KNOWLEDGE DISTILLATION FOR IMPROVED GENERALIZATION IN TWO-STAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale classification often uses a two-stage architecture: a lightweight candidate generator selects a small subset of classes, and a stronger target classifier scores them. Recent work has improved these systems through cross-stage training, using techniques such as knowledge distillation and joint optimization. Despite these empirical advances, the effect of the candidate-generator distribution on target-classifier training remains poorly understood. To explain this mechanism, we formulate sampled softmax training as importance sampling SGD, using the candidate generator distribution as the proposal. We also develop a variance reduction analysis of cross-stage distillation. In the context of cross-stage information asymmetry, the output distribution of the target classifier is minimax-optimal for the unbiased estimator. For the practical self-normalized estimator, alignment controls the variance and bias bounds. These results demonstrate how distillation enhances training efficiency and convergence. In addition to providing training samples, the candidate generator determines which classes are available during inference. To analyze this role, we derive conditions for teacher quality and candidate generator capacity under which positive weight distillation strictly tightens the coverage loss upper bound relative to independent training. Combining these results, we obtain an end-to-end generalization bound that considers both coverage loss and target-classifier error. Experiments on real-world datasets demonstrate that distillation yields the lowest gradient variance among non-Oracle proposals. It also improves final accuracy across three model combinations. Convergence experiments on two combinations further demonstrate faster training.

1 INTRODUCTION

Large-scale classification arises in extreme multi-label classification Prabhu et al. (2018); You et al. (2019), recommender systems Covington et al. (2016); Koren et al. (2009), dense retrieval Karpukhin et al. (2020); Nogueira et al. (2019), and language modeling Jean et al. (2015); Bengio & Sénécal (2008). Computing the full softmax cross-entropy gradient requires summing over all classes. Therefore, its cost grows linearly with the number of classes and becomes substantial when the number of classes reaches the millions or billions. At this scale, full softmax is limited by memory and computing power. It is also largely unnecessary since most classes are unrelated to a given input. The two-stage candidate-ranking architecture (first select candidates, then rank them) has therefore become the mainstream solution. A lightweight candidate generator selects a small set of candidates, which a stronger target classifier then scores. Target-classifier training uses the sampled softmax loss Jean et al. (2015); Bengio & Sénécal (2008); Gutmann & Hyvärinen (2010), which approximates the full gradient over the candidate set. However, its effectiveness depends directly on the distribution of the candidate generator. Therefore, determining the optimal candidate-generator distribution for training the target classifier remains an open design problem.

Recent work has sought to improve target-classifier training by coordinating across stages using knowledge distillation Hinton et al. (2015), adaptive negative sampling Lian et al. (2020); Jiang et al. (2021), and joint training Gallagher et al. (2019); Qin et al. (2022); Wang et al. (2025). These approaches have demonstrated empirical benefits. The existing theory Zhang et al. (2024); Hron et al. (2021) primarily addresses generalization error. However, the effect of the candidate-generator distribution on the target classifier's gradient updates remains unclear.

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.

To explain this training mechanism, we consider the sampled softmax training of the target classifier to be a stochastic gradient estimation. In this estimation, the candidate generator-related distribution is the proposal, and the full softmax gradient is the target. Unlike in classical importance sampling Owen (2013); Rubinstein & Kroese (2016), where the proposal is chosen freely, here the candidate generator is both the object being sampled and an independently trained model that has no access to the target classifier's per-class gradients. This information asymmetry narrows the feasible proposals. We revisit the relationship through two complementary aspects: training variance and retrieval coverage. This results in the first theory of variance reduction in cross-stage co-training for large-scale classification. Distribution alignment controls the variance of target-classifier training and, under explicit conditions, tightens an upper bound on the coverage loss of the candidate generator. The end-to-end generalization bound incorporates both the training and coverage contributions. This principle connects knowledge distillation, adaptive negative sampling, and joint training. Our main contributions are:

