \mathcal{X}_\delta$, and the solution is $\mathbf{h} = \mathbf{d}$.

If $\lambda_1 > 0$ and $\lambda_2 = 0$, then $\mathcal{X}_\delta^-$ is satisfied but $\mathcal{X}_\delta^+$ is violated. Hence, we obtain $\mathbf{h} = \mathbf{d} - \lambda_1\mathbf{a}$. From complementary slackness, we have $\mathbf{a}^T\mathbf{h} = \delta$. Solving for $\lambda_1$, we obtain

$$\mathbf{h} = \mathbf{d} - \frac{(\mathbf{a}^T\mathbf{d} - \delta)}{\|\mathbf{a}\|_2^2}\mathbf{a} \tag{91}$$

If $\lambda_1 = 0$ and $\lambda_2 > 0$, then $\mathcal{X}_\delta^+$ is satisfied but $\mathcal{X}_\delta^-$ is violated. Solving for $\lambda_2$, we obtain

$$\mathbf{h} = \mathbf{d} - \frac{(\mathbf{a}^T\mathbf{d} + \delta)}{\|\mathbf{a}\|_2^2}\mathbf{a} \tag{92}$$

Note that the condition $\lambda_1 > 0$ and $\lambda_2 > 0$ is impossible, as it would mean that both halfpsace constraints are simultaneously violated. Hence, the overall solution is

$$\mathbf{h} = \mathbf{d} - \frac{(\mathbf{a}^T\mathbf{d} - \delta)_+}{\|\mathbf{a}\|_2^2}\mathbf{a} - \frac{(-\mathbf{a}^T\mathbf{d} - \delta)_+}{\|\mathbf{a}\|_2^2}\mathbf{a} = \mathbf{d} - [(\mathbf{a}^T\mathbf{d} - \delta)_+ - (-\mathbf{a}^T\mathbf{d} - \delta)_+]\frac{\mathbf{a}}{\|\mathbf{a}\|_2^2} \tag{93}$$

### D.2 Backward Pass

Recall from the forward analysis that the projection of a single column $\mathbf{h}$ onto $\mathcal{X}_\delta$ is

$$\mathbf{x} = \mathbf{\Pi}_{\mathcal{X}_\delta}(\mathbf{h}) = \mathbf{h} - \big[(\mathbf{a}^\top \mathbf{h} - \delta)_+ - (-\mathbf{a}^\top \mathbf{h} - \delta)_+\big] \frac{\mathbf{a}}{\|\mathbf{a}\|_2^2}. \tag{94}$$

Writing $s := \mathbf{a}^\top \mathbf{h}$ and collapsing the two hinges into a single map, this is soft-thresholding at $\delta$:

$$\mathbf{x} = \mathbf{h} - g(s, \delta)\, \frac{\mathbf{a}}{\|\mathbf{a}\|_2^2}, \qquad g(s, \delta) := (s - \delta)_+ - (-s - \delta)_+ = \operatorname{sign}(s)\,(|s| - \delta)_+. \tag{95}$$

▷ **Jacobian.** Differentiating and using $\frac{\partial s}{\partial \mathbf{h}} = \mathbf{a}^\top$ gives, and $g'(s, \delta) = \mathbf{1}[\,|s| > \delta\,]$ we get,

$$\mathbf{J} := \frac{\partial \mathbf{x}}{\partial \mathbf{h}} = \mathbf{I} - \mathbf{1}[\,|s| > \delta\,]\, \mathbf{P}_\mathbf{a}, \qquad \mathbf{P}_\mathbf{a} := \frac{\mathbf{a}\mathbf{a}^\top}{\|\mathbf{a}\|_2^2}, \tag{96}$$

that is,

$$\mathbf{J} = \begin{cases} \mathbf{I}, & |s| \le \delta \quad \text{(inside the slab)}, \\ \mathbf{I} - \mathbf{P}_\mathbf{a}, & |s| > \delta \quad \text{(outside)}. \end{cases} \tag{97}$$

▷ **Backward pass.** As $\mathbf{P}_\mathbf{a}$ is symmetric, $\mathbf{J} = \mathbf{J}^\top$, so calculating the gradient

$$d\mathbf{h} = \mathbf{J}^\top d\mathbf{x} = d\mathbf{x} - \mathbf{1}[\,|s| > \delta\,]\, \frac{\mathbf{a}\mathbf{a}^\top}{\|\mathbf{a}\|_2^2} d\mathbf{x}. \tag{98}$$

### D.3 Matrix Form

For $\mathbf{H} \in \mathbb{R}^{n \times d}$, let $\mathbf{s} = \mathbf{H}^\top \mathbf{a} \in \mathbb{R}^d$, and let $\mathbf{S}_{\text{diag}} = \operatorname{diag}\big(\mathbf{1}[\,|\mathbf{s}| > \delta\,]\big) \in \{0, 1\}^{d \times d}$ be the diagonal active-set matrix over columns. For a vector argument $\mathbf{s} \in \mathbb{R}^d$, $g$ acts entrywise, $[g(\mathbf{s}, \delta)]_j := \operatorname{sign}(s_j)\,(|s_j| - \delta)_+$, equivalently

$$g(\mathbf{s}, \delta) = \operatorname{sign}(\mathbf{s}) \odot \big(|\mathbf{s}| - \delta \mathbf{1}\big)_+, \tag{99}$$

the element-wise soft-thresholding operator ($\odot$: element-wise multiplication). Then,

$$\text{Forward:} \quad \mathbf{X} = \mathbf{H} - \frac{\mathbf{a}}{\|\mathbf{a}\|_2^2} g(\mathbf{s}, \delta)^\top, \tag{100}$$

$$\text{Backward:} \quad d\mathbf{H} = d\mathbf{X} - \frac{\mathbf{a}\mathbf{a}^\top}{\|\mathbf{a}\|_2^2} d\mathbf{X} \cdot \mathbf{S}_{\text{diag}}. \tag{101}$$

### D.4 Multiple Sensitive Subgroups

For $K > 2$, the relaxed constraint set $\mathcal{X}_\delta$ is the intersection of $K - 1$ slabs of the two-group form, $\mathcal{X}_\delta := \bigcap_{k=1}^{K-1} \mathcal{X}_\delta^k$, where $\mathcal{X}_\delta^k := \{\mathbf{h} : |\mathbf{h}^\top \mathbf{a}_k| \le \delta\}$, with $\mathbf{a}_k$ as defined in equation 5. For $K = 3$, this constrains the gaps between $\mathcal{S}_0 - \mathcal{S}_1$ and $\mathcal{S}_0 - \mathcal{S}_2$. While each individual slab $\mathcal{X}_\delta^k$ admits the closed-form projection equation 84, their intersection $\mathcal{X}_\delta$ forms a polytope. Because multiple constraints can be violated simultaneously, equation 84 does not apply directly. However, since each $\mathcal{X}_\delta^k$ is closed, convex, and contains $\mathbf{0}$, the projection $\operatorname{Proj}_{\mathcal{X}_\delta}$ can be computed via Dykstra's algorithm (Boyle & Dykstra, 1986) by cycling through the slabs with individual correction terms. For the backward pass, we can differentiate though the KKT conditions of the problem using implicit differentiation (Blondel et al., 2022).

## E Variance Control Inside the Diffusion (Proof of Proposition 7)

*Proof.* Let $\Phi(\mathbf{H}) := \mathbf{P}[\mathbf{T}_{\eta,\lambda}\mathbf{H} + \eta\gamma\mathbf{H}^{(0)}] = \operatorname{Proj}_{\mathcal{X}}(\mathbf{T}_{\eta,\lambda}\mathbf{H} + \eta\gamma\mathbf{H}^{(0)})$ denote the update of equation 9. Because $\mathcal{X}$ is a closed convex set, the projection on it is non-expansive, hence we have that

$$\begin{aligned} &\|\Phi(\mathbf{H}) - \Phi(\mathbf{Y})\|_F = \|\operatorname{Proj}_{\mathcal{X}}(\mathbf{T}_{\eta,\lambda}\mathbf{H} + \eta\gamma\mathbf{H}^{(0)}) - \operatorname{Proj}_{\mathcal{X}}(\mathbf{T}_{\eta,\lambda}\mathbf{Y} + \eta\gamma\mathbf{H}^{(0)})\|_F \\ &\le \|(\mathbf{T}_{\eta,\lambda}\mathbf{H} + \eta\gamma\mathbf{H}^{(0)}) - (\mathbf{T}_{\eta,\lambda}\mathbf{Y} + \eta\gamma\mathbf{H}^{(0)})\|_F = \|\mathbf{T}_{\eta,\lambda}(\mathbf{H} - \mathbf{Y})\|_F \\ &\le \|\mathbf{T}_{\eta,\lambda}\|_2 \|\mathbf{H} - \mathbf{Y}\|_F. \end{aligned} \tag{102}$$

We know that $\mathbf{T}_{\eta,\lambda} = \mathbf{I} - \eta\mathbf{Q}_\lambda$ is a contractive operator iff for all eigenvalues $\mu$ of $\mathbf{Q}_\lambda$ we get

$$|1 - \eta\mu| < 1 \Leftrightarrow 0 < \eta\mu < 2. \tag{103}$$

Hence, it should hold that

$$0 < \eta < \frac{2}{\boldsymbol{\mu}_{\max}(\mathbf{Q}_\lambda)}. \tag{104}$$

We know that the Hessian of $f_\lambda$ is $\mathbf{Q}_\lambda = \gamma\mathbf{I} + (1-\gamma)\mathbf{L}_\mathbf{S} + 2\lambda\mathbf{M}$ and is symmetric, so its eigenvalues are real. We also know that, $\mathbf{L}_\mathbf{S} \succeq 0$ and $\mathbf{M} \succeq 0$ give $\mathbf{Q}_\lambda \succeq \gamma\mathbf{I}$, hence $\mu \geq \gamma = \boldsymbol{\mu}_{\min}(\mathbf{Q}_\lambda) > 0$ for every eigenvalue $\mu$ of $\mathbf{Q}_\lambda$. Moreover, we know:

- $\lambda_{\max}(\mathbf{L}_\mathbf{S}) \leq 2$, since $-1 \leq \sigma(\mathbf{S}) \leq 1$. (equality holds for bipartite graphs)
- $\|\mathbf{M}\|_2 = \frac{1}{n_{\min}}$, because:
  $\|\mathbf{M}_k\|_2 = 1 \Rightarrow \|\mathbf{B}_k^\top \mathbf{M}_k \mathbf{B}_k\|_2 = 1$.
  Also, $\mathbf{M}$ is a block diagonal with $\frac{1}{n_k}\mathbf{M}_i$ at each block. Hence, $\|\mathbf{M}\|_2 = \max_k \|\frac{1}{n_k}\mathbf{M}_k\|_2 = \max_k\{\frac{1}{n_k}\} = \frac{1}{\min_i n_k} = \frac{1}{n_{\min}}$

Hence,

$$\boldsymbol{\mu}_{\max}(\mathbf{Q}_\lambda) \leq \gamma + 2(1-\gamma) + \frac{2\lambda}{n_{\min}} = 2 - \gamma + \frac{2\lambda}{n_{\min}} \Rightarrow \frac{2}{2 - \gamma + \frac{2\lambda}{n_{\min}}} < \frac{2}{\boldsymbol{\mu}_{\max}(\nabla^2 g)} \tag{105}$$

meaning that a safe choice for $\eta$ is

$$\eta < \frac{2}{2 - \gamma + \frac{2\lambda}{n_{\min}}}. \tag{106}$$

Let's define $\rho := \|\mathbf{T}_{\eta,\lambda}\|_2 = \max\{|1 - \eta\boldsymbol{\mu}_{\min}|, |1 - \eta\boldsymbol{\mu}_{\max}|\} < 1$. Hence, equation 102 results in

$$\|\Phi(\mathbf{H}) - \Phi(\mathbf{Y})\|_2 \leq \rho\|\mathbf{H} - \mathbf{Y}\|_2. \tag{107}$$

By the Banach fixed-point theorem (Bauschke & Combettes, 2020), $\Phi$ admits a unique fixed point $\mathbf{H}^*$ and the iterates $\mathbf{H}^{(t+1)} = \Phi(\mathbf{H}^{(t)})$ converge to it from any initialization. The rate follows by applying equation 107 with $\mathbf{Y} = \mathbf{H}^*$: since $\Phi(\mathbf{H}^*) = \mathbf{H}^*$,

$$\|\mathbf{H}^{(t+1)} - \mathbf{H}^*\|_F = \|\Phi(\mathbf{H}^{(t)}) - \Phi(\mathbf{H}^*)\|_F \leq \rho\,\|\mathbf{H}^{(t)} - \mathbf{H}^*\|_F, \tag{108}$$

and induction on $t$ gives

$$\|\mathbf{H}^{(t)} - \mathbf{H}^*\|_F \leq \rho^t\,\|\mathbf{H}^{(0)} - \mathbf{H}^*\|_F. \tag{109}$$

Reaching accuracy $\varepsilon$ requires $t = O(\log(1/\varepsilon)/\log(1/\rho))$ iterations.

**The fixed point is the solution of the constrained problem.** Writing the update as $\Phi(\mathbf{H}) = \mathbf{P}[\mathbf{H} - \eta\nabla f_\lambda(\mathbf{H})]$, which is equivalent to $\mathrm{Proj}_\mathcal{X}(\mathbf{H} - \eta\nabla f_\lambda(\mathbf{H}))$, and using the fact that the projection onto a closed convex set $\mathcal{C}$ is characterized by $\langle \Pi_\mathcal{C}(\mathbf{z}) - \mathbf{z},\ \Pi_\mathcal{C}(\mathbf{z}) - \mathbf{y}\rangle \leq 0$ for all $\mathbf{y} \in \mathcal{C}$ (Bauschke & Combettes, 2020), $\mathbf{H}^* = \Phi(\mathbf{H}^*)$ holds if and only if

$$\left\langle \mathbf{H}^* - \eta\nabla f_\lambda(\mathbf{H}^*) - \mathbf{H}^*,\ \mathbf{Y} - \mathbf{H}^* \right\rangle \leq 0 \quad \text{for all } \mathbf{Y} \in \mathcal{X}. \tag{110}$$

This is equivalent to $\langle \nabla f_\lambda(\mathbf{H}^*), \mathbf{Y} - \mathbf{H}^* \rangle \geq 0$ for all $\mathbf{Y} \in \mathcal{X}$, using $\eta > 0$. This is precisely the first-order optimality condition for $\min_{\mathbf{H}\in\mathcal{X}} f_\lambda(\mathbf{H})$, which is necessary and sufficient because $f_\lambda$ is convex and $\mathcal{X}$ is convex. Since $f_\lambda$ is strictly convex the minimizer is unique, so $\mathbf{H}^* = \arg\min_{\mathbf{H}\in\mathcal{X}} f_\lambda(\mathbf{H})$.

□

# F PER-STEP VERSUS END-ONLY PROJECTION

**Theorem 14.** *Let $\mathbf{Y}^*$ denote the fixed point of the APPNP propagation, and let $\mathcal{X}$ be the closed convex set defined by the inequality and equality constraints $\mathcal{X} = \{\mathbf{X} \in \mathbb{R}^{n\times d} \mid w_i(\mathbf{X}) \leq 0,\ i = 1,\dots,p;\ u_j(\mathbf{X}) = 0,\ j = 1,\dots,q\}$. Then the solution obtained by projecting once at the end is characterized by the KKT conditions as*

$$\mathbf{X}^* = \mathbf{Y}^* - \tfrac{1}{2} z(\mathbf{X}^*), \tag{111}$$

*whereas the solution of FairProp satisfies*

$$\mathbf{X}^* = \mathbf{Y}^* - [\mathbf{I} - (1-\gamma)\mathbf{S} + 2\lambda\mathbf{M}]^{-1}\, z(\mathbf{X}^*), \tag{112}$$

*where $z(\mathbf{X}) = \sum_{i=1}^{p} \nu_i^* \nabla w_i(\mathbf{X}) + \sum_{j=1}^{q} \boldsymbol{\mu}_j^* \nabla u_j(\mathbf{X})$, with $\nu_i^*$ and $\boldsymbol{\mu}_j^*$ the optimal dual variables associated with the inequality and equality constraints, respectively.*

*Proof.* The proof is strongly based on the proof of Theorem 5.1 by Kariotakis & Konar (2026) It comprises two parts. First, we show how just projection once after all the propagation steps produces results that are agnostic to the graph structure, and then show that for FairProp propagation is not. We make use of the fact that the constraint set $\mathcal{X}$ is convex.

• **Propagation + Projection:** Given that $\mathbf{Y}^* = \gamma[\mathbf{I} - (1-\gamma)\mathbf{S} + 2\lambda\mathbf{M}]^{-1}\mathbf{H}$ is the fixed point of the regularized PPNP, and since $\mathcal{X}$ is convex, the optimization problem that this method is solving, comprises the following inequality and equality constraints:

$$\begin{aligned} \min_{\mathbf{X}} \quad & \|\mathbf{X} - \mathbf{Y}^*\|_2^2 \\ \text{s.t.} \quad & w_i(\mathbf{X}) \leq 0,\ i = 1,\dots,p \\ & u_j(\mathbf{X}) = 0,\ j = 1,\dots,q \end{aligned} \tag{113}$$

From the stationarity KKT condition we get that

$$\nabla\mathcal{L}(\mathbf{X}^*, \boldsymbol{\nu}, \boldsymbol{\mu}) = 0 \Leftrightarrow \nabla\{\|\mathbf{X}^* - \mathbf{Y}^*\|_2^2\} + \sum_{i=1}^{p} \nu_i \nabla w_i(\mathbf{X}^*) + \sum_{j=1}^{q} \boldsymbol{\mu}_j \nabla u_j(\mathbf{x}^*) = 0 \tag{114}$$

$$2(\mathbf{X}^* - \mathbf{Y}^*) + \sum_{i=1}^{p} \nu_i \nabla w_i(\mathbf{X}^*) + \sum_{j=1}^{q} \boldsymbol{\mu}_j \nabla u_j(\mathbf{X}^*) = 0 \tag{115}$$

$$\Leftrightarrow \mathbf{X}^* = \mathbf{Y}^* - \frac{1}{2}\left(\sum_{i=1}^{p} \nu_i \nabla w_i(\mathbf{X}^*) + \sum_{j=1}^{q} \boldsymbol{\mu}_j \nabla u_j(\mathbf{X}^*)\right). \tag{116}$$

As shown in the results above, applying a single projection after all propagation steps yields an outcome that is agnostic to the graph structure.

• **FairProp:** We have shown that FairProp is solving the following constrained, strongly convex QP

$$\min\{f_\lambda(\mathbf{X}) \text{ s.t. } \mathbf{X} \in \mathcal{X}\}, \tag{117}$$

where

$$f_\lambda(\mathbf{X}) := \frac{1}{2}\gamma\|\mathbf{X} - \mathbf{H}\|^2 + \frac{1}{2}(1-\gamma)\mathrm{tr}(\mathbf{X}^\top \mathbf{L}_\mathbf{S}\mathbf{X}) + \lambda\,\mathrm{tr}\big(\mathbf{X}^\top\mathbf{M}\mathbf{X}\big), \tag{118}$$

with

$$\nabla f_\lambda(\mathbf{X}) = [\mathbf{I} - (1-\gamma)\mathbf{S} + 2\lambda\mathbf{M}]\mathbf{X} - \gamma\mathbf{H}. \tag{119}$$

Since $\mathcal{X}$ is convex, the optimization problem comprises of the following inequality and equality constraints:

$$\begin{aligned} \min_{\mathbf{X}} \quad & f_\lambda(\mathbf{X}) \\ \text{s.t.} \quad & w_i(\mathbf{X}) \leq 0,\ i = 1,\dots,p \\ & u_j(\mathbf{X}) = 0,\ j = 1,\dots,q \end{aligned} \tag{120}$$

From the stationarity KKT condition we get that

$$\nabla \mathcal{L}(\mathbf{X}^*, \boldsymbol{\nu}, \boldsymbol{\mu}) = 0 \Leftrightarrow \nabla f(\mathbf{X}^*) + \sum_{i=1}^{p} \nu_i \nabla w_i(\mathbf{X}^*) + \sum_{j=1}^{q} \boldsymbol{\mu}_j \nabla u_j(\mathbf{X}^*) = 0 \tag{121}$$

$$\Leftrightarrow [\mathbf{I} - (1-\gamma)\mathbf{S} + 2\lambda \mathbf{M}]\mathbf{X}^* - \gamma \mathbf{H} + \sum_{i=1}^{p} \nu_i \nabla w_i(\mathbf{X}^*) + \sum_{j=1}^{q} \boldsymbol{\mu}_j \nabla u_j(\mathbf{X}^*) = 0 \tag{122}$$

$$\Leftrightarrow \mathbf{X}^* = \mathbf{Y}^* - [\mathbf{I} - (1-\gamma)\mathbf{S} + 2\lambda \mathbf{M}]^{-1} \left( \sum_{i=1}^{p} \nu_i \nabla w_i(\mathbf{X}^*) + \sum_{j=1}^{q} \boldsymbol{\mu}_j \nabla u_j(\mathbf{X}^*) \right) \tag{123}$$

$$\tag{124}$$

where $\mathbf{Y}^* = \gamma[\mathbf{I} - (1-\gamma)\mathbf{S} + 2\lambda \mathbf{M}]^{-1}\mathbf{H}$.
Re-expressing $[\mathbf{I} - (1-\gamma)\mathbf{S} + 2\lambda \mathbf{M}]^{-1} = \sum_{j=0}^{\infty}[(1-\gamma)\mathbf{S} - 2\lambda \mathbf{M}]^j$ shows that FairProp corrects $\mathbf{Y}^*$ via a graph diffusion term in order to arrive at the solution, hence it is clear that it makes use of the graph structure. For $\lambda = 0$, i.e., only mean-equalization, we get $[\mathbf{I} - (1-\gamma)\mathbf{S}]^{-1} = \sum_{j=0}^{\infty}(1-\gamma)^j \mathbf{S}^j$. □

### F.1 One Dimensional Embeddings and Mean-Equalization

#### F.1.1 End-Only Projection

**Proposition 15.** *Let $\mathcal{S}_0, \mathcal{S}_1$ be the sensitive groups of nodes, with $n_k = |\mathcal{S}_k| > 0$ and $n_0 + n_1 = n$, and let the number of features be $d = 1$. Define the following fairness constraint set*

$$\mathcal{X}_0 := \left\{ \mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{a}^\top \mathbf{x} = 0 \right\}, \qquad \mathbf{a} := \frac{\mathbf{e}_{\mathcal{S}_0}}{n_0} - \frac{\mathbf{e}_{\mathcal{S}_1}}{n_1} \in \mathbb{R}^n. \tag{125}$$

*Let $\mathbf{y} \in \mathbb{R}^n$ be the given embedding vector and let $\delta := \mathbf{a}^\top \mathbf{y} = \bar{\mathbf{y}}_0 - \bar{\mathbf{y}}_1$, with $\bar{\mathbf{y}}_k := \frac{1}{n_k}\sum_{i \in \mathcal{S}_k} \mathbf{y}(i)$, be the difference of the two group means. Then the Euclidean projection of $\mathbf{y}$ onto $\mathcal{X}_0$ is unique and given entry-wise by*

$$\mathbf{x}^*(i) = \begin{cases} \mathbf{y}(i) - \dfrac{n_1}{n}\,\delta, & i \in \mathcal{S}_0 \\ \mathbf{y}(i) + \dfrac{n_0}{n}\,\delta, & i \in \mathcal{S}_1, \end{cases} \tag{126}$$

*equivalently $\mathbf{x}^* = \left(\mathbf{I} - \frac{\mathbf{a}\mathbf{a}^\top}{\|\mathbf{a}\|_2^2}\right)\mathbf{y}$.*

*Proof.* $\mathcal{X}_0$ is a hyperplane through the origin and $\|\mathbf{x} - \mathbf{y}\|_2^2$ is strictly convex, so the problem

$$\min_{\mathbf{x} \in \mathcal{X}_0} \quad \|\mathbf{x} - \mathbf{y}\|_2^2 \tag{127}$$

has a unique solution, characterized by stationarity of the Lagrangian $\mathcal{L}(\mathbf{x}, \lambda) = \|\mathbf{x} - \mathbf{y}\|_2^2 + 2\lambda\, \mathbf{a}^\top \mathbf{x}$:

$$\nabla_{\mathbf{x}} \mathcal{L} = 2(\mathbf{x} - \mathbf{y}) + 2\lambda \mathbf{a} = \mathbf{0} \quad \Longrightarrow \quad \mathbf{x} = \mathbf{y} - \lambda \mathbf{a}. \tag{128}$$

Imposing feasibility gives $\mathbf{a}^\top \mathbf{y} - \lambda \|\mathbf{a}\|_2^2 = 0$, and since $\mathbf{e}_{\mathcal{S}_0}$ and $\mathbf{e}_{\mathcal{S}_1}$ have disjoint supports,

$$\|\mathbf{a}\|_2^2 = \frac{1}{n_0} + \frac{1}{n_1} = \frac{n}{n_0 n_1}, \qquad \text{hence} \qquad \lambda^* = \frac{\delta}{\|\mathbf{a}\|_2^2} = \frac{n_0 n_1}{n}\,\delta. \tag{129}$$

Substituting $\lambda^*$ into equation 128 and using $\mathbf{a}(i) = 1/n_0$ for $i \in \mathcal{S}_0$ and $\mathbf{a}(i) = -1/n_1$ for $i \in \mathcal{S}_1$ yields equation 126. □

**Remark 16.** *The correction in equation 126 is constant within each group. The groups are shifted in opposite directions by amounts inversely proportional to their sizes, with the two displacements standing in ratio $n_1 : n_0$. The minority is thus moved further, and in the limit of a small minority the majority is left essentially untouched, while the minority absorbs the entire mean gap. The constraint equalizes the group means but not the price paid for doing so. Utility is sacrificed almost entirely on the minority side, so a criterion imposed for fairness reasons is itself satisfied unfairly.*

### F.1.2 PER-STEP PROJECTION OF FAIRPROP

**Proposition 17.** *Let $\mathcal{S}_0, \mathcal{S}_1$ be the sensitive groups of nodes, with $n_k = |\mathcal{S}_k| > 0$ and $n_0 + n_1 = n$, and let the number of features be $d$. Define the following fairness constraint set*

$$\mathcal{X}_0 := \left\{ \mathbf{X} \in \mathbb{R}^n \ : \ \mathbf{a}^\top \mathbf{X} = 0 \right\}, \qquad \mathbf{a} := \frac{\mathbf{e}_{\mathcal{S}_0}}{n_0} - \frac{\mathbf{e}_{\mathcal{S}_1}}{n_1} \in \mathbb{R}^n. \tag{130}$$

*Let $\mathbf{Y}^*$ denote the fixed point of the APPNP propagation, then the solution of FairProp (for $\lambda = 0$) satisfies:*

$$\mathbf{X}^* = \mathbf{Y}^* - \frac{\mathbf{Q}^{-1}\mathbf{a}\mathbf{a}^\top \mathbf{Y}^*}{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{a}} = \left[ \mathbf{I} - \frac{\mathbf{Q}^{-1}\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{a}} \right] \mathbf{Y}^*, \tag{131}$$

*with $\mathbf{Q} = [\mathbf{I} - (1-\gamma)\mathbf{S}]$.*

*Proof.* We have shown that FairProp is solving the following constrained, strongly convex QP

$$\min \left\{ f(\mathbf{X}) \text{ s.t. } \mathbf{X} \in \mathcal{X}_0 \right\}, \tag{132}$$

where

$$f(\mathbf{X}) := \frac{1}{2}\gamma \|\mathbf{X} - \mathbf{H}\|^2 + \frac{1}{2}(1-\gamma)\mathrm{tr}(\mathbf{X}^\top \mathbf{L}_\mathbf{S} \mathbf{X}), \tag{133}$$

with

$$\nabla f(\mathbf{X}) = \mathbf{Q}\mathbf{X} - \gamma \mathbf{H}, \qquad \mathbf{Q} = [\mathbf{I} - (1-\gamma)\mathbf{S}] \succ 0. \tag{134}$$

Since $\mathcal{X}_0$ is convex, we have the Lagrangian

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\mu}) = f(\mathbf{X}) + \mathrm{tr}(\mathbf{X}^\top \mathbf{a}\boldsymbol{\mu}^\top) \tag{135}$$

the following KKT conditions:
• (Stationarity): $\nabla_\mathbf{X} \mathcal{L} = \nabla_\mathbf{X} f(\mathbf{X}) + \mathbf{a}\boldsymbol{\mu}^\top = 0 \Leftrightarrow \mathbf{Q}\mathbf{X} = \gamma \mathbf{H} - \mathbf{a}\boldsymbol{\mu}^\top$
• (Primal Feasibility): $\mathbf{a}^\top \mathbf{X} = 0$
Since $\mathbf{Q} \succ 0$,

$$\mathbf{X} = \mathbf{Q}^{-1}[\gamma \mathbf{H} - \mathbf{a}\boldsymbol{\mu}^\top]. \tag{136}$$

Substituting back to $\mathbf{a}^\top \mathbf{X} = 0$, we get:

$$\gamma \mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{H} - \mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{a}\boldsymbol{\mu}^\top = 0 \Rightarrow \boldsymbol{\mu}^\top = \gamma \frac{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{H}}{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{a}}. \tag{137}$$

Then, we have the following closed form for the solution $\mathbf{X}^*$:

$$\mathbf{X}^* = \gamma \mathbf{Q}^{-1}\mathbf{H} - \gamma \mathbf{Q}^{-1}\mathbf{a} \frac{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{H}}{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{a}}. \tag{138}$$

The solution of the unconstrained problem of APPNP $\min_\mathbf{X} \left\{ f(\mathbf{X}) \right\}$, is $\mathbf{Y}^* = \gamma \mathbf{Q}^{-1}\mathbf{H}$. Hence, we can express $\mathbf{X}^*$ it in terms of the unconstrained solution, as

$$\mathbf{X}^* = \mathbf{Y}^* - \frac{\mathbf{Q}^{-1}\mathbf{a}\mathbf{a}^\top \mathbf{Y}^*}{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{a}} = \left[ \mathbf{I} - \frac{\mathbf{Q}^{-1}\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{a}} \right] \mathbf{Y}^*. \tag{139}$$

□

**Corollary 18.** *Let $\mathcal{S}_0, \mathcal{S}_1$ be the sensitive groups of nodes, with $n_k = |\mathcal{S}_k| > 0$ and $n_0 + n_1 = n$, and let the number of features be $d = 1$. Define the following fairness constraint set*

$$\mathcal{X}_0 := \left\{ \mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{a}^\top \mathbf{x} = 0 \right\}, \qquad \mathbf{a} := \frac{\mathbf{e}_{\mathcal{S}_0}}{n_0} - \frac{\mathbf{e}_{\mathcal{S}_1}}{n_1} \in \mathbb{R}^n. \tag{140}$$

*Let $\mathbf{y}^*$ denote the fixed point of the APPNP propagation, and let $\delta := \mathbf{a}^\top \mathbf{y}^* = \bar{\mathbf{y}}_0 - \bar{\mathbf{y}}_1$, with $\bar{\mathbf{y}}_k := \frac{1}{n_k} \sum_{i \in \mathcal{S}_k} \mathbf{y}^*(i)$, be the difference of the two group means. Then the solution of FairProp satisfies:*

$$\mathbf{x}^* = \mathbf{y}^* - \frac{\mathbf{Q}^{-1}\mathbf{a}}{\mathbf{a}^\top \mathbf{Q}^{-1}\mathbf{a}} \delta. \tag{141}$$

In Figure 4 we can see an illustration of the Projection versus FairProp (proj) on a simulated two-group graph ($n_0 = 40$, $n_1 = 240$, $d = 1$, $\gamma = 0.1$, no variance regularization). The horizontal axis is the single feature and the vertical axis is a display-only jitter, identical in all panels. **Left:** the APPNP fixed point $\mathbf{y}^*$ has a mean gap $\delta = \mathbf{a}^\top \mathbf{y}^*$ between the minority and majority groups. **Middle:** the Euclidean projection onto $\{\mathbf{x} : \mathbf{a}^\top \mathbf{x} = 0\}$ shifts every node in a group by the same amount ($\frac{n_1}{n}\delta$ and $\frac{n_0}{n}\delta$). The group means coincide, but each group keeps its spread. **Right:** FairProp shifts node $i$ proportionally to $(\mathbf{Q}^{-1}\mathbf{a})_i$, so nodes with many cross-group neighbours move less and nodes deep inside their group move more. This cancels the uneven propagation of the input mean gap through the graph, so the group means coincide and the within-group spread happens to shrink as well. Red brackets show the minority spread (mean $\pm 2$ std).

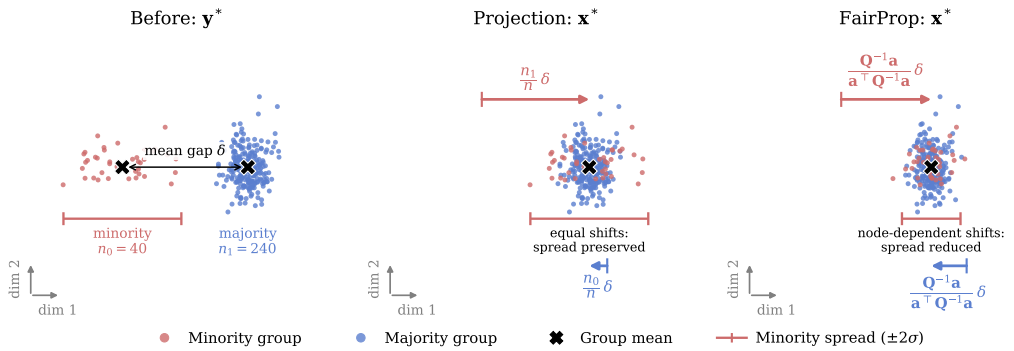



Figure 4: Left: APPNP output $\mathbf{y}^*$ with mean gap $\delta$. Middle: the Euclidean projection shifts each group uniformly, equalizing the means but preserving within-group spread. Right: FairProp shifts node $i$ proportionally to $(\mathbf{Q}^{-1}\mathbf{a})_i$, equalizing the means while also reducing within-group spread. Brackets: minority mean $\pm 2$ std ($\gamma = 0.1$, $\lambda = 0$).

In Figure 5 we can see the graph on which the previous experiments were run on. The minority contains bridge nodes (outlined), which have many edges to the majority, and core nodes, which have almost none; this makes $(\mathbf{Q}^{-1}\mathbf{a})_i$ vary within the minority and hence makes FairProp's shifts node-dependent.

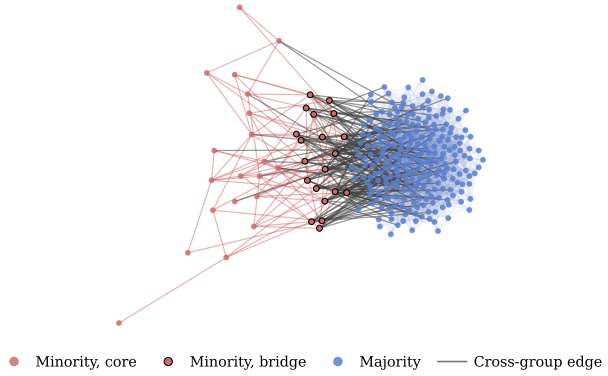



Figure 5: The simulated graph. Outlined minority nodes have edges to the majority; the remaining minority nodes have almost none.

# G EXPERIMENTS

## G.1 EXPERIMENTAL SETUP

All experiments were performed on a single machine, with an Intel i7-13700K CPU @ 3.4GHz and 128GB of main memory, and an NVIDIA GeForce RTX 3070 GPU (8GB VRAM), running Python 3.10.19 and PyTorch 2.6.0.

**FairProp Details:**
*MLP Architecture.* All tasks, i.e., node classification (NC), link prediction (LP), node regression (NR), share a two-layer MLP from the original APPNP architecture (Gasteiger et al., 2018),

$$\text{Dropout}(x) \;\to\; \text{Linear}(d_{\text{in}}, 64) \;\to\; \text{ReLU} \;\to\; \text{Dropout} \;\to\; \text{Linear}(64, d_{\text{out}}),$$

where $d_{\text{out}}$ equals the number of classes for NC, 64 for LP, and 1 for NR, and dropout probability is $p = 0.5$ throughout. After the MLP, APPNP-style propagation runs for $T = 10$ steps.

*Teleportation $\gamma$.* We fix $\gamma = 0.1$ across all runs, unless stated otherwise.

*Optimizer and Training.* We train all models with the Adam optimizer for 200 epochs at a learning rate of $10^{-2}$. For NC and NR we apply weight decay $5 \times 10^{-4}$; for LP we use no weight decay. All results are averaged over 20 independent runs with different random seeds.

*Loss Functions.* For NC, we use negative log-likelihood on the training nodes. For LP, binary cross-entropy with logits and negative sampling, and for NR, the mean-squared error on the training nodes.

## G.2 DATASETS

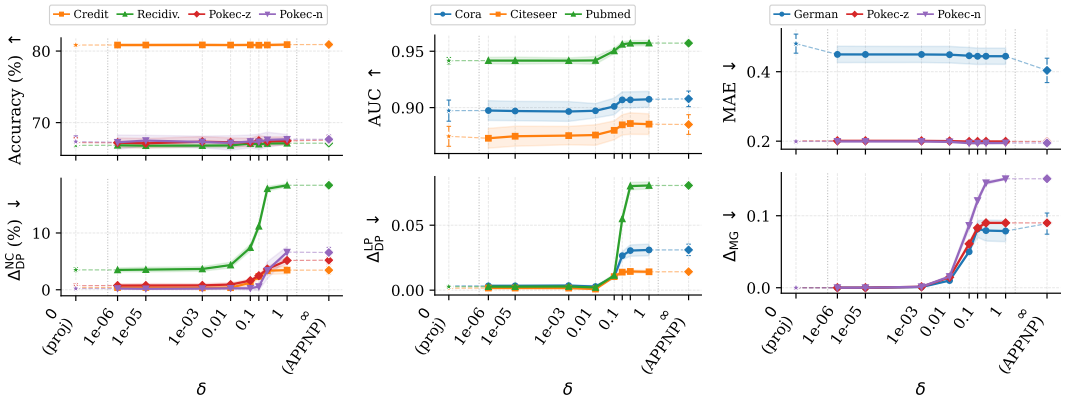
For node classification, we use Credit (Yeh & Lien, 2009), a consumer lending network where nodes are clients connected by payment-behaviour similarity, with age as the sensitive attribute and payment default as the prediction target; Recidivism (Jordan & Freiburger, 2015; Agarwal et al., 2021), where nodes are criminal defendants linked by similar criminal profiles, with race as the sensitive attribute and re-offending as the target; and Pokec-z and Pokec-n (Dai & Wang, 2021), two regional subgraphs of the Slovak social network Pokec (Zilina and Nitra regions), where nodes are users, edges are friendships, and gender is the sensitive attribute. For link prediction, we use the citation networks Cora, Citeseer, and Pubmed (Yang et al., 2016), where nodes are academic papers and edges are citations; we adopt label parity (topic class mod 2) as a proxy sensitive attribute following standard LP fairness convention (Spinelli et al., 2021). For node regression, we use GermanCredit (Asuncion et al., 2007), a loan applicant network where nodes are connected by feature similarity and the target is credit score; and again Pokec-z and Pokec-n, where the regression target is profile completion percentage, with gender as the sensitive attribute. For the dataset statistics please refer to Table 4.

## G.3 ABLATION STUDY: EFFECT OF THE MEAN-CENTERING TOLERANCE PARAMETER $\delta$.

The parameter $\delta$ controls the tightness of the mean-centering constraints in the FairProp projections, and as a consequence, the tightness of the upper bound on the fairness parity gap. Setting $\delta = 0$ enforces exact mean-centering, requires projecting onto a subspace, and yields the tightest bound on the parity gap (exact for node regression). Increasing $\delta > 0$ relaxes the mean-centering constraint, requires projecting onto a slab, and also loosens the upper-bound on the fairness parity gap. In the limit, as $\delta \to \infty$, the constraint vanishes, and the FairProp propagations become equivalent to its unconstrained APPNP counterpart. Hence, $\delta$ can be viewed as a tunable parameter which interpolates between these two regimes. Our goal is to investigate how task utility and fairness are affected as $\delta$ is varied. From our results in the previous subsection, we observe that setting $\delta = 0$ in FairProp causes a small drop in predictive accuracy compared to APPNP ($\delta \to \infty$) while significantly reducing the parity gap. By varying $\delta > 0$, we test if we can close the accuracy gap to APPNP while possibly incurring modest increase in the parity gap. Figure 6 showcases the results across different tasks and datasets. As $\delta$ is increased from 0, the general trend is that both accuracy and parity gap stay roughly the same until a tipping point is reached, which usually occurs around $\delta = 10^{-2}$. Beyond this value, $\delta$ begins to exhibit an influence on the results, with a noticeable

Table 4: Datasets used in the experiments. Edges are counted without multiplicity (undirected). For node regression (NR) the *Target* column lists the predicted continuous quantity; for link prediction (LP) node classes are not used as a task label but are listed for reference.

| Dataset | Task | Nodes | Edges | Features | Sensitive attr. | Target / Classes |
|---|---|---|---|---|---|---|
| *Node Classification (NC)* | | | | | | |
| Credit | NC | 30,000 | 150,000 | 22 | Age ($> 25$) | 2 classes |
| Recidivism | NC | 6,172 | 30,860 | 7 | Race | 2 classes |
| Pokec-z | NC | 67,796 | 617,958 | 276 | Gender | 2 classes |
| Pokec-n | NC | 66,569 | 517,047 | 265 | Gender | 2 classes |
| *Link Prediction (LP)* | | | | | | |
| Cora | LP | 2,708 | 5,278 | 1,433 | Class parity | 7 classes |
| Citeseer | LP | 3,327 | 4,552 | 3,703 | Class parity | 6 classes |
| Pubmed | LP | 19,717 | 44,324 | 500 | Class parity | 3 classes |
| *Node Regression (NR)* | | | | | | |
| German Credit | NR | 1,000 | 10,122 | 33 | Gender | Credit score |
| Pokec-z | NR | 67,796 | 617,958 | 276 | Gender | Profile completion |
| Pokec-n | NR | 66,569 | 517,047 | 265 | Gender | Profile completion |




(a) Node Classification (b) Link Prediction (c) Node Regression

Figure 6: Showing the effect of relaxing the mean-gap $\delta$ by using the slab projection, on different tasks and datasets.

increase in the fairness parity gap, while accuracy remains largely unchanged. The only exception is the GermanCredit dataset used for node regression, where relaxing the constraint reduces the MAE by $7\%$ from $0.481$ to $0.448$. At $\delta \to \infty$, we recover the APPNP solution, empirically confirming that the slab constraint reduces to the unconstrained problem in the limit. This indicates that by varying $\delta$ in the FairProp projections, offers a way to manage the utility-fairness trade-off in cases where exact projection leads to a significant drop in utility.

### G.4 Detailed Experiments

Note that for link prediction, we also use methods originally proposed for node classification. We discard the classification head and replace it with an inner-product decoder applied to the learned node embeddings, optimizing binary cross-entropy loss. Moreover, for node regression, since there is only one available baseline, we also use methods originally proposed for node classification. We replace the classification head with a single linear output unit and optimized mean squared error loss in place of cross-entropy.

Table 5: **Node classification** results. Best and second best per column in **bold** and underlined, respectively, excluding the unfair baseline. Dashed results indicate out-of-memory (OOM) or time-consuming experiments. All results are presented by $\times 100$.

| | Credit | | Recidivism | | Pokec-z | | Pokec-n | |
|---|---|---|---|---|---|---|---|---|
| | Acc ↑ | $\Delta_{\text{DP}}^{\text{NC}}$ ↓ | Acc ↑ | $\Delta_{\text{DP}}^{\text{NC}}$ ↓ | Acc ↑ | $\Delta_{\text{DP}}^{\text{NC}}$ ↓ | Acc ↑ | $\Delta_{\text{DP}}^{\text{NC}}$ ↓ |
| APPNP | 80.9 ±0.2 | 3.5 ±0.7 | 67.1 ±0.2 | 18.4 ±0.3† | 67.6 ±0.6 | 5.2 ±0.6 | 67.7 ±0.6 | 6.6 ±0.9 |
| FairProp (proj) | 80.8 ±0.2 | **0.4 ±0.3** | 66.8 ±0.3 | **3.5 ±0.4**† | 67.2 ±0.7 | **0.7 ±0.3** | 67.3 ±0.8 | **0.2 ±0.2** |
| FairGNN | 79.9 ±1.1 | 2.9 ±1.9 | 54.0 ±10.0 | 10.8 ±7.6 | 53.9 ±3.9 | 5.0 ±3.3 | 52.8 ±2.7 | 4.4 ±4.8 |
| NIFTY | 80.2 ±0.3 | 10.1 ±0.4 | 68.1 ±1.0 | 19.8 ±0.5 | 65.5 ±1.7 | 11.5 ±3.1 | 64.7 ±1.1 | 4.7 ±3.2 |
| RawlsGCN | 79.6 ±0.7 | 1.7 ±0.7 | 67.9 ±0.9 | 18.7 ±1.1 | 57.1 ±6.4 | 2.5 ±4.5 | 54.5 ±6.6 | 2.6 ±4.7 |
| FairGAT | 79.4 ±0.5 | 1.2 ±0.3 | 67.8 ±1.1 | 15.1 ±0.5 | **69.4 ±0.9** | 2.0 ±1.2 | **68.6 ±0.8** | 2.3 ±3.0 |
| EDITS | — | — | 65.8 ±2.8 | 19.1 ±5.5 | — | — | — | — |
| FairEdit | — | — | 68.1 ±0.9 | 19.2 ±0.8 | — | — | — | — |
| FairVGNN | 80.0 ±0.3 | 0.8 ±0.1 | 65.5 ±2.3 | 19.5 ±0.4 | — | — | — | — |
| FairWalk | 73.4 ±0.4 | 1.5 ±0.8 | 60.8 ±1.2 | 14.0 ±1.4 | 63.1 ±0.9 | 4.3 ±3.7 | 61.6 ±1.0 | 4.4 ±3.3 |
| CrossWalk | 73.6 ±0.4 | 3.1 ±1.3 | 60.6 ±1.2 | 16.1 ±1.3 | 63.1 ±1.1 | 5.2 ±3.7 | 62.3 ±0.7 | 4.9 ±3.5 |
| FMP | 81.2 ±0.1 | 3.5 ±0.3 | **69.8 ±0.3** | 19.1 ±0.3 | 65.8 ±0.4 | 3.9 ±0.3 | 65.3 ±0.5 | 1.7 ±0.5 |
| FairINV | 79.2 ±1.0 | 7.2 ±4.4 | 67.3 ±1.6 | 21.1 ±6.2 | 63.1 ±1.0 | 8.0 ±4.8 | 62.4 ±1.1 | 5.6 ±4.2 |
| FairSAD | 78.1 ±0.5 | **0.4 ±0.5** | 63.4 ±3.2 | 10.0 ±3.8 | — | — | — | — |
| DAB-GNN | 79.6 ±1.1 | 4.6 ±2.0 | 63.3 ±6.3 | 8.7 ±6.3 | 66.5 ±2.0 | 3.9 ±3.3 | 65.1 ±1.3 | 4.6 ±3.8 |
| FairGP | **81.6 ±0.3** | 3.3 ±0.8 | 68.0 ±0.9 | 6.4 ±2.5 | — | — | — | — |

† Recidivism results for APPNP / FairProp (proj) use $\gamma = 0.3$; all other datasets use $\gamma = 0.1$.

Table 6: **Link prediction** results. Best and second best per column in **bold** and underlined, respectively, excluding the unfair baseline.

| | Cora | | Citeseer | | Pubmed | |
|---|---|---|---|---|---|---|
| | AUC ↑ | $\Delta_{\text{DP}}^{\text{LP}}$ ↓ | AUC ↑ | $\Delta_{\text{DP}}^{\text{LP}}$ ↓ | AUC ↑ | $\Delta_{\text{DP}}^{\text{LP}}$ ↓ |
| APPNP | 0.908 ±0.007 | 0.031 ±0.004 | 0.885 ±0.009 | 0.014 ±0.001 | 0.957 ±0.002 | 0.081 ±0.002 |
| FairProp (proj) | 0.897 ±0.009 | 0.003 ±0.001 | 0.875 ±0.009 | 0.002 ±0.000 | 0.942 ±0.003 | 0.003 ±0.001 |
| FairGNN | 0.683 ±0.070 | 0.008 ±0.006 | 0.808 ±0.041 | 0.019 ±0.009 | 0.873 ±0.065 | 0.007 ±0.008 |
| EDITS | 0.905 ±0.008 | 0.009 ±0.002 | 0.879 ±0.007 | 0.008 ±0.001 | 0.966 ±0.001 | 0.028 ±0.005 |
| FairEdit | 0.799 ±0.049 | 0.015 ±0.010 | 0.808 ±0.032 | 0.012 ±0.004 | 0.914 ±0.025 | 0.004 ±0.004 |
| FairVGNN | 0.852 ±0.057 | 0.051 ±0.021 | 0.851 ±0.012 | 0.036 ±0.010 | 0.745 ±0.089 | 0.032 ±0.036 |
| FairWalk | 0.908 ±0.006 | 0.024 ±0.002 | 0.883 ±0.007 | 0.014 ±0.002 | 0.961 ±0.002 | 0.071 ±0.002 |
| CrossWalk | 0.907 ±0.007 | 0.024 ±0.002 | 0.884 ±0.008 | 0.014 ±0.001 | 0.961 ±0.002 | 0.072 ±0.002 |
| FairGAT | 0.915 ±0.006 | 0.050 ±0.017 | 0.919 ±0.007 | 0.027 ±0.016 | 0.845 ±0.017 | 0.066 ±0.034 |
| FairAdj | 0.915 ±0.006 | 0.019 ±0.001 | 0.888 ±0.006 | 0.013 ±0.001 | **0.969 ±0.001** | 0.050 ±0.001 |
| FairDrop | 0.922 ±0.006 | 0.022 ±0.003 | 0.900 ±0.007 | 0.015 ±0.002 | 0.967 ±0.001 | 0.060 ±0.002 |
| FairWire | 0.811 ±0.013 | **0.000 ±0.000** | 0.817 ±0.013 | **0.000 ±0.000** | 0.917 ±0.002 | **0.001 ±0.001** |
| FairINV | **0.928 ±0.006** | 0.027 ±0.001 | **0.924 ±0.007** | 0.014 ±0.001 | 0.953 ±0.001 | 0.080 ±0.001 |
| DAB-GNN | 0.843 ±0.007 | 0.181 ±0.016 | 0.771 ±0.013 | 0.131 ±0.014 | 0.872 ±0.004 | 0.180 ±0.010 |
| FairSAD | 0.848 ±0.019 | 0.001 ±0.000 | 0.817 ±0.028 | 0.001 ±0.000 | 0.743 ±0.012 | **0.001 ±0.001** |
| FairGP | 0.746 ±0.021 | **0.000 ±0.000** | 0.865 ±0.010 | 0.001 ±0.000 | 0.835 ±0.009 | **0.001 ±0.001** |

Table 7: **Node regression** results. Best and second best per column in **bold** and underlined, respectively, excluding the unfair baseline.

| | GermanCredit | | Pokec-z | | Pokec-n | |
|---|---|---|---|---|---|---|
| | MAE ↓ | $\Delta_{\mathrm{MG}}$ ↓ | MAE ↓ | $\Delta_{\mathrm{MG}}$ ↓ | MAE ↓ | $\Delta_{\mathrm{MG}}$ ↓ |
| APPNP | 0.404 ±0.035 | 0.089 ±0.015 | 0.199 ±0.001 | 0.091 ±0.001† | 0.199 ±0.001 | 0.152 ±0.001† |
| FairProp (proj) | 0.481 ±0.027 | **0.000 ±0.000** | **0.200 ±0.001** | **0.000 ±0.000**† | **0.199 ±0.001** | **0.000 ±0.000**† |
| FairGNN | 0.415 ±0.016 | 0.070 ±0.024 | 0.600 ±0.021 | 0.100 ±0.011 | 1.826 ±5.353 | 0.453 ±1.322 |
| FairGAT | 0.442 ±0.012 | 0.017 ±0.012 | 0.652 ±0.052 | 0.061 ±0.007 | 0.654 ±0.042 | 0.094 ±0.008 |
| EDITS | 0.414 ±0.013 | 0.024 ±0.019 | 0.587 ±0.004 | 0.006 ±0.002 | 0.586 ±0.005 | 0.020 ±0.001 |
| FairVGNN | **0.407 ±0.024** | 0.049 ±0.021 | — | — | — | — |
| FairEdit | 0.429 ±0.029 | 0.019 ±0.015 | — | — | — | — |
| FairWalk | 0.445 ±0.008 | 0.017 ±0.012 | 0.478 ±0.004 | 0.055 ±0.001 | 0.472 ±0.004 | 0.100 ±0.002 |
| CrossWalk | 0.429 ±0.010 | 0.021 ±0.013 | 0.478 ±0.003 | 0.057 ±0.001 | 0.471 ±0.004 | 0.109 ±0.002 |
| FMP | 0.558 ±0.006 | 0.074 ±0.008 | 0.204 ±0.002 | 0.096 ±0.001 | 0.201 ±0.002 | 0.154 ±0.001 |
| FairINV | 0.464 ±0.000 | 0.209 ±0.001 | 0.635 ±0.000 | 0.085 ±0.000 | 0.640 ±0.000 | 0.150 ±0.000 |
| DAB-GNN | 0.750 ±0.118 | 0.107 ±0.081 | 0.722 ±0.231 | 0.034 ±0.020 | 0.640 ±0.159 | 0.081 ±0.034 |
| FairSAD | 0.456 ±0.020 | 0.135 ±0.027 | — | — | — | — |
| FairGP | 0.524 ±0.017 | 0.163 ±0.036 | — | — | — | — |
| FnRGNN | 0.521 ±0.011 | 0.022 ±0.018 | 0.249 ±0.021 | 0.004 ±0.003 | 0.239 ±0.005 | 0.020 ±0.008 |

† Pokec-z/-n results for APPNP / FairProp (proj) use $\gamma = 0.5$; GermanCredit use $\gamma = 0.1$.

## G.5 RUNTIME RESULTS

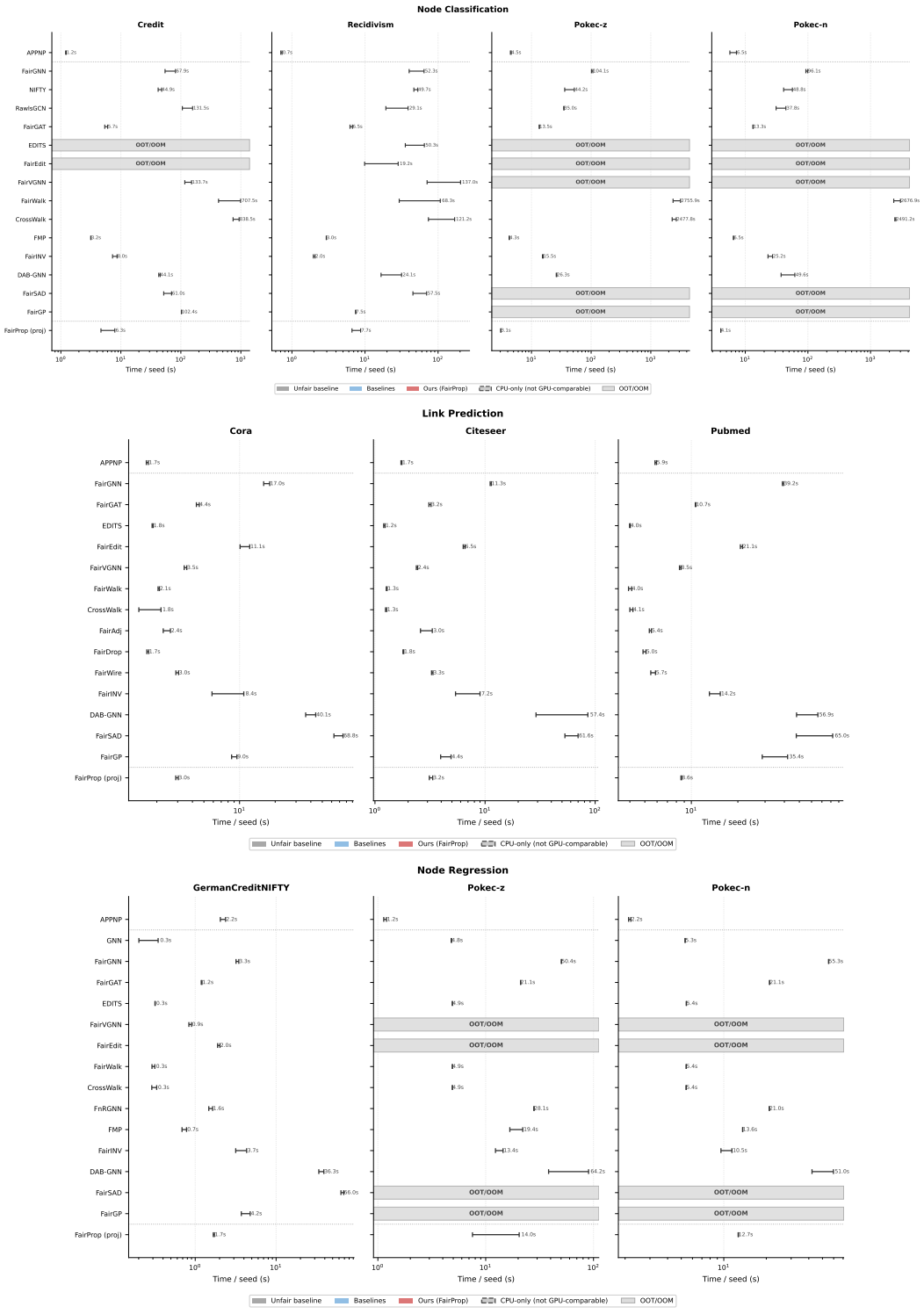



Figure 7: Runtime Results. OOT stands for Out-of-Time and OOM for Out-of-Memory.

### G.6 Utility-Fairness Trade-Off Plots

#### G.6.1 Node Classification Scatter Plots

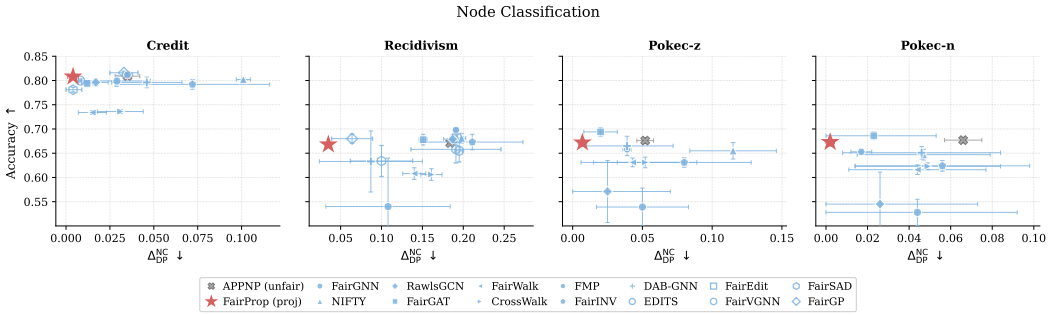



Figure 8: Utility-fairness trade-off on node classification. Each point shows mean accuracy (higher is better) against $\Delta_{DP}^{NC}$ (lower is better) over 20 runs. Top-left is optimal.

#### G.6.2 Link Prediction Scatter Plots

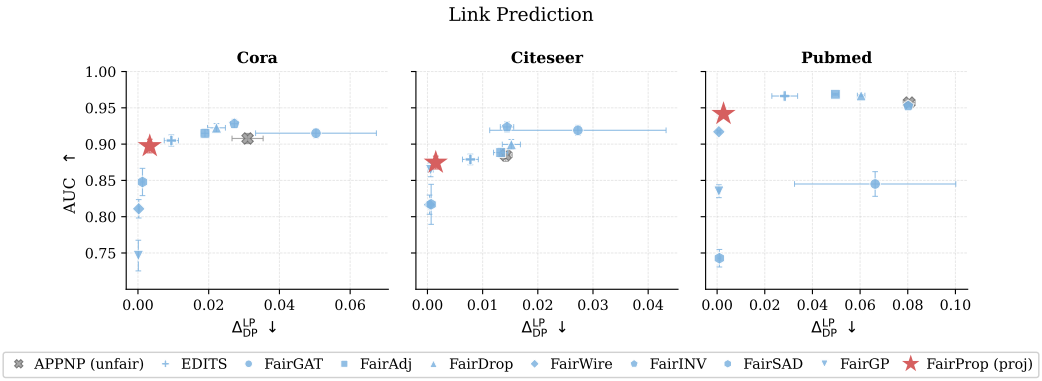



Figure 9: Utility-fairness trade-off on link prediction. Each point shows mean AUC (higher is better) against $\Delta_{DP}^{LP}$ (lower is better) over 20 runs. Top-left is optimal.

### G.7 Node Classification Bound Comparison

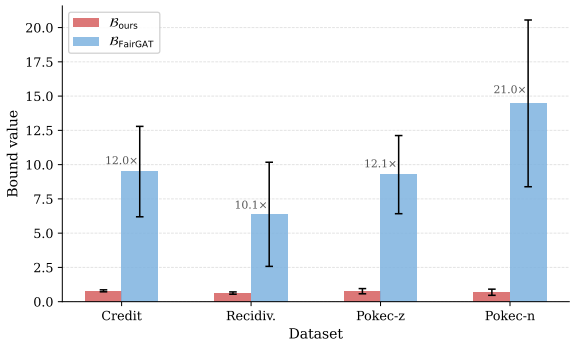



Figure 10: Comparison of our bound on the soft prediction gap against that of FairGAT (Kose & Shen, 2024a).