BV-SAMPLER: BIAS–VARIANCE-AWARE SAMPLING FOR GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neighborhood sampling enables scalable graph neural networks but introduces aggregation error under finite computation budgets. We propose BV-SAMPLER, a bias–variance-aware framework combining feature-aware sampling, contribution weighting, and sampler-induced contrastive learning under sparse supervision. We characterize optimal weights and proposals when the other is fixed, and develop an efficient diagonal surrogate. We also analyze layer-wise error accumulation and connect independent sampler-induced view disagreement to aggregation variance, yielding an encoded alignment bound. Experiments demonstrate competitive prediction across rich- and sparse-label settings, with the largest gains under sampled inference. On low-label OGBN-arXiv, BV-SAMPLER improves Micro-F1 by 6.77 percentage points over LABOR and 2.95 points over GraphSAINT with DGI. Ablations and direct diagnostics support the contributions of sampling and weighting and demonstrate reduced MSE through a bias–variance trade-off.

1 INTRODUCTION

Graph neural networks (GNNs) have emerged as a powerful framework for learning on graph-structured data (Veličković et al., 2017; Xu et al., 2018), with applications spanning social networks (Fan et al., 2019), recommendation systems (Wu et al., 2022), molecular biology (Hu et al., 2020), and citation analysis (Liu et al., 2022a). A central challenge in scaling GNNs to large graphs is the computational cost of neighborhood aggregation (Chen et al., 2018; Zeng et al., 2020; Zou et al., 2019): recursively expanding neighborhoods can cause the computation to grow exponentially with network depth. Sampling-based methods such as GraphSAGE (Hamilton et al., 2017), FastGCN (Chen et al., 2018), GraphSAINT (Zeng et al., 2020), and LADIES (Zou et al., 2019) address this by estimating representations from sampled nodes or subgraphs, enabling mini-batch training on large-scale graphs.

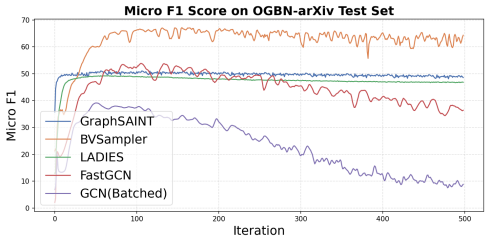


Figure 1: Micro-F1 trajectories on OGBN-arXiv under the sparse-label setting. FastGCN and GCN deteriorate substantially after early peaks, while GraphSAINT and LADIES exhibit milder declines.

An important question is how to preserve representation quality when only a limited number of neighbors can be sampled. Many sampling analyses emphasize unbiased estimation and variance reduction. However, aggregation mean squared error (MSE) depends on both squared bias and variance, and an unbiased estimator need not minimize this error under a finite budget. Classical estimation theory shows that introducing controlled bias can reduce squared-error risk (James & Stein, 1961; Hesterberg, 1995; Ionides, 2008). For GNN aggregation, this trade-off also depends on the magnitudes and directions of neighboring representations. These considerations motivate a joint design question: *how should neighborhood selection and contribution weighting be chosen together to improve aggregation quality under limited computation?*

A further challenge is sparse supervision. Acquiring dense label annotations at scale can require substantial human effort, specialized expertise, or costly experimental pipelines (Mo et al., 2022; Liu et al., 2022b; Bai et al., 2019). Figure 1 illustrates this challenge in our low-label OGBN-arXiv experiment: several baseline trajectories decline after their early peaks, with particularly pronounced

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.

deterioration for FastGCN and batched GCN. Although these trajectories do not isolate the cause of degradation, they motivate investigating how sampling and representation learning can work together when labels are scarce. In this setting, unlabeled nodes offer an additional source of information beyond the supervised training signal.

A natural way to exploit unlabeled nodes is graph self-supervised learning (You et al., 2020; Zhu et al., 2020; You et al., 2021; Zhu et al., 2021; Suresh et al., 2021). GraphCL and GRACE, for example, generate views through augmentations such as edge dropout and feature masking. Neighborhood sampling itself also produces stochastic views, raising a second question: *can the same sampling mechanism support both aggregation and learning from unlabeled nodes?* Two independent views from a fixed aggregation sampler have expected squared disagreement equal to twice the estimator variance. Through a fixed Lipschitz encoder, this identity yields an upper bound on encoded view disagreement. This connection motivates a sampler-induced contrastive objective that encourages representation consistency under the perturbations introduced by neighborhood sampling.

In this paper, we present BV-SAMPLER, a **bias–variance-aware sampling framework** that combines feature-aware neighborhood selection, contribution weighting, and sampler-induced contrastive learning. We formulate aggregation MSE in terms of sampling probabilities and per-neighbor weights, retaining the interactions between neighboring features. The analysis characterizes optimal weights for a fixed proposal and an optimal proposal for fixed weights, while a tractable diagonal surrogate provides an efficient weighting rule. We further account for fresh sampling error across GNN layers and connect the aggregation sampler to an auxiliary learning signal under sparse supervision. This formulation distinguishes exact conditional results from practical approximations while preserving a common framework for estimator design and representation learning. Our contributions are summarized as follows:

- We derive conditional optima for contribution weights and sampling probabilities, and distinguish these results from an efficient diagonal surrogate. The analysis also characterizes feature interactions and layer-wise error accumulation.
- We construct contrastive views from the aggregation sampler and establish their connection to conditional sampling variance, yielding an encoded alignment bound that motivates learning from unlabeled nodes.
- We evaluate BV-SAMPLER across supervision and inference settings, with controlled component ablations and direct bias–variance diagnostics supporting the framework's benefits in the evaluated sparse-label regimes.

2 BIAS–VARIANCE-AWARE SAMPLING

In this section, we present the theoretical foundation of bias–variance-aware sampling. We retain the aggregation-MSE formulation, derive conditional optimal weights and proposals, and distinguish these exact results from a practical surrogate. We then analyze error propagation and the variance-based motivation for sampler-induced contrastive views.

2.1 SAMPLING-BASED AGGREGATION ESTIMATOR VIA MEAN SQUARE ERROR

Let $G = (V, E)$ be a graph. Consider a node v with neighbor set $\mathcal{N}(v)$. The exact aggregation in a graph neural network layer is

$$h_v = \sum_{u \in \mathcal{N}(v)} A_{vu} h_u, \tag{1}$$

where A_{vu} is the normalized adjacency matrix and h_u is the feature vector of neighbor node u. Since computing the full aggregation is prohibitively expensive for high-degree nodes, we estimate it via sampling. Let $p(u)$ be the sampling probability of neighbor u, and suppose we sample k neighbors independently. The sampling estimator is

$$\hat{h}_v = \frac{1}{k} \sum_{i=1}^{k} \frac{A_{vu_i}}{p(u_i)} \alpha(u_i) h_{u_i}, \tag{2}$$

where $\alpha(u) \in \mathbb{R}$ is a per-node bias-control weight, whose role will become clear shortly. We measure the quality of this estimator by its mean squared error (MSE):

$$\mathrm{MSE}(\hat{h}_v) = \mathbb{E}\left[\|\hat{h}_v - h_v\|^2 \right]. \tag{3}$$

Our goal is to minimize this quantity by jointly optimizing $p(u)$ and $\alpha(u)$. The following lemma shows that the MSE of any sampling estimator decomposes additively into a (squared) bias term and a variance term, providing the foundation for the joint optimization that follows.

Lemma 2.1 (Bias–Variance Decomposition of MSE)**.** *Let $\hat{h}_v$ be any estimator of h_v. Its mean squared error decomposes as*

$$\mathrm{MSE}(\hat{h}_v) \;=\; \|\mathrm{Bias}(\hat{h}_v)\|^2 \;+\; \mathrm{Var}(\hat{h}_v), \tag{4}$$

where $\|\mathrm{Bias}(\hat{h}_v)\|^2 = \|\mathbb{E}[\hat{h}_v] - h_v\|^2$ and $\mathrm{Var}(\hat{h}_v) = \mathbb{E}\|\hat{h}_v - \mathbb{E}[\hat{h}_v]\|^2$.

See proof in Appendix C.1. Notice that when $\alpha(u) = 1$, the estimator reduces to the standard unbiased importance sampling estimator. This unbiased formulation underlies sampling methods that seek to improve aggregation accuracy by reducing estimator variance (Chen et al., 2018; Zou et al., 2019; Huang et al., 2018). Let $x_u = A_{vu}h_u$ and $m_\alpha = \sum_u \alpha(u)x_u$, the exact conditional variance is

$$\mathrm{Var}(\hat{h}_v) = \frac{1}{k}\left(\sum_u \frac{\alpha(u)^2\|x_u\|^2}{p(u)} - \|m_\alpha\|^2\right). \tag{5}$$

The mean-dependent term is retained when optimizing the exact MSE.

2.2 Optimal Biased Estimator and Sampling Distribution

Next, we want to derive optimal biased weight $\alpha^*(u)$ by plugging Equation 2 to the biased term, (i.e., $\|\mathrm{Bias}\|^2 = \|\mathbb{E}[\hat{h}_v] - h_v\|^2$):

$$\|\mathrm{Bias}\|^2 = \left\| \sum_{u\in\mathcal{N}(v)} A_{vu}\,(\alpha(u) - 1)\,h_u \right\|^2, \tag{6}$$

which yields a quadratic form involving inner products between the feature vectors of distinct neighbors:

$$\|\mathrm{Bias}\|^2 = \sum_u \sum_{u'} A_{vu}\,A_{vu'}\,(\alpha(u) - 1)\,(\alpha(u') - 1)\,\langle h_u,\, h_{u'}\rangle. \tag{7}$$

We now derive the *exact* MSE minimizer that accounts for these interactions.

Theorem 2.2 (Non-Diagonal Optimal Bias Weight)**.** *Define the Gram matrix $\mathbf{G} \in \mathbb{R}^{|\mathcal{N}(v)|\times|\mathcal{N}(v)|}$ and the diagonal regularisation matrix $\mathbf{D}$ by*

$$G_{uu'} := A_{vu}\,A_{vu'}\,\langle h_u,\, h_{u'}\rangle, \qquad D_{uu} := \frac{A_{vu}^2\,\|h_u\|^2}{k\,p(u)}, \tag{8}$$

with $D_{uu'} = 0$ for $u \neq u'$. The vector of bias weights $\boldsymbol{\alpha}^* \in \mathbb{R}^{|\mathcal{N}(v)|}$ that minimizes the exact conditional MSE for fixed p over real weights satisfies the linear system

$$((1 - 1/k)\mathbf{G} + \mathbf{D})\,\boldsymbol{\alpha}^* = \mathbf{G}\,\mathbf{1}, \tag{9}$$

whose unique solution is

$$\boldsymbol{\alpha}^* = ((1 - 1/k)\mathbf{G} + \mathbf{D})^{-1}\,\mathbf{G}\,\mathbf{1}. \tag{10}$$

See Proof in Appendix C.2. Notice that on the active set, the corrected matrix is positive definite. The solution need not be coordinatewise positive; nonnegative or bounded weights instead require a constrained quadratic optimization. Solving the linear system (9) requires constructing $\mathbf{G}$ in $O(|\mathcal{N}(v)|^2 d)$ time and solving an $|\mathcal{N}(v)| \times |\mathcal{N}(v)|$ system in $O(|\mathcal{N}(v)|^3)$ time per node v. For high-degree nodes in large-scale graphs this is prohibitive, motivating the diagonal approximation as a practical default and the non-diagonal solution as a conditional MSE oracle for low-degree nodes or offline analysis. The high computational cost of the non-diagonal solution motivates a tractable relaxation. In this subsection, we introduce a biased estimator that minimizes a diagonal surrogate.

Theorem 2.3 (Optimal Weight for a Diagonal Surrogate)**.** *Consider the sampling estimator $\hat{h}_v = \frac{1}{k}\sum_{i=1}^{k} \frac{A_{vu_i}}{p(u_i)}\alpha(u_i)h_{u_i}$. For fixed p, define $J(\alpha; p) = \sum_u A_{vu}^2\|h_u\|^2[(\alpha(u)-1)^2 + \alpha(u)^2/(kp(u))]$. The weight minimizing this surrogate satisfies*

$$\alpha^*(u) = \frac{kp(u)}{1 + kp(u)}. \tag{11}$$

This surrogate drops the negative mean-dependent variance term in Equation (5) *and uses the approximation of the squared bias*

$$\|\textit{Bias}\|^2 \approx \sum_u A_{vu}^2(\alpha(u)-1)^2\|h_u\|^2 \tag{12}$$

See proof in Appendix C.3. Replacing the Gram matrix by its diagonal in the *exact* objective gives

$$\alpha_{\mathrm{diag,exact}}(u) = \frac{kp(u)}{1+(k-1)p(u)}. \tag{13}$$

Thus, Equation (11) minimizes J, rather than the exact risk even for orthogonal contributions. The additional diagonal approximation means that J is not a general upper bound on MSE for correlated features. Appendix C.10 gives an explicit counterexample. Under diagonal G, the relative discrepancy between the exact and surrogate weights is $p(u)/[1+(k-1)p(u)] \le p(u)$.

We now derive the optimal sampling distribution that minimizes the variance of the estimator.

Theorem 2.4 (Optimal Sampling Distribution)**.** *Consider the estimator*

$$\hat{h}_v = \frac{1}{k}\sum_{i=1}^{k}\frac{A_{vu_i}}{p(u_i)}\alpha(u_i)h_{u_i}. \tag{14}$$

For fixed bias weights $\alpha(u)$, the sampling distribution that minimizes the variance satisfies

$$p^*(u) \propto A_{vu}|\alpha(u)|\|h_u\|. \tag{15}$$

See proof in Appendix C.4. The proposal is normalized over nonzero weighted contributions. These are conditional optima, not a closed-form jointly optimal pair for the exact risk. Joint optimization of the surrogate is characterized in Appendix C.11.

2.3 When an Unbiased Estimator Is Suboptimal

In Section 2.1, we show in Lemma 2.1 that the MSE of any sampling estimator is jointly determined by its bias and variance. Fixing $\alpha(u) = 1$ restricts the design to unbiased importance estimation, while allowing bias enlarges the class of estimators. This raises a natural question: *does enforcing unbiasedness actually minimize the MSE?* The following theorem answers this question when sampling variance is positive.

Theorem 2.5 (Unbiased Estimator is Suboptimal)**.** *For a fixed proposal, let Y be the unbiased aggregation estimator with target h_v and variance $V > 0$. There exists a scalar $a \in [0,1)$ such that aY has strictly lower MSE. In particular,*

$$a^* = \frac{\|h_v\|^2}{\|h_v\|^2+V}, \qquad \mathrm{MSE}(a^*Y) = \frac{\|h_v\|^2 V}{\|h_v\|^2+V} < V. \tag{16}$$

When $V = 0$, the unbiased estimator already has zero risk.

See proof in Appendix C.5. Theorem 2.5 establishes an oracle existence result under squared-error risk. The scalar depends on the target and variance, so it does not guarantee improvement by the practical weights or in downstream accuracy. It motivates considering both p and α without imposing unbiasedness as a necessary design constraint.

2.4 Representation Error Propagation in Deep GNNs

Now, we analyze how estimation error accumulates across multiple GNN layers. Consider an L-layer GNN:

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} A_{vu}\, h_u^{(l)}\, W^{(l)}\right). \tag{17}$$

Let $\hat{h}^{(l)}$ denote the representation computed using sampling, and define the *total expected error* at layer l as $\epsilon^{(l)} := \max_v \mathbb{E}\left\|\hat{h}_v^{(l)} - h_v^{(l)}\right\|$. Note that $\epsilon^{(l)}$ accounts for both bias and the effect of sampling variance propagated from earlier layers, and satisfies $\epsilon^{(l)} \ge \|\mathbb{E}[\hat{h}^{(l)}] - h^{(l)}\|$ by the triangle inequality.

Theorem 2.6 (Representation Error Propagation). *Assume σ is L_σ-Lipschitz and the network weights are fixed. Let $a = \max_v \sum_u A_{vu}$ and let*

$$\delta_l = \max_v \left(\mathbb{E} \left\| \hat{\mu}_v^{(l)} - \sum_u A_{vu} \hat{h}_u^{(l)} \right\|^2 \right)^{1/2}$$

measure fresh aggregation error on the current approximate features, including local bias and variance. Then

$$\epsilon^{(l+1)} \leq L_\sigma \|W^{(l)}\| (a\epsilon^{(l)} + \delta_l). \tag{18}$$

Let $\gamma_l = L_\sigma \|W^{(l)}\| a$ and $b_l = L_\sigma \|W^{(l)}\| \delta_l$, we obtain

$$\epsilon^{(L)} \leq \left(\prod_{l=0}^{L-1} \gamma_l \right) \epsilon^{(0)} + \sum_{j=0}^{L-1} b_j \prod_{l=j+1}^{L-1} \gamma_l. \tag{19}$$

See proof in Appendix C.6. Fresh error cannot be omitted when every layer samples anew. Symmetric adjacency normalization does not in general imply $a = 1$. This bound motivates local error control without claiming global optimality of the trained network.

Theorem 2.7 (Unbiased Risk and Bounds). *Let $c_u := A_{vu}\|h_u\|$ and $C := \sum_{u \in \mathcal{N}(v)} c_u$.*
(i) Lower bound for the unbiased estimator. *For any unbiased estimator of the form in Equation* (2) *($\alpha(u) \equiv 1$) using k i.i.d. samples under any sampling distribution p,*

$$\mathrm{MSE}_{\mathrm{unbiased}}(\hat{h}_v) \;\geq\; \frac{C^2 - \|h_v\|^2}{k}. \tag{20}$$

(ii) Surrogate upper bound. *With $p_0(u) = c_u/C$ and J as defined above,*

$$\min_{p,\alpha} J(\alpha; p) \leq J(\alpha^*(p_0); p_0) = \sum_u \frac{c_u^2}{1 + k c_u / C} \leq \frac{C^2}{k}. \tag{21}$$

This is a bound on the surrogate, which also bounds the exact risk of the surrogate-optimal pair when the Gram matrix is diagonal.

(iii) Achievable exact risk. *Let $M = \|h_v\|^2$ and $V_{\min} = (C^2 - M)/k$. If $V_{\min} > 0$, an oracle biased estimator satisfies*

$$\inf_{p,\alpha} \mathrm{MSE}(\hat{h}_v) \leq \frac{M V_{\min}}{M + V_{\min}} < V_{\min}. \tag{22}$$

The numerator $C^2 - M = 2\sum_{u<u'} (\|x_u\|\|x_{u'}\| - \langle x_u, x_{u'} \rangle)$ exposes the role of feature directions. It can be zero, so no positive universal $\Theta(1/k)$ lower rate or strict dominance by the practical weights is asserted.

See proof in Appendix C.7.

2.5 Sampler-Augmented Contrastive Learning under Label Scarcity

The proposed framework studies sampling and weighting under an aggregation-MSE objective. In realistic settings, labels may be scarce, motivating the use of unlabeled nodes through graph self-supervised learning (You et al., 2020; Zhu et al., 2020). GraphCL and GRACE construct views through augmentations such as edge dropping and feature masking, while adaptive methods refine those choices (You et al., 2021; Zhu et al., 2021; Suresh et al., 2021). We instead investigate views generated by the aggregation sampler itself, whose conditional variability can be analyzed directly.

To exploit this connection, we propose a contrastive learning framework that is intrinsically coupled with the importance sampler itself rather than built upon external augmentations. Instead of generating views through random masking or edge perturbation, we construct two stochastic views directly from independent samples drawn from a fixed proposal p. Concretely, two independent forward passes through the sampler produce

$$\hat{h}_v^{(s)} = \frac{1}{k} \sum_{i=1}^{k} \frac{A_{v u_i^{(s)}}}{p(u_i^{(s)})} \alpha(u_i^{(s)}) h_{u_i^{(s)}}, \quad s = 1, 2, \tag{23}$$

where $u_i^{(s)}$ are i.i.d. samples drawn from p. The sampled representations are then passed through an encoder f_θ and projection head g_ϕ:

$$z_v^{(s)} = g_\phi\left(f_\theta\left(\hat{h}_v^{(s)}\right)\right), \quad z_v^{(s)} \in \mathbb{S}^{d-1}. \tag{24}$$

This construction is motivated by the alignment component of contrastive learning (Oord et al., 2018; Wang & Isola, 2020). For normalized outputs, a standard symmetric paired-view objective is

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2N}\sum_{v,s}\log\frac{\exp(\langle z_v^{(s)}, z_v^{(3-s)}\rangle/\tau)}{\sum_{(w,t)\neq(v,s)}\exp(\langle z_v^{(s)}, z_w^{(t)}\rangle/\tau)}. \tag{25}$$

Its negative positive-pair similarity term encourages alignment; the remaining log-partition term includes the positive view and the other nodes' views. We analyze positive-pair disagreement rather than identify the full loss with aggregation variance. The following proposition formalizes the connection. The supplied implementation uses a cross-view objective variant; the identity below concerns the sampled views themselves.

Proposition 2.8 (View Disagreement and Estimator Variance)**.** *Let the two views $\hat{h}_v^{(1)}, \hat{h}_v^{(2)}$ be drawn independently from the same importance sampler, conditional on fixed features, weights, and proposal, with finite second moments. Then the unencoded view disagreement equals exactly twice the estimator variance,*

$$\mathbb{E}\left[\|\hat{h}_v^{(1)} - \hat{h}_v^{(2)}\|^2\right] \;=\; 2\cdot\mathrm{Var}(\hat{h}_v). \tag{26}$$

Furthermore, if the composite map $g_\phi \circ f_\theta$ is L-Lipschitz on the view support, including output normalization, and held fixed, then the encoded view disagreement is bounded by

$$\mathbb{E}\left[\|z_v^{(1)} - z_v^{(2)}\|^2\right] \;\le\; 2L^2\cdot\mathrm{Var}(\hat{h}_v). \tag{27}$$

The proof is provided in Appendix C.8. Proposition 2.8 establishes an exact identity for unencoded views and an upper bound after encoding. The bound cannot be reversed to conclude that minimizing the encoded alignment loss reduces unencoded estimator variance. It motivates sampler-induced learning, whose downstream benefit is evaluated empirically. The next proposition identifies a conditional optimum for the variance-induced alignment bound.

Proposition 2.9 (Optimal Sampling Minimizes the Alignment-Loss Bound)**.** *Suppose the composite map $g_\phi \circ f_\theta$ is L-Lipschitz with output on the unit sphere $\mathbb{S}^{d-1}$. Then the InfoNCE alignment loss admits the variance-induced upper bound*

$$\mathbb{E}\left[1 - \mathrm{sim}(z_v^{(1)}, z_v^{(2)})\right] \;\le\; L^2\,\mathrm{Var}_p(\hat{h}_v), \tag{28}$$

For fixed features, weights, encoder, and a common valid Lipschitz bound, the proposal in Theorem 2.4 minimizes this upper bound over p.

See proof in Appendix C.9. Equation (28) shows that the optimal estimator-theoretic distribution p^* and the sampling distribution that minimizes the variance bound on the alignment loss coincide. We note that this is a bound, not an exact characterization: minimizing the upper bound does not generally minimize the alignment loss itself, and a small Lipschitz constant alone does not establish tightness.

3 EXPERIMENTS

In this section, we experimentally analyze and compare BV-SAMPLER with graph sampling baselines. We aim to answer the following research questions: *Q1: How does* BV-SAMPLER *compare with existing methods in effectiveness and efficiency? Q2: How does it perform under sparse supervision? Q3: How does it perform when inference also uses sampled neighborhoods? Q4: How does each component contribute to the framework?*

3.1 EXPERIMENTAL SETUP

Datasets: We conduct our main experiments on Cora (Sen et al., 2008), PubMed (Sen et al., 2008), Reddit (Hamilton et al., 2017), and OGBN-arXiv (Hu et al., 2020), and report an additional sampled-inference comparison on PPI-large. Dataset statistics and the custom training, validation, and test splits are provided in Table 5 and Appendix D.2.
Baseline Methods: We compare BV-SAMPLER with GraphSAGE (Hamilton et al., 2017), FastGCN (Chen et al., 2018), GraphSAINT (Zeng et al., 2020), LADIES (Zou et al., 2019), LABOR (Balin & Çatalyürek, 2023), and MVS-GNN (Cong et al., 2020). We also include full-graph GCN as a reference and DGI (Veličković et al., 2018) as an additional self-supervised objective in the sampled-inference comparison.
Metrics: Following FastGCN Chen et al. (2018), we report Micro-F1 in percent and training time per epoch, including sampling, to evaluate effectiveness and efficiency. The average columns give unweighted means across datasets. Full-graph and sampled-inference results are reported separately.
Implementation Details: We cache node feature norms and refresh entries for visited nodes during training to reduce proposal-construction overhead. Algorithm 1 and Appendix D.2 provide the algorithm and implementation settings. The code is available at Anonymous GitHub[1].

Table 1: Rich-label comparison with full-graph inference. Micro-F1 is in percent and time is seconds per epoch including sampling.

Methods	Cora		PubMed		Reddit		OGBN-arXiv		Average	
	Micro-F1	Time	Micro-F1	Time	Micro-F1	Time	Micro-F1	Time	Micro-F1	Time
GCN (Full Graph)	86.86±0.15	0.0031	88.20±0.13	0.0036	**94.77±0.11**	0.3228	**71.24±0.20**	0.0488	85.27	0.0945
GraphSAGE	85.86±0.76	0.0157	**88.80±0.36**	0.0152	94.48±0.25	0.1871	69.59±0.38	0.1266	84.68	0.0861
FastGCN	**87.78±0.15**	0.0053	87.90±0.10	0.0063	93.83±0.08	0.0598	70.29±0.30	0.0689	84.95	0.0350
GraphSAINT	80.28±0.37	0.0074	88.18±0.51	0.0495	93.66±0.38	0.1218	70.50±0.43	0.0816	83.16	0.0650
LADIES	86.68±0.63	0.0456	86.30±0.50	0.0554	91.78±0.18	3.3384	55.15±0.50	0.2648	79.98	0.9260
LABOR	87.08±0.30	0.0054	86.80±0.77	0.0067	90.91±0.58	**0.0591**	62.90±0.29	0.0528	81.92	0.0310
MVS-GNN	87.51±0.36	0.0025	86.60±0.33	0.0022	93.53±0.30	0.1710	64.34±0.36	0.0383	83.00	0.0535
BV-SAMPLER	**87.78±0.08**	**0.0021**	88.20±0.42	**0.0020**	94.29±0.11	0.0783	71.01±0.31	**0.0295**	**85.32**	**0.0280**

3.2 EFFECTIVENESS AND EFFICIENCY OF BV-SAMPLER (Q1 & Q2)

We evaluate all methods under both rich and sparse supervision using full-graph inference. Tables 1 and 2 report Micro-F1 and per-epoch training time. In the rich-label setting, BV-SAMPLER achieves an average Micro-F1 of 85.32%, compared with 85.27% for full-graph GCN and 84.95% for FastGCN. It ties FastGCN on Cora and remains competitive on the other datasets, although GraphSAGE achieves the highest PubMed score and full-graph GCN achieves the highest Reddit and OGBN-arXiv scores. These results show that the proposed framework maintains competitive predictive performance across the evaluated datasets; the small average differences do not establish statistical superiority.

In the sparse-label setting (Table 2), BV-SAMPLER achieves the highest Micro-F1 among the sampling-based methods on all three datasets: 84.92% on PubMed, 93.32% on Reddit, and 67.63% on OGBN-arXiv. Relative to FastGCN, these correspond to improvements of 0.74, 0.25, and 0.88 percentage points, respectively. The average score increases from 81.33% for FastGCN to 81.96% for BV-SAMPLER. Full-graph GCN remains a strong reference, including a higher score on Reddit. Thus, the benefit under sparse supervision is supported by the comparisons with sampling-based methods without requiring superiority to full-graph training on every dataset.

The reported runtimes also indicate competitive efficiency. In the rich-label table, BV-SAMPLER has the lowest per-epoch time on Cora, PubMed, and OGBN-arXiv, while LABOR is faster on Reddit. In the sparse-label table, BV-SAMPLER is fastest on PubMed; FastGCN is faster on Reddit and LABOR is faster on OGBN-arXiv. Across datasets and two label settings, BV-SAMPLER has the lowest arithmetic-average runtime in both tables. These averages summarize the recorded configurations rather than establish a uniform speedup across graph sizes.

[1] https://anonymous.4open.science/r/BV-Sampler-129A

Table 2: Sparse-label comparison with full-graph inference, using the recorded custom splits. Micro-F1 is in percent and time is seconds per epoch.

Methods	PubMed		Reddit		OGBN-arXiv		Average	
	Micro-F1	Time	Micro-F1	Time	Micro-F1	Time	Micro-F1	Time
GCN (Full Graph)	84.50±0.07	0.0040	93.66±0.03	0.3307	67.58±0.12	0.0409	81.91	0.1252
GraphSAGE	83.10±0.51	0.0151	91.73±0.21	0.1879	66.05±0.09	0.1266	80.29	0.1098
FastGCN	84.18±0.08	0.0062	93.07±0.10	**0.0523**	66.75±0.30	0.0602	81.33	0.0395
GraphSAINT	82.00±0.80	0.0083	91.66±0.17	0.1282	63.49±0.29	0.0881	79.05	0.0748
LADIES	82.92±0.55	0.0556	90.92±0.45	3.0416	53.03±0.55	0.2429	75.62	1.1133
LABOR	81.95±0.41	0.0030	90.42±0.51	0.2577	59.80±0.34	0.0123	77.39	0.0910
MVS-GNN	81.95±0.41	0.0029	92.97±0.31	0.1080	59.90±0.35	0.0300	78.27	0.0469
BV-SAMPLER	**84.92 ± 0.19**	**0.0020**	**93.32 ± 0.04**	0.0788	**67.63 ± 0.30**	**0.0294**	**81.96**	**0.0367**

Table 3: Batch-only inference under the sparse-label setting on OGBN-arXiv and PPI-large datasets.

Method	OGBN-arXiv		PPI-large	
	Micro-F1	With additional DGI	Micro-F1	Time (s)
GCN (Batch)	39.31±0.47	52.58±0.31	67.06±0.12	0.0662
GraphSAGE	54.42±0.34	61.63±0.27	66.52±0.36	0.0841
FastGCN	53.50±0.17	62.95±0.52	67.27±0.34	0.0634
GraphSAINT	56.92±0.11	63.62±0.46	66.71±0.43	0.0461
LADIES	48.69±0.53	60.72±0.34	67.90±0.12	0.0820
MVS-GNN	55.73±0.55	61.49±0.41	62.27±0.77	0.193
LABOR	59.80±0.34	61.47±0.29	68.08±0.19	0.0364
BV-SAMPLER	66.57±0.23	66.48±0.21	69.06±0.02	0.0478

3.3 PERFORMANCE UNDER SAMPLED INFERENCE (Q3)

The sparse-label training trajectories in Figure 1 motivate evaluating the framework when inference is also restricted to sampled neighborhoods. This setting measures prediction without full-graph aggregation at evaluation time. Table 3 reports the OGBN-arXiv results, and Appendix D.3 provides the additional Reddit comparison. On OGBN-arXiv, BV-SAMPLER achieves 66.57% Micro-F1, exceeding LABOR without additional DGI by 6.77 percentage points. Adding DGI improves the baseline scores, but the strongest baseline with DGI, GraphSAINT, reaches 63.62%, leaving a 2.95-point gap. Adding DGI to BV-SAMPLER yields 66.48%, providing no additional improvement in this experiment. On Reddit, the reported batch-only score of 92.79% exceeds the strongest listed baseline, GraphSAGE, by 2.53 points. These results support the usefulness of the full framework in the evaluated sparse-label, sampled-inference settings. They do not by themselves isolate which component causes the improvement or establish an inherent failure of the baseline methods. As for the batch-only inference on PPI-large dataset, BV-SAMPLER achieves 69.06% Micro-F1, compared with 68.08% for LABOR, at reported times of 0.0478 and 0.0364 seconds, respectively. This comparison illustrates an accuracy–runtime trade-off rather than a gain in both measures.

3.4 ABLATION STUDY AND SENSITIVITY (Q4)

To examine the contribution of each component, we conduct the ablation study on OGBN-arXiv dataset and the results are presented in Table 4. The first group crosses the sampling proposal with contribution weighting while retaining the sampler-induced contrastive objective. Starting from structure-only sampling with $\alpha \equiv 1$, feature-aware sampling increases Micro-F1 from 65.28% to 67.21%, a gain of 1.93%. Under the same structure-only proposal, biased weighting increases the score to 66.99%, a gain of 1.71 points. Combining feature-aware sampling and biased weighting yields the highest score, 67.63%, improving on the common reference by 2.35 %. The matched comparisons further show that biased weighting adds 0.42 points to feature-aware sampling, while feature-aware sampling adds 0.64 points when biased weights are already used. Both components therefore contribute in the evaluated setting. Their combined gain is smaller than the sum of their separate gains from the common reference, indicating overlapping benefits rather than a super-additive interaction. The second group compares auxiliary objectives while retaining feature-aware sampling and biased weighting. Replacing the sampler-induced objective with GraphCL or DGI gives 66.07% and 67.23%, respectively, compared with 67.63% for the full model. Removing sampling InfoNCE

Table 4: Sparse-label OGBN-arXiv ablation with full-graph inference. The upper block separates sampling proposal and contribution weighting while retaining the sampler-induced contrastive objective. The lower block varies the auxiliary objective with feature-aware sampling and biased weighting.

Sampling proposal	**Contribution weighting**	**Micro-F1 (%)**
Structure-only ($p \propto A_{vu}$)	$\alpha \equiv 1$	65.28 ± 0.44
Feature-aware	$\alpha \equiv 1$	67.21 ± 0.24
Structure-only ($p \propto A_{vu}$)	Biased weights	66.99 ± 0.12
Feature-aware (full BV-SAMPLER)	Biased weights	67.63 ± 0.30
Auxiliary objective variant		
BV-SAMPLER-GraphCL		66.07 ± 0.19
BV-SAMPLER-DGI		67.23 ± 0.32
BV-SAMPLER-S (no sampling InfoNCE)		66.98 ± 0.17

gives 66.98%, a decrease of 0.65 points. The results indicate a useful, modest contribution from sampler-induced learning in this setting, rather than a universal advantage over other self-supervised objectives.

Sensitivity and Budget Analysis: Due to space limit, we report the performance and runtime across different sampling budgets and contrastive coefficients in Appendix D.6.

4 Related Work

GNN sampling and variance reduction. GraphSAGE, FastGCN, adaptive sampling, LADIES, Cluster-GCN, and GraphSAINT address different sampling units and computational constraints (Hamilton et al., 2017; Chen et al., 2018; Huang et al., 2018; Zou et al., 2019; Chiang et al., 2019; Zeng et al., 2020). MVS-GNN analyzes forward embedding and backward gradient variance and uses adaptive sampling and variance reduction (Cong et al., 2020). LABOR addresses neighborhood explosion through layer-neighbor sampling (Balin & Çatalyürek, 2023). Norm-dependent importance proposals therefore have substantial prior work. Our emphasis is the aggregation-MSE formulation connecting proposal design and contribution weighting, its conditional and approximate solutions, and its connection to sampler-induced views. Norm-proportional proposals are an established ingredient of this design.

Shrinkage and biased importance sampling. The bias–variance trade-off is established in statistical estimation. James–Stein estimation provides a classical example under multivariate normal mean estimation and quadratic loss (James & Stein, 1961). Weighted and self-normalized importance sampling, bias-reduced variants, and truncation address related stability and risk issues (Hesterberg, 1995; Mahmood et al., 2014; Cardoso et al., 2022; Ionides, 2008). Our per-neighbor factor is a smooth function of expected sample count. It differs operationally from a common self-normalization denominator or hard clipping, but these differences do not imply a universal ordering of their risks. We study these principles within a GNN framework that considers proposal design, contribution weighting, and sampler-induced representation learning together.

Graph self-supervision. GraphCL (You et al., 2020) and GRACE (Zhu et al., 2020) use augmented views for contrastive learning. GCA (Zhu et al., 2021), JOAO (You et al., 2021), and AD-GCL (Suresh et al., 2021) adapt augmentation choices. We instead reuse the stochastic aggregation mechanism to generate views. Our conditional variance identity motivates this construction but does not show that other augmentations are incompatible with importance sampling or that their learning objectives cannot respond to sampling noise.

5 Conclusion

In this paper, we presented BV-SAMPLER, a bias–variance-aware GNN framework combining feature-aware sampling, contribution weighting, and sampler-induced contrastive learning. Our analysis distinguishes exact conditional MSE optima from a practical surrogate, accounts for layer-wise error accumulation, and connects sampling variance to representation consistency. Experiments demonstrate competitive performance under rich supervision and gains in the evaluated sparse-label, sampled-inference settings. Component ablations and direct bias–variance diagnostics further illustrate how sampling and weighting contribute to prediction and aggregation quality.

REFERENCES

Yunsheng Bai, Hao Ding, Yang Qiao, Agustin Marinovic, Ken Gu, Ting Chen, Yizhou Sun, and Wei Wang. Unsupervised inductive graph-level representation learning via graph-graph proximity. *arXiv preprint arXiv:1904.01098*, 2019.

Muhammed Fatih Balin and Ümit Çatalyürek. Layer-neighbor sampling—defusing neighborhood explosion in gnns. In *Advances in Neural Information Processing Systems*, volume 36, pp. 25819–25836, 2023.

Gabriel Cardoso, Sergey Samsonov, Achille Thin, Eric Moulines, and Jimmy Olsson. BR-SNIS: Bias reduced self-normalized importance sampling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 716–729, 2022.

Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: Fast learning with graph convolutional networks via importance sampling. In *International Conference on Learning Representations*, 2018.

Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 257–266, 2019.

Weilin Cong, Rana Forsati, Mahmut Kandemir, and Mehrdad Mahdavi. Minimal variance sampling with provable guarantees for fast training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1393–1403, 2020. doi: 10.1145/3394486.3403192.

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The world wide web conference*, pp. 417–426, 2019.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995. doi: 10.1080/00401706.1995.10484303.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Wenbing Huang, Tong Zhang, Yu Rong, and Junzhou Huang. Adaptive sampling towards fast graph representation learning. *Advances in neural information processing systems*, 31, 2018.

Edward L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008. doi: 10.1198/106186008X320456.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 361–379. University of California Press, 1961.

Jiaying Liu, Feng Xia, Xu Feng, Jing Ren, and Huan Liu. Deep graph learning for anomalous citation detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2543–2557, 2022a.

Yixin Liu, Yu Zheng, Daokun Zhang, Hongxu Chen, Hao Peng, and Shirui Pan. Towards unsupervised deep graph structure learning. In *Proceedings of the ACM web conference 2022*, pp. 1392–1403, 2022b.

A. Rupam Mahmood, Hado van Hasselt, and Richard S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Yujie Mo, Liang Peng, Jie Xu, Xiaoshuang Shi, and Xiaofeng Zhu. Simple unsupervised graph representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 7797–7805, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

Susheel Suresh, Pan Li, Cong Hao, and Jennifer Neville. Adversarial graph augmentation to improve graph contrastive learning. *Advances in Neural Information Processing Systems*, 34:15920–15933, 2021.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

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.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *International conference on machine learning*, pp. 12121–12132. PMLR, 2021.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. Graphsaint: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the web conference 2021*, pp. 2069–2080, 2021.

Difan Zou, Ziniu Hu, Yewen Wang, Song Jiang, Yizhou Sun, and Quanquan Gu. Layer-dependent importance sampling for training deep and large graph convolutional networks. *Advances in neural information processing systems*, 32, 2019.

A AI USE STATEMENT

In this work, we used generative AI tools for writing, editing, or formatting purposes, including improving readability and checking the typos and notation consistency. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have reviewed all AI-assisted work. We checked LLM-generated polishing of the writing and manually verified the content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

B TIME COMPLEXITY

For k sampled contributions, computing the correction adds $O(k)$ arithmetic to an $O(kd)$ aggregation. If a proposal is constructed over n_v candidates from cached scores, score construction and normalization cost $O(n_v)$; alias-table construction and k draws cost $O(n_v + k)$. Thus the aggregation and proposal work costs $O(n_v + kd)$, excluding the remaining GNN transformations. One scalar norm per node requires $O(|V|)$ storage, $O(|V|d)$ initialization, and $O(d)$ per refreshed entry. Layer-specific caches multiply storage by the number of cached layers. Contrastive learning requires two sampled views for participating nodes, with overhead depending on forward-pass reuse.

C PROOF

C.1 PROOF OF LEMMA 2.1

Proof. We start from the definition of mean squared error:

$$\text{MSE} = \mathbb{E}\left[\|\hat{h}_v - h_v\|^2\right]. \tag{29}$$

Add and subtract $\mathbb{E}[\hat{h}_v]$:

$$\hat{h}_v - h_v = (\hat{h}_v - \mathbb{E}[\hat{h}_v]) + (\mathbb{E}[\hat{h}_v] - h_v). \tag{30}$$

Then,

$$\|\hat{h}_v - h_v\|^2 = \|\hat{h}_v - \mathbb{E}[\hat{h}_v]\|^2 + \|\mathbb{E}[\hat{h}_v] - h_v\|^2 + 2(\hat{h}_v - \mathbb{E}[\hat{h}_v])^\top(\mathbb{E}[\hat{h}_v] - h_v). \tag{31}$$

Taking expectation:

$$\mathbb{E}\left[(\hat{h}_v - \mathbb{E}[\hat{h}_v])\right] = 0, \tag{32}$$

so the cross term vanishes:

$$\mathbb{E}\left[(\hat{h}_v - \mathbb{E}[\hat{h}_v])^\top(\mathbb{E}[\hat{h}_v] - h_v)\right] = 0. \tag{33}$$

Therefore,

$$\mathbb{E}\|\hat{h}_v - h_v\|^2 = \|\mathbb{E}[\hat{h}_v] - h_v\|^2 + \mathbb{E}\|\hat{h}_v - \mathbb{E}[\hat{h}_v]\|^2. \tag{34}$$

This completes the proof. □

C.2 PROOF OF THEOREM 2.2

Proof. Substituting the true squared bias (7) and the variance expression from Section 2.1, the MSE is

$$\text{MSE}(\boldsymbol{\alpha}) = \sum_{u,u'} G_{uu'}\,(\alpha(u) - 1)\,(\alpha(u') - 1) + \sum_u D_{uu}\,\alpha(u)^2 - \frac{1}{k}\boldsymbol{\alpha}^\top \mathbf{G}\boldsymbol{\alpha}. \tag{35}$$

In matrix notation, writing $\boldsymbol{\beta} := \boldsymbol{\alpha} - \mathbf{1}$:

$$\mathrm{MSE} = \boldsymbol{\beta}^\top \mathbf{G}\, \boldsymbol{\beta} + (\boldsymbol{\beta} + \mathbf{1})^\top \mathbf{D}\, (\boldsymbol{\beta} + \mathbf{1}) - \frac{1}{k} (\boldsymbol{\beta} + \mathbf{1})^\top \mathbf{G} (\boldsymbol{\beta} + \mathbf{1}). \tag{36}$$

Expanding and collecting terms in $\boldsymbol{\alpha} = \boldsymbol{\beta} + \mathbf{1}$:

$$\mathrm{MSE} = \boldsymbol{\alpha}^\top ((1 - 1/k)\mathbf{G} + \mathbf{D})\, \boldsymbol{\alpha} - 2\, \boldsymbol{\alpha}^\top \mathbf{G}\, \mathbf{1} + \mathbf{1}^\top \mathbf{G}\, \mathbf{1}. \tag{37}$$

Since $\mathbf{G}$ is a Gram matrix it is positive semi-definite, and $\mathbf{D}$ is positive definite (as $A_{vu} > 0$, $\|h_u\| > 0$, and $p(u) > 0$ for active neighbors), with $k \geq 1$, so $(1 - 1/k)\mathbf{G} + \mathbf{D}$ is positive definite and hence invertible. The MSE in (37) is therefore a strictly convex quadratic in $\boldsymbol{\alpha}$. Setting its gradient to zero:

$$\nabla_{\boldsymbol{\alpha}} \mathrm{MSE} = 2\, ((1 - 1/k)\mathbf{G} + \mathbf{D})\, \boldsymbol{\alpha} - 2\, \mathbf{G}\, \mathbf{1} = \mathbf{0}, \tag{38}$$

which yields the linear system (9) and its solution (10). □

C.3 Proof of Theorem 2.3

Proof. The MSE can be written as

$$\mathrm{MSE} = \|\mathrm{Bias}\|^2 + \mathrm{Var}. \tag{39}$$

The squared bias is approximately

$$\|\mathrm{Bias}\|^2 \approx \sum_u A_{vu}^2 (\alpha(u) - 1)^2 \|h_u\|^2. \tag{40}$$

Dropping the negative mean-dependent term gives the variance upper approximation

$$\widetilde{V} = \frac{1}{k} \sum_u \frac{A_{vu}^2 \alpha(u)^2}{p(u)} \|h_u\|^2. \tag{41}$$

Combining this with the diagonal bias approximation defines the surrogate

$$J = \sum_u A_{vu}^2 (\alpha(u) - 1)^2 \|h_u\|^2 + \frac{1}{k} \sum_u \frac{A_{vu}^2 \alpha(u)^2}{p(u)} \|h_u\|^2. \tag{42}$$

We minimize this with respect to $\alpha(u)$. Take derivative:

$$\frac{\partial J}{\partial \alpha(u)} = 2 A_{vu}^2 (\alpha(u) - 1) \|h_u\|^2 + \frac{2}{k} \frac{A_{vu}^2 \alpha(u)}{p(u)} \|h_u\|^2. \tag{43}$$

Set derivative to zero:

$$A_{vu}^2 (\alpha(u) - 1) + \frac{1}{k} \frac{A_{vu}^2 \alpha(u)}{p(u)} = 0. \tag{44}$$

Solve for $\alpha(u)$:

$$\alpha(u) \left(1 + \frac{1}{kp(u)}\right) = 1. \tag{45}$$

Thus,

$$\alpha^*(u) = \frac{1}{1 + \frac{1}{kp(u)}} = \frac{kp(u)}{1 + kp(u)}. \tag{46}$$

which completes the proof. □

C.4 Proof of Theorem 2.4

Proof. From the variance expression:

$$\mathrm{Var}(\hat{h}_v) = \frac{1}{k}\left[\sum_u \frac{A_{vu}^2 \alpha(u)^2}{p(u)} \|h_u\|^2 - \left\|\sum_u A_{vu}\alpha(u) h_u\right\|^2\right]. \tag{47}$$

Since the second term does not depend on $p(u)$, minimizing the variance is equivalent to solving

$$\min_{p(u)} \sum_u \frac{c_u}{p(u)}, \qquad \text{subject to} \quad \sum_u p(u) = 1, \quad p(u) \geq 0, \tag{48}$$

where we define

$$c_u := A_{vu}^2\, \alpha(u)^2\, \|h_u\|^2 \;\geq\; 0. \tag{49}$$

If $c_u = 0$ for some u (e.g., $A_{vu} = 0$ or $h_u = 0$), then the term $c_u/p(u)$ contributes zero to the objective for any $p(u) \geq 0$, and assigning $p^*(u) = 0$ is optimal. We therefore restrict attention to the active set $\mathcal{S} := \{u \in \mathcal{N}(v) : c_u > 0\}$ and solve over $p(u) > 0$ for $u \in \mathcal{S}$.

Form the Lagrangian for the restricted problem:

$$\mathcal{L} = \sum_{u\in\mathcal{S}} \frac{c_u}{p(u)} + \lambda\left(\sum_{u\in\mathcal{S}} p(u) - 1\right). \tag{50}$$

Taking the partial derivative with respect to $p(u)$ and setting it to zero:

$$\frac{\partial \mathcal{L}}{\partial p(u)} = -\frac{c_u}{p(u)^2} + \lambda = 0 \quad \Longrightarrow \quad p(u)^2 = \frac{c_u}{\lambda}. \tag{51}$$

Since $p(u) > 0$ and $c_u > 0$, we require $\lambda > 0$, and we take the positive square root:

$$p(u) = \frac{\sqrt{c_u}}{\sqrt{\lambda}}. \tag{52}$$

Applying the normalization constraint $\sum_{u\in\mathcal{S}} p(u) = 1$ determines $\sqrt{\lambda} = \sum_{u'\in\mathcal{S}} \sqrt{c_{u'}}$, giving

$$p^*(u) = \frac{\sqrt{c_u}}{\sum_{u'\in\mathcal{S}} \sqrt{c_{u'}}}. \tag{53}$$

Substituting back $c_u = A_{vu}^2\, \alpha(u)^2\, \|h_u\|^2$:

$$p^*(u) \propto \sqrt{A_{vu}^2\, \alpha(u)^2\, \|h_u\|^2} = A_{vu}\, |\alpha(u)|\, \|h_u\|. \tag{54}$$

This completes the proof. □

C.5 Proof of Theorem 2.5

Proof. The MSE as a function of α is

$$\mathrm{MSE}(\alpha) = \mathrm{Bias}^2(\alpha) + \mathrm{Var}(\alpha). \tag{55}$$

The optimal weight α^* is defined as

$$\alpha^* = \arg\min_{\alpha} \mathrm{MSE}(\alpha). \tag{56}$$

Since α^* achieves the minimum MSE over *all* choices of α, and since $\alpha(u) = 1$ is one particular choice, it follows immediately that

$$\mathrm{MSE}(\alpha^*) \leq \mathrm{MSE}(\alpha = 1) = \mathrm{MSE}_{\text{unbiased}}. \tag{57}$$

Let $\mu = h_v$ and $V = \mathbb{E}\|Y - \mu\|^2$. Unbiasedness of Y gives

$$\mathbb{E}\|aY - \mu\|^2 = (a-1)^2\|\mu\|^2 + a^2 V.$$

For $V > 0$, this strictly convex quadratic is minimized at $a^* = \|\mu\|^2/(\|\mu\|^2 + V)$. Substitution yields the stated risk, which is strictly less than V. If $\mu = 0$, the optimum is $a^* = 0$; if strictly positive factors are required, any $a \in (0, 1)$ still improves on $a = 1$. This completes the proof. □

C.6 Proof of Theorem 2.6

Proof. Let

$$\mu_v^{(l)} = \sum_u A_{vu} h_u^{(l)}, \qquad \widetilde{\mu}_v^{(l)} = \sum_u A_{vu} \hat{h}_u^{(l)}$$

denote the exact aggregation of the reference and sampled representations, respectively. Let $\hat{\mu}_v^{(l)}$ be the freshly sampled aggregation of the current sampled representations. The next-layer representations are

$$h_v^{(l+1)} = \sigma\left(\mu_v^{(l)} W^{(l)}\right), \qquad \hat{h}_v^{(l+1)} = \sigma\left(\hat{\mu}_v^{(l)} W^{(l)}\right).$$

All expectations below include randomness from previous layers and the fresh sampling at layer l. We assume the relevant second moments are finite.

Adding and subtracting $\widetilde{\mu}_v^{(l)}$ gives

$$\hat{\mu}_v^{(l)} - \mu_v^{(l)} = \left(\hat{\mu}_v^{(l)} - \widetilde{\mu}_v^{(l)}\right) + \sum_u A_{vu} \left(\hat{h}_u^{(l)} - h_u^{(l)}\right). \tag{58}$$

Since $A_{vu} \geq 0$, the triangle inequality implies

$$\begin{aligned}
\mathbb{E}\left\|\hat{\mu}_v^{(l)} - \mu_v^{(l)}\right\| &\leq \mathbb{E}\left\|\hat{\mu}_v^{(l)} - \widetilde{\mu}_v^{(l)}\right\| + \sum_u A_{vu} \mathbb{E}\left\|\hat{h}_u^{(l)} - h_u^{(l)}\right\| \\
&\leq \left(\mathbb{E}\left\|\hat{\mu}_v^{(l)} - \widetilde{\mu}_v^{(l)}\right\|^2\right)^{1/2} + \left(\sum_u A_{vu}\right) \epsilon^{(l)} \\
&\leq \delta_l + a\epsilon^{(l)}.
\end{aligned} \tag{59}$$

The second inequality follows from Cauchy–Schwarz and the definition of $\epsilon^{(l)}$; the last inequality uses the definitions of δ_l and $a = \max_v \sum_u A_{vu}$. This argument does not require independence between the fresh and propagated errors. By the L_σ-Lipschitz property of σ and the induced matrix norm,

$$\begin{aligned}
\mathbb{E}\left\|\hat{h}_v^{(l+1)} - h_v^{(l+1)}\right\| &= \mathbb{E}\left\|\sigma\left(\hat{\mu}_v^{(l)} W^{(l)}\right) - \sigma\left(\mu_v^{(l)} W^{(l)}\right)\right\| \\
&\leq L_\sigma \, \mathbb{E}\left\|\left(\hat{\mu}_v^{(l)} - \mu_v^{(l)}\right) W^{(l)}\right\| \\
&\leq L_\sigma \|W^{(l)}\| \, \mathbb{E}\left\|\hat{\mu}_v^{(l)} - \mu_v^{(l)}\right\| \\
&\leq L_\sigma \|W^{(l)}\| \left(a\epsilon^{(l)} + \delta_l\right),
\end{aligned} \tag{60}$$

where the weights are fixed. Taking the maximum over v yields

$$\epsilon^{(l+1)} \leq L_\sigma \|W^{(l)}\| \left(a\epsilon^{(l)} + \delta_l\right). \tag{61}$$

Define $\gamma_l = L_\sigma \|W^{(l)}\| a$ and $b_l = L_\sigma \|W^{(l)}\| \delta_l$. Equation (61) becomes $\epsilon^{(l+1)} \leq \gamma_l \epsilon^{(l)} + b_l$. We prove by induction that

$$\epsilon^{(L)} \leq \left(\prod_{l=0}^{L-1} \gamma_l\right) \epsilon^{(0)} + \sum_{j=0}^{L-1} b_j \prod_{l=j+1}^{L-1} \gamma_l, \tag{62}$$

where an empty product equals one.

For $L = 1$, this is precisely $\epsilon^{(1)} \leq \gamma_0 \epsilon^{(0)} + b_0$. Suppose Equation (62) holds for L. Since $\gamma_L \geq 0$,

$$\begin{aligned}
\epsilon^{(L+1)} &\leq \gamma_L \epsilon^{(L)} + b_L \\
&\leq \left(\prod_{l=0}^{L} \gamma_l\right) \epsilon^{(0)} + \sum_{j=0}^{L-1} b_j \prod_{l=j+1}^{L} \gamma_l + b_L \\
&= \left(\prod_{l=0}^{L} \gamma_l\right) \epsilon^{(0)} + \sum_{j=0}^{L} b_j \prod_{l=j+1}^{L} \gamma_l.
\end{aligned} \tag{63}$$

This establishes the induction step and proves the stated multi-layer bound.

Finally, if $\gamma_l \le \gamma < 1$ and $b_l \le b$ for every layer, then

$$\begin{aligned}\epsilon^{(L)} &\le \gamma^L \epsilon^{(0)} + b \sum_{j=0}^{L-1} \gamma^{L-1-j} \\ &= \gamma^L \epsilon^{(0)} + b \, \frac{1-\gamma^L}{1-\gamma}. \end{aligned} \tag{64}$$

This completes the proof. □

C.7 Proof Of Theorem 2.7

Proof. **Part (i): Lower bound for unbiased estimators.** Since $\alpha(u) \equiv 1$ the estimator is unbiased, so $\mathrm{MSE} = \mathrm{Var}(\hat{h}_v)$. With k i.i.d. samples,

$$k \cdot \mathrm{Var}(\hat{h}_v) \;=\; \sum_{u \in \mathcal{N}(v)} \frac{A_{vu}^2 \, \|h_u\|^2}{p(u)} - \|h_v\|^2. \tag{65}$$

Since $\sum_u p(u) = 1$, the Cauchy–Schwarz inequality gives

$$\sum_u \frac{A_{vu}^2 \, \|h_u\|^2}{p(u)} \;=\; \sum_u p(u) \cdot \frac{c_u^2}{p(u)^2} \;\ge\; \Big(\sum_u c_u\Big)^2 = C^2, \tag{66}$$

which yields (20).

Part (ii): Upper bound for the surrogate. Substituting $\alpha^*(u) = kp(u)/(1 + kp(u))$ from Theorem 2.3 into the surrogate J:

$$\alpha^*(u) - 1 = -\frac{1}{1+kp(u)}, \qquad \big(\alpha^*(u)\big)^2 = \frac{k^2 p(u)^2}{(1+kp(u))^2}. \tag{67}$$

The per-neighbor contribution to J is

$$A_{vu}^2 \|h_u\|^2 \left[\frac{1}{(1+kp(u))^2} + \frac{1}{k} \cdot \frac{k^2 p(u)}{(1+kp(u))^2}\right] = \frac{A_{vu}^2 \|h_u\|^2}{(1+kp(u))^2} \cdot (1+kp(u)) = \frac{A_{vu}^2 \|h_u\|^2}{1+kp(u)}, \tag{68}$$

Taking the feasible proposal $p_0(u) = c_u/C$, rather than identifying it with a joint optimum, and using $1 + kc_u/C \ge kc_u/C$:

$$\sum_u \frac{A_{vu}^2 \, \|h_u\|^2}{1 + kc_u/C} \;\le\; \sum_u \frac{A_{vu}^2 \, \|h_u\|^2 \cdot C}{k \, c_u} = \frac{C}{k} \sum_u A_{vu} \|h_u\| = \frac{C^2}{k}, \tag{69}$$

establishing the second inequality in (21).

Part (iii): Achievable exact risk. The proposal $p_0(u) = c_u/C$ attains the unbiased lower bound $V_{\min} = (C^2 - \|h_v\|^2)/k$. Applying the scalar factor in Theorem 2.5 to that estimator gives risk $MV_{\min}/(M + V_{\min}) < V_{\min}$ when $V_{\min} > 0$. This feasible oracle bounds the infimum over all weights and proposals. Expanding the two squared norms gives the feature-direction identity. When G is diagonal, the exact risk equals $J - \|m_\alpha\|^2/k \le J$, which justifies the restricted transfer of the surrogate bound. No corresponding transfer holds for arbitrary correlated contributions.

□

C.8 Proof of Proposition 2.8

Proof. **Part 1: Unencoded equality.** Since $\hat{h}_v^{(1)}$ and $\hat{h}_v^{(2)}$ are independent and identically distributed,

$$\begin{aligned}\mathbb{E}\Big[\|\hat{h}_v^{(1)} - \hat{h}_v^{(2)}\|^2\Big] &= \mathbb{E}\Big[\|\hat{h}_v^{(1)}\|^2\Big] - 2\,\mathbb{E}\big[\hat{h}_v^{(1)}\big]^\top \mathbb{E}\big[\hat{h}_v^{(2)}\big] + \mathbb{E}\Big[\|\hat{h}_v^{(2)}\|^2\Big] \\ &= 2\Big(\mathbb{E}\Big[\|\hat{h}_v\|^2\Big] - \|\mathbb{E}[\hat{h}_v]\|^2\Big) \;=\; 2\,\mathrm{Var}(\hat{h}_v), \end{aligned} \tag{70}$$

where the cross term factorises by independence and the marginals coincide by identical distribution. This establishes Equation (26).

Part 2: Encoded inequality. Let $\Phi := g_\phi \circ f_\theta$ denote the composite encoder. Since Φ is L-Lipschitz,

$$\|z_v^{(1)} - z_v^{(2)}\|^2 \;=\; \|\Phi(\hat{h}_v^{(1)}) - \Phi(\hat{h}_v^{(2)})\|^2 \;\le\; L^2\,\|\hat{h}_v^{(1)} - \hat{h}_v^{(2)}\|^2. \tag{71}$$

Taking expectations and using Part 1,

$$\mathbb{E}\Big[\|z_v^{(1)} - z_v^{(2)}\|^2\Big] \;\le\; L^2\,\mathbb{E}\Big[\|\hat{h}_v^{(1)} - \hat{h}_v^{(2)}\|^2\Big] \;=\; 2L^2\,\mathrm{Var}(\hat{h}_v), \tag{72}$$

which establishes Equation (27). □

C.9 Proof of Proposition 2.9

Proof. Since both $z_v^{(1)}$ and $z_v^{(2)}$ lie on the unit sphere $\mathbb{S}^{d-1}$,

$$\|z_v^{(1)} - z_v^{(2)}\|^2 \;=\; 2 \;-\; 2\,\mathrm{sim}(z_v^{(1)}, z_v^{(2)}), \tag{73}$$

so the alignment loss can be written as

$$\mathbb{E}\Big[1 - \mathrm{sim}(z_v^{(1)}, z_v^{(2)})\Big] \;=\; \tfrac{1}{2}\,\mathbb{E}\Big[\|z_v^{(1)} - z_v^{(2)}\|^2\Big]. \tag{74}$$

By the encoded inequality of Proposition 2.8 (Equation (27)),

$$\mathbb{E}\Big[1 - \mathrm{sim}(z_v^{(1)}, z_v^{(2)})\Big] \;\le\; \tfrac{1}{2}\cdot 2L^2\,\mathrm{Var}_p(\hat{h}_v) \;=\; L^2\,\mathrm{Var}_p(\hat{h}_v), \tag{75}$$

which establishes the upper bound (28). Holding the features, weights, encoder (including normalization), and a common Lipschitz bound fixed, the right-hand side of (75) depends on the sampling distribution p only through $\mathrm{Var}_p(\hat{h}_v)$. By Theorem 2.4, p^* minimizes this variance over all distributions p, and therefore p^* minimizes the variance-induced upper bound on the alignment loss. □

C.10 Why the Diagonal Surrogate Is Not a General Upper Bound

Let $R(\alpha; p) = \mathrm{MSE}(\hat{h}_v)$ and $c_u = \|x_u\|$. Let $U(\alpha; p) = \|\sum_u (\alpha_u - 1)x_u\|^2 + \sum_u c_u^2\alpha_u^2/(kp_u)$. Then

$$U(\alpha; p) - R(\alpha; p) = \frac{1}{k}\left\|\sum_u \alpha_u x_u\right\|^2 \ge 0.$$

Thus retaining the full bias and discarding only the negative variance term gives an upper bound. The diagonal objective J does not generally share this property. Take $x_1 = x_2 = 1$, $p_1 = p_2 = 1/2$, $k = 1$, and $\alpha_1 = \alpha_2 = 1/3$. The target is 2 and every sampled estimate is $2/3$, so

$$R = (2/3 - 2)^2 = 16/9, \qquad J = 2(1/3 - 1)^2 + 2(1/3)^2/(1/2) = 4/3 < R.$$

This example also has a zero-variance unbiased estimator. It demonstrates why neither a universal majorization claim nor strict dominance of the practical correction is appropriate.

C.11 Joint Optimization of the Diagonal Surrogate

This subsection concerns J, not the exact correlated-feature risk R. Allow $p_u = 0$ together with $\alpha_u = 0$ using the limiting perspective convention: the unsampled contribution is omitted, its variance contribution is zero, and its diagonal squared-bias contribution is c_u^2. Optimizing J first over α reduces the problem to

$$\min_{p\in\Delta} F(p) = \sum_u \frac{c_u^2}{1 + kp_u}, \qquad \Delta = \{p : p_u \ge 0,\ \sum_u p_u = 1\}. \tag{76}$$

For $c_u > 0$, F is strictly convex. The optimality conditions yield

$$p_u^\dagger = \frac{1}{k}\left(\frac{c_u}{t} - 1\right)_+, \qquad \sum_u \left(\frac{c_u}{t} - 1\right)_+ = k, \qquad \alpha_u^\dagger = \frac{kp_u^\dagger}{1 + kp_u^\dagger}. \tag{77}$$

There is a unique $t > 0$ solving the scalar normalization equation: its left side is continuous and strictly decreasing wherever positive, diverges as $t \downarrow 0$, and vanishes for $t \geq \max_u c_u$. To verify the optimality conditions, use $F'_u(p_u) = -kc_u^2/(1+kp_u)^2$. For an active coordinate the common multiplier is kt^2; for an inactive coordinate $c_u \leq t$, satisfying the complementary inequality. Substituting the minimizing weights shows that this is the joint minimum of the surrogate. Boundary solutions mean that an interior fixed point need not exist under a strict $p_u > 0$ constraint.

For a fixed finite active set with $c_u > 0$, all coordinates eventually become active as k grows. With n coordinates and $C = \sum_u c_u$, $t = C/(k+n)$ and

$$p_u^\dagger = \frac{(k+n)c_u}{kC} - \frac{1}{k}, \qquad p_u^\dagger - \frac{c_u}{C} = \frac{nc_u/C - 1}{k}. \tag{78}$$

Thus the surrogate-optimal distribution approaches the norm-proportional proposal at rate $O(1/k)$ for this fixed set. This explains a possible limited gain from proposal iteration; it does not prove identical classification performance or convergence of an arbitrary alternating implementation. It also does not yield a $c_u^{2/3}$ rule. The experiment labeled "joint fixed point" must be tied to its implemented update and stopping rule before being identified with this solution.

D Additional Experimental results

Table 5: Dataset Statistics

Dataset	Nodes	Edges	Classes	Features	Train/Val/Test (Rich label)	Train/Val/Test (Sparse label)
Cora	2,708	5,429	7	1,433	1,208 / 500 / 1,000	-
Pubmed	19,717	44,338	3	500	18,217 / 500 / 1,000	910 / 500 / 18,307
Reddit	232,965	11,606,919	41	602	152,410 / 23,831 / 55,334	7,671 / 23,831 / 201,463
OGBN-arXiv	169,343	1,166,243	40	128	90,941 / 29,799 / 48,603	4,547 / 14,899 / 149,897
PPI	56,944	2,818,716	14	50	-	1,122 / 6,514 / 49,308

D.1 Dataset Details

Cora is a citation network where nodes represent research papers and edges denote citation links, with the task of classifying papers into research topics. PubMed is a large-scale biomedical citation graph, where each node corresponds to a scientific publication and the goal is to categorize papers into disease-related classes. Reddit is a social interaction graph constructed from user comments, where nodes correspond to posts and edges capture interactions; the task is to predict the community (subreddit) to which each post belongs. OGBN-arXiv is a directed citation network of computer science papers, where nodes represent arXiv papers and edges indicate citations; each node is associated with a 128-dimensional text-derived feature vector, and the task is to classify papers into 40 subject areas, reflecting their primary research domains. The statistics of four datasets are summarized in Table 5.

D.2 Implementation Details

To reduce the computational cost, we cache the layer-0 feature norm for every node before training and gradually update the cache for the sampled nodes at each epoch. The pseudo-code is shown in Algorithm 1. The experiments are done in a Linux system with a RTX 4090 GPU and I9-13900F CPU. The hyperparameters are summarized in Table 6.

Table 6: Hyperparameter configurations per dataset.

Hyperparameter	**Cora**	**PubMed**	**Reddit**	**OGBN-arXiv**	PPI-large
`Hidden feature dimension`	16	16	128	256	128
`Initial batch size`	256	256	1024	1024	1024
`Size of neighbor nodes`	400	400	5120	10240	10240
`Coefficient of contrastive loss`	0.5	0.5	0.5	0.5	0.5
`Epochs`	500	500	1000	2000	2000
`Number of hidden layers`	1	1	1	2	2

Algorithm 1 BV-SAMPLER: Bias–Variance Optimal Sampling (one epoch)

Require: Graph $G = (\mathcal{V}, \mathcal{E})$, normalized adjacency A, node features $\{h_u\}$, labeled set $\mathcal{V}_L$, unlabeled set $\mathcal{V}_U$, sampling budget k, labeled ratio p, contrastive weight λ_c, learning rate η

Ensure: Encoder parameters θ, projection head parameters ϕ

for each node $u \in \mathcal{V}$ in batches **do**
  $\mathcal{C}[u] \leftarrow \|f_\theta^{(0)}(h_u)\|$ ▷ Cache Layer-0 feature norm for every node
end for
for each mini-batch **do**
  Sample $\mathcal{V}_L^{\text{batch}}$: $\lfloor p \cdot |\mathcal{B}| \rfloor$ nodes from $\mathcal{V}_L$
  Sample $\mathcal{V}_U^{\text{batch}}$: remaining $\lceil (1-p) \cdot |\mathcal{B}| \rceil$ nodes from $\mathcal{V}_U$
  Run biased forward pass on $\mathcal{V}_L^{\text{batch}} \cup \mathcal{V}_U^{\text{batch}}$ using p^* and α^*
  Compute supervised loss on labeled nodes: $\mathcal{L}_{\text{CE}} \leftarrow \text{CrossEntropy}(\hat{y}_v,\, y_v)_{v \in \mathcal{V}_L^{\text{batch}}}$
  Draw two views for unlabeled nodes: $z_v^{(1)},\, z_v^{(2)}$ for $v \in \mathcal{V}_U^{\text{batch}}$
  Compute joint loss:

$$\mathcal{L}_{\text{total}} \;\leftarrow\; \mathcal{L}_{\text{CE}}(\mathcal{V}_L^{\text{batch}}) \;+\; \lambda_c\, \mathcal{L}_{\text{InfoNCE}}(\mathcal{V}_U^{\text{batch}})$$

  $\theta, \phi \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}, \quad \phi - \eta \nabla_\phi \mathcal{L}_{\text{total}}$ ▷ SGD step
  Update cache: $\mathcal{C}[u] \leftarrow \|h_u^{(l)}\|$ for all visited nodes u
end for

D.3 ADDITIONAL EXPERIMENTAL RESULTS

A central motivation of our work is the observation that existing samplers fail under label scarcity. As shown in Figure 1, Micro-F1 scores of FastGCN, GraphSAINT, and LADIES on OGBN-arXiv collapse after approximately 100 training iterations at the sparse-label setting, revealing an instability that is absent under dense supervision. To stress-test this failure mode, we follow a batch-only inference protocol: at evaluation time, each model is restricted to the same sampled mini-batch rather than the full graph. This protocol targets the deployment-relevant regime where computing full neighborhoods for all test nodes is prohibitive. Table 7 reports Micro-F1 under this protocol at the sparse-label setting. Two findings are noteworthy: (1) all unbiased samplers degrade substantially, with GCN (Batch) collapsing to 39.31 on OGBN-arXiv and even the strongest baseline, GraphSAINT, reaching only 56.92; (2) BV-SAMPLER achieves 66.57% on OGBN-arXiv and 92.79% on Reddit, surpassing the respective strongest baselines by 9.65% and 2.50%. We attribute the superior performance of BV-SAMPLERto two coupled mechanisms: (i) under unbiased sampling, the per-step gradient variance is inflated (Theorem 2.7), so the supervised signal from a small labeled set is no longer sufficient to keep the encoder in a good basin; and (ii) unlike traditional graph contrastive learning methods that are decoupled from sampling, BV-SAMPLER couples variance control with a sampler-aware contrastive objective that explicitly regularizes the encoder against sampling noise (Proposition 2.8). Comparing Table 3 with Table 2 reveals that under full-graph evaluation BV-SAMPLER is competitive with all baselines, but under batch-only evaluation the margin widens dramatically.

Table 7: Stability evidence under batch-only inference on Reddit and OGBN-arXiv in the sparse-label setting. The best score is marked in bold.

Methods	OGBN-arXiv	Reddit
GCN (Batch)	39.31±0.47	85.44±0.29
GraphSAGE	54.42±0.34	90.26±0.28
FastGCN	53.50±0.17	90.29±0.25
GraphSAINT	56.92±0.11	87.07±0.75
LADIES	48.69±0.53	89.91±0.39
BV-SAMPLER	**66.57 ± 0.23**	**92.79 ± 0.30**

Table 8: Correction transfer on PubMed, three paired seeds. Micro-F1 is converted from the reported fractional scores to percent. Epochs refer to a training-loss target, not a test-accuracy target.

Condition	Micro-F1 (%)	Epochs to loss target	p vs. (A)
(A) MVS-GNN, $\alpha \equiv 1$	82.10±0.58	31.3±8.7	–
(B) Same sampler, one correction	83.43±0.67	27.7±8.3	0.002
(C) Joint fixed-point iteration	83.28±0.11	30.7±6.8	0.088

Table 9: Frozen-model Cora diagnostics. Bias^2, variance, and MSE are normalized by reference-representation energy. Values are preserved from the supplied experimental report.

Depth	Evaluation weight	Bias^2	Variance	MSE
1	$\alpha \equiv 1$	0.007±0.001	1.510±0.056	1.525±0.057
1	Shrinkage	0.359±0.002	0.014±0.000	0.374±0.002
2	$\alpha \equiv 1$	0.008±0.001	1.965±0.092	1.984±0.092
2	Shrinkage	0.569±0.003	0.013±0.000	0.582±0.003
3	$\alpha \equiv 1$	0.011±0.002	2.874±0.278	2.902±0.282
3	Shrinkage	0.705±0.004	0.012±0.000	0.718±0.003
4	$\alpha \equiv 1$	0.021±0.008	3.960±0.457	4.003±0.457
4	Shrinkage	0.760±0.006	0.016±0.001	0.776±0.005

D.4 Contribution Weighting and Joint Updates

To isolate contribution weighting from proposal design, we hold the MVS-GNN sampler fixed and apply the proposed correction on PubMed. Table 8 shows that the correction increases Micro-F1 from 82.10% to 83.43%, with a reported paired-test p-value of 0.002. Joint fixed-point iteration achieves 83.28% and does not provide a detectable improvement over the one-step correction ($p = 0.767$). Thus, an inexpensive weighting step captures the observed benefit in this experiment; three seeds are insufficient to establish equivalence between the two variants.

The epochs-to-target comparison uses a training-loss threshold equal to 1.05 times the unweighted baseline's final loss for each seed. The one-step correction reaches this target in 27.7 epochs on average, compared with 31.3 for the baseline, although variation across seeds is substantial. This controlled study uses a separate run collection from the full-graph benchmark table and should not be interpreted as a common-accuracy wall-clock comparison.

D.5 Direct Bias–Variance Analysis

To examine the estimation trade-off directly, we freeze converged Cora models at depths one through four and compare sampled representations with full-graph references over 100 Monte Carlo draws. Table 9 reports normalized squared bias, variance, and MSE. With $\alpha \equiv 1$, variance increases from 1.510 at depth one to 3.960 at depth four. Biased weighting increases squared bias but reduces variance to between 0.012 and 0.016, yielding reported MSE reductions of approximately 71–81% across the four depths.

Table 10: Reported Micro-F1 (%) versus contrastive coefficient. The arXiv scores correspond to the low-label batch-only experiment; PubMed and Reddit use the reported 5% settings.

λ_c	OGBN-arXiv	PubMed	Reddit
0	65.85±0.13	84.69±0.28	93.36±0.05
0.1	66.28±0.25	84.81±0.25	93.34±0.06
0.3	66.39±0.27	85.14±0.35	93.26±0.03
0.5	66.57±0.23	84.97±0.21	93.36±0.07
1	66.52±0.20	84.95±0.18	93.32±0.06

Table 11: Micro-F1 (%) and time (s) versus sampling budget. PubMed and Reddit use the reported 5% label settings. Timing scope must be confirmed before comparison with per-epoch results.

k	Cora		PubMed		Reddit	
	Micro-F1	Time	Micro-F1	Time	Micro-F1	Time
10	87.40±0.18	0.0033	84.57±0.49	0.0033	92.82±0.06	0.077
20	87.23±0.41	0.0035	84.67±0.14	0.0035	93.21±0.16	0.106
50	87.55±0.38	0.0037	84.94±0.21	0.0036	93.31±0.12	0.130
100	87.70±0.61	0.0037	85.04±0.50	0.0038	93.17±0.04	0.130
200	87.80±0.41	0.0038	85.11±0.25	0.0039	93.25±0.14	0.131

D.6 Sampling Budget and sensitivity analysis

In Table 11, we report the performance and runtime with respect to different budget. As the budget increases from 10 to 200, PubMed Micro-F1 increases from 84.57% to 85.11%, while Cora and Reddit exhibit non-monotonic changes. On OGBN-arXiv with batch-only inference, increasing λ_c from 0 to 0.5 improves Micro-F1 from 65.85% to 66.57%; the Reddit sweep shows no consistent gain. These results indicate that additional samples and auxiliary supervision have dataset-dependent benefits.

In Table 10, we analyze the sensitivity of λ_c. By observation, increasing λ_c from 0 to 0.5 improves Micro-F1 on OGBN-arXiv from 65.85% to 66.57%, with little change at $\lambda_c = 1$. PubMed achieves its highest mean score at $\lambda_c = 0.3$, improving by 0.45 percentage points over $\lambda_c = 0$. Reddit remains within a narrow range of 93.26–93.36%, showing no consistent benefit. These results indicate that moderate contrastive weighting can improve prediction, while its benefit depends on the dataset.

E Limitations

We highlight several limitations of the present work to guide future research. (i) The closed-form α^* relies on a diagonal approximation of the squared bias. Theorem 2.3 drops cross-neighbor inner products $\langle h_u, h_{u'} \rangle$ from the bias term; the non-diagonal optimum is given by Theorem 2.2 but costs $O(|\mathcal{N}(v)|^3)$ per node and is therefore impractical for high-degree nodes. (ii) The feature-norm cache introduces staleness. Our practical algorithm caches $\|h_u\|$ from earlier mini-batches and updates lazily, which can produce slightly out-of-date sampling probabilities for nodes that are visited infrequently. (iii) The contrastive component links variance to an upper bound on the alignment loss, not to the alignment loss itself.