- We derive an end-to-end generalization bound that separates the irreducible Bayes error from the training and retrieval errors that are controlled by distribution alignment. We also identify the conditions under which distillation strictly tightens the coverage-loss upper bound relative to independent training, including the teacher quality and candidate generator capacity.
- We formulate sampled softmax training as an importance sampling stochastic optimization problem. We characterize the minimax optimal proposal under information asymmetry between the two stages and demonstrate how alignment controls the variance and bias of the practical self-normalized estimator, accelerating convergence.
- Experiments on real-world datasets and three model combinations demonstrate that the target-classifier distribution results in the lowest gradient variance. Furthermore, lower-variance proposals enhance convergence and final accuracy, while distilled candidate generators consistently outperform those trained independently.

2 Related Work

This work analyzes how distillation-induced distribution alignment between the candidate generator and the target classifier reduces gradient-variance and improves candidate-set coverage in a two-stage architecture. We review three bodies of related work: two-stage classification architectures and cross-stage co-training, knowledge distillation theory, and importance-sampling variance reduction.

2.1 Multi-stage Classification Architectures

The standard solution for large-scale classification is the two-stage candidate-ranking pipeline, which is a lightweight model that selects candidates and a stronger model that re-ranks them. In recommender systems, matrix factorization Koren et al. (2009) and deep candidate generators Covington et al. (2016) supply ranking models Guo et al. (2017); Zhou et al. (2018). The industry has expanded upon this concept by designing multi-stage cascades under algorithm-system co-optimization constraints Wang et al. (2020). In extreme multi-label classification, label-tree selectors Prabhu & Varma (2014) and Transformer-based pipelines Chang et al. (2020) come before the ranking step. Later work couples these two stages through dynamic negative sampling Jiang et al. (2021); Kharbanda et al. (2022). In information retrieval, dense retrievers Karpukhin et al. (2020) act as learned candidate generators in retrieve-then-rerank pipelines Hofstätter et al. (2020); Nogueira et al. (2019).

Cross-stage co-training is a recurring theme across all these settings. From a theoretical perspective, Gallagher et al. (2019) analyze the joint optimization of cascaded rankers in terms of the efficiency-effectiveness trade-off. Hron et al. (2021) prove that independent training is suboptimal under certain conditions. Finally, Zhang et al. (2024) establish a Rademacher complexity bound for tree-structured two-stage systems. Empirically, CoRR Huang et al. (2023) uses cross-stage distillation, while RankFlow Qin et al. (2022) and Wang et al. (2025) jointly optimize multi-stage cascades. From the perspective of the training process, none of the above characterizes how the candidate-generator distribution affects the per-step gradient variance of the target classifier.

2.2 KNOWLEDGE DISTILLATION

Hinton et al. (2015) proposed distillation as supervising a student with a teacher's soft output distribution. This concept has been expanded to a two-stage approach that uses a stronger target classifier to guide candidate generators Tang & Wang (2018); Kang et al. (2020); Chen et al. (2023). Theoretical accounts include a generalization bound for linear classifiers Phuong & Lampert (2019), a multi-view feature-transfer interpretation Allen-Zhu & Li (2023), and a characterization of distillation as partial SVRG where soft labels act as a control variate Safaryan et al. (2023). This work complements the above by analyzing distillation from the perspective of both sampling-distribution alignment and candidate-set coverage loss.

2.3 IMPORTANCE SAMPLING AND VARIANCE REDUCTION

As the class space increases to the millions, the standard approximations to the full softmax are sampled softmax Jean et al. (2015); Bengio & Sénécal (2008), noise-contrastive estimation Gutmann & Hyvärinen (2010), and adaptive softmax Grave et al. (2017). In the field of stochastic optimization, Zhao & Zhang (2015) prove that sampling proportional to per-example smoothness constants minimizes gradient-estimator variance. Additionally, Katharopoulos & Fleuret (2018) derive a practical, gradient-norm-based strategy that has been empirically verified to speed up convergence. The complementary control-variate line represented by SVRG Johnson & Zhang (2013) reduces variance by periodically anchoring stochastic gradients to a full-gradient snapshot. The two approaches reduce variance by changing the proposal distribution and correcting the estimator, respectively. The above studies all focus on reducing variance within a single model. This work extends the importance-sampling analysis to a two-stage setting in which the distribution of the candidate generator determines the proposal for the target classifier's gradient estimator.

3 PRELIMINARIES

3.1 PROBLEM SETUP AND NOTATION

We consider a two-stage candidate-ranking architecture trained using the softmax cross-entropy objective, which is approximated using sampled softmax. We require that the output of the candidate generator admit a model as probabilistic sampling from a parametric distribution. Below we introduce the notation and training objectives of the two components in turn. The full class set is a finite set $\mathcal{I}$ with $|\mathcal{I}| = N$. For an input c, the target classifier (with parameters ϕ) assigns to each class $j \in \mathcal{I}$ a score $s_\phi(j, c) \in \mathbb{R}$, which induces the softmax distribution

$$P_\phi(j|c) = \frac{\exp(s_\phi(j,c))}{\sum_{k\in\mathcal{I}} \exp(s_\phi(k,c))}, \quad j \in \mathcal{I}, \tag{1}$$

with partition function denoted by $Z_\phi(c) = \sum_{k\in\mathcal{I}} \exp(s_\phi(k,c))$. The parallel candidate generator (with parameters θ) defines a distribution $Q_\theta(\cdot|c)$, whose parametric form is given in Section 3.2.

Given a positive example i^+, the training objective of the target classifier is the full-softmax cross-entropy loss:

$$\mathcal{L}(\phi) = -s_\phi(i^+, c) + \log Z_\phi(c). \tag{2}$$

Its gradient is

$$\nabla_\phi \mathcal{L}(\phi) = -\nabla_\phi s_\phi(i^+, c) + \sum_{j\in\mathcal{I}} P_\phi(j|c)\, \nabla_\phi s_\phi(j, c). \tag{3}$$

The second term is the full-class-set expectation under P_ϕ; for N in the millions it is too costly to compute exactly and must be approximated by sampling, whose quality hinges on the candidate generator's distribution Q_θ, formalized in the next section.

3.2 IMPORTANCE SAMPLING FOR TWO-STAGE TRAINING

Sampled softmax approximates the full-softmax loss by importance sampling Jean et al. (2015); Bengio & Sénécal (2008). In a two-stage system, the trainable Q_θ is both candidate generator and

proposal; with score $s_\theta(j, c)$, its fully supported form is

$$Q_\theta(j|c) = \frac{\exp(s_\theta(j,c))}{\sum_{k\in\mathcal{I}} \exp(s_\theta(k,c))}, \quad j \in \mathcal{I}. \tag{4}$$

Drawing K i.i.d. samples from Q_θ gives the unbiased estimator

$$\hat{g}(\phi) = -\nabla_\phi s_\phi(i^+, c) + \frac{1}{K}\sum_{k=1}^{K} \frac{P_\phi(j_k|c)}{Q_\theta(j_k|c)} \nabla_\phi s_\phi(j_k, c), \quad j_k \overset{\text{i.i.d.}}{\sim} Q_\theta(\cdot|c), \tag{5}$$

with $\mathbb{E}_{Q_\theta}[\hat{g}(\phi)] = \nabla_\phi \mathcal{L}(\phi; c, i^+)$. Because P_ϕ requires the full partition function Z_ϕ, sampled softmax Bengio & Sénécal (2008) instead self-normalizes the weights over $\mathcal{S} = \{i^+, j_1, \ldots, j_K\}$ and defines $\hat{\mathcal{L}}(\phi) = -\log \hat{P}_\mathcal{S}(i^+)$ with

$$\hat{P}_\mathcal{S}(j) = \frac{\exp(s_\phi(j,c)) \,/\, Q_\theta(j|c)}{\sum_{l\in\mathcal{S}} \exp(s_\phi(l,c)) \,/\, Q_\theta(l|c)}, \quad j \in \mathcal{S}. \tag{6}$$

The partition function cancels; differentiating with Q_θ independent of ϕ gives

$$\hat{g}^{\text{SN}} := \nabla_\phi \hat{\mathcal{L}}(\phi) = -\nabla_\phi s_\phi(i^+, c) + \sum_{j\in\mathcal{S}} \hat{P}_\mathcal{S}(j)\, \nabla_\phi s_\phi(j, c), \tag{7}$$

This SNIS estimator replaces the constant K in Eq. (5) by the in-set weight sum and has $O(1/K)$ finite-sample bias Owen (2013).

3.3 Regularity Assumptions

Our convergence analysis relies on the following standard regularity of the scoring function.

Assumption 1 (Regularity of the scoring function)**.**

(a) **Bounded scores:** there exists a constant $B > 0$ such that $|s_\phi(j, c)| \le B$ for all $j \in \mathcal{I}$ and all ϕ.

(b) **Lipschitz-continuous gradient:** there exists a constant $\beta > 0$ such that $\|\nabla_\phi s_\phi(j, c)\big|_{\phi_1} - \nabla_\phi s_\phi(j, c)\big|_{\phi_2}\| \le \beta \,\|\phi_1 - \phi_2\|$ for all $j \in \mathcal{I}$ and all ϕ_1, ϕ_2.

(c) **Bounded gradient:** there exists a constant $G > 0$ such that $\|\nabla_\phi s_\phi(j, c)\| \le G$ for all $j \in \mathcal{I}$ and all ϕ.

These conditions hold for typical neural scorers with bounded parameters and smooth activations (e.g., sigmoid, tanh, or GELU). Under them, the global loss $\mathcal{L}(\phi) = -s_\phi(i^+, c) + \log Z_\phi(c)$ (Eq. (2)) is L-smooth with $L = 2\beta + G^2$ (Lemma B.1, proof in Appendix B).

4 Variance Analysis of the Gradient Estimator

We analyze how the candidate generator's proposal distribution Q_θ governs the variance of the target classifier's gradient estimator, and which Q_θ minimizes it: first for the strictly unbiased estimator $\hat{g}$ (Eq. (5)), whose conditional variance admits an exact closed form, and then for the self-normalized estimator $\hat{g}^{\text{SN}}$ used in practice.

4.1 Variance and Convergence

From the L-smoothness of $\mathcal{L}$ (Lemma B.1), standard nonconvex SGD analysis bounds the target classifier's convergence by its gradient-estimator variance.

Theorem 4.1 (Convergence rate)**.** *Under Assumption 1 ($L = 2\beta + G^2$), run the target classifier for T SGD steps with learning rate $\eta \in (0, 2/L)$ and gradient estimator $\hat{g}_t$ on the population loss $\bar{\mathcal{L}}(\phi) = \mathbb{E}_{(c,i^+)\sim\mathcal{D}}[\mathcal{L}(\phi; c, i^+)]$, and let $\sigma^2 = \sup_t \operatorname{Tr}(\mathbb{V}[\hat{g}_t \mid \phi_t])$ be the total estimator variance. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla\bar{\mathcal{L}}(\phi_t)\|^2 \le \underbrace{\frac{\bar{\mathcal{L}}(\phi_0) - \bar{\mathcal{L}}^*}{\eta(1-\eta L/2)\,T}}_{\to 0 \text{ as } T\to\infty} + \underbrace{\frac{\eta L}{2(1-\eta L/2)}\,\sigma^2}_{\text{variance residual}}, \tag{8}$$

where $\bar{\mathcal{L}}^* = \inf_\phi \bar{\mathcal{L}}(\phi)$; the standard step size $\eta^* = \min\big(1/L,\ \sqrt{2(\bar{\mathcal{L}}(\phi_0) - \bar{\mathcal{L}}^*)/(L\sigma^2 T)}\big)$ gives the rate $O(\sigma/\sqrt{T})$ (proof in Appendix B).

The total variance σ^2 decomposes exactly, for any K, into a sampling part $V_{\mathrm{IS}}(\phi_t, Q_\theta) = \mathbb{E}_{Y\sim\mathcal{D}}[\mathrm{Tr}(\mathbb{V}_{Q_\theta}[\hat{g}_t \mid Y])]$ that depends on Q_θ, and a data part $V_{\mathrm{data}}(\phi_t) = \mathrm{Tr}(\mathbb{V}_{Y\sim\mathcal{D}}[\nabla\mathcal{L}(\phi_t; c, i^+)])$ that does not (Proposition B.2, Appendix B); only V_{IS} is controllable through the proposal distribution.

4.2 The Optimal Proposal Distribution

The residual satisfies $\sigma^2 \le \sigma^2_{Q_\theta} + \sigma^2_{\mathrm{data}}$, where $\sigma^2_{Q_\theta} = \sup_t V_{\mathrm{IS}}(\phi_t, Q_\theta)$ is the only proposal-controllable term and $\sigma^2_{\mathrm{data}} = \sup_t V_{\mathrm{data}}(\phi_t)$ is irreducible mini-batch-SGD variance Ghadimi & Lan (2013). Conditional-variance minimization separates over (c, i^+); after dropping the Q_θ-independent term in the exact variance identity (Appendix B), it reduces to

$$\min_{Q_\theta\in\Delta_N} M(Q_\theta) := \sum_{j\in\mathcal{I}} \frac{P_\phi(j|c)^2}{Q_\theta(j|c)} \|\nabla_\phi s_\phi(j,c)\|^2, \tag{9}$$

where Δ_N is the probability simplex, and this minimization admits a closed form (proof in Appendix B).

Theorem 4.2 (Optimal proposal distribution)**.** *Optimizing over the interior simplex $\Delta_N^\circ = \{Q \in \Delta_N : Q_j > 0\ \forall j\}$, if $\|\nabla_\phi s_\phi(j,c)\| > 0$ for all $j \in \mathcal{I}$, then the unique minimizer of $M(Q_\theta)$ is*

$$Q^*_\theta(j|c) = \frac{P_\phi(j|c)\,\|\nabla_\phi s_\phi(j,c)\|}{\sum_{i\in\mathcal{I}} P_\phi(i|c)\,\|\nabla_\phi s_\phi(i,c)\|}, \tag{10}$$

with minimum value $M(Q^*_\theta) = \big(\sum_{j\in\mathcal{I}} P_\phi(j|c)\,\|\nabla_\phi s_\phi(j,c)\|\big)^2$.

Implementing this optimum requires the target classifier's per-class gradient norms, which are unavailable to the candidate generator. This information constraint motivates the minimax formulation below.

4.2.1 Minimax Optimality under Information Asymmetry

To address this, this section adopts a minimax criterion: treating the unknown gradient norm as an adversarial perturbation, we minimize the variance in the worst case. Let $g_j = \|\nabla_\phi s_\phi(j,c)\| \ge 0$ be the unknown gradient norm. By Assumption 1(c), $g_j \le G$ holds for each j, and hence $\sum_{j\in\mathcal{I}} P_\phi(j|c)\, g_j^2 \le G^2$. In what follows we adopt the latter weighted second-moment constraint as the adversary's uncertainty set. This constraint is a (strictly looser) relaxation of the pointwise bound $g_j \le G$. Writing $C = G^2$, the minimax problem is:

$$\min_{Q_\theta\in\Delta_N^\circ}\ \max_{g_j\ge 0,\ \sum_{j\in\mathcal{I}} P_\phi(j)\, g_j^2 \le C}\ \sum_{j\in\mathcal{I}} \frac{P_\phi(j)^2}{Q_\theta(j)}\, g_j^2, \tag{11}$$

where Δ_N° restricts Q_θ to the interior simplex so as to keep the importance weights well defined. This minimax problem admits a closed-form solution (proof in Appendix B):

Theorem 4.3 (Minimax-optimal proposal distribution)**.** *The solution of the minimax problem* (11) *is $Q^{\mathrm{mm}}_\theta = P_\phi$, with minimax value C. The same solution ($Q_\theta = P_\phi$, with value C) also holds under the stricter pointwise uncertainty set $\{g_j \le G\}$.*

Thus, in the worst case where the gradient norm is unknown, $Q_\theta = P_\phi$ gives the smallest variance bound. Moreover, if the gradient norm is approximately uniform across classes ($\|\nabla_\phi s_\phi(j,c)\| \approx g$ for all j), the optimal proposal (10) itself simplifies to $Q^*_\theta(j|c) \approx P_\phi(j|c)$, agreeing with the minimax solution.

4.3 Variance and Bias of the Self-Normalized Estimator

Since the unbiased estimator $\hat{g}$ depends on the intractable partition function Z_ϕ, practice instead uses the self-normalized estimator $\hat{g}^{\mathrm{SN}}$ (Eq. (7)), which lets Z_ϕ cancel during normalization at the cost of exact unbiasedness; all conclusions of the exact case carry over to this practical estimator. An analogous convergence guarantee holds for $\hat{g}^{\mathrm{SN}}$: the same $O(1/T)$ optimization decay and σ^2 variance

residual up to constants, plus an additional $O(b^2)$ bias term, where b bounds the full-expectation bias of $\hat{g}^{\mathrm{SN}}$ (Theorem B.3, Appendix B). The variance admits a data/sampling decomposition with a bias-correction term. At $Q_\theta = P_\phi$, uniform importance weights eliminate the χ^2-driven bias component; the positive-example normalization contributes a structural bias of order $O(1/K)$ (Proposition B.4, Appendix B). The following bound relates conditional variance to the same χ^2 divergence that controls the bias and motivates the variance measurements in the experiments (proof in Appendix B):

Proposition 4.4 (χ^2 bound on the self-normalized variance)**.** *Under Assumption 1(c) ($\|\nabla_\phi s_\phi(j,c)\| \le G$), let $\chi^2_c = \chi^2(P_\phi(\cdot|c)\|Q_\theta(\cdot|c))$ and $w_{\max} := \sup_{c,j} P_\phi(j|c)/Q_\theta(j|c)$. For any finite K and any fully supported Q_θ,*

$$\mathrm{Tr}(\mathbb{V}_{Q_\theta}[\hat{g}^{\mathrm{SN}} \mid c, i^+]) \le \frac{16G^2 w_{\max}^2}{K^2} + \frac{16G^2(1+2\chi^2_c)}{K}. \tag{12}$$

Substituting Proposition 4.4 and the asymptotic bias bound (Appendix, Eq. (25)) into the convergence residual of Theorem B.3 and dropping $O(K^{-2})$ remainders, the Q_θ-controllable dominant part is governed by the average divergence $\bar{\chi}^2 = \mathbb{E}_{(c,i^+)}[\chi^2(P_\phi\|Q_\theta)]$ and $w_{\max}$, both minimized at $Q_\theta = P_\phi$. Thus, as in the exact case (Theorems 4.2–4.3), driving $Q_\theta \to P_\phi$ tightens the convergence bound in a first-order asymptotic sense.

4.4 Implications for Cross-Stage Training

These conclusions unify several cross-stage co-training methods. Knowledge distillation Hinton et al. (2015); Tang & Wang (2018); Kang et al. (2020) optimizes the KL divergence, driving $Q_\theta \to P_\phi$, which makes the candidate generator the minimax-robust optimal proposal (Theorem 4.3) and tightens the variance residual of the convergence bound (Theorem 4.1); this is why distillation benefits the target classifier itself, not just the candidate generator, by supplying lower-variance gradients. Adaptive negative sampling Lian et al. (2020) concentrates Q_θ on high-scoring classes, closer to P_ϕ than uniform and thus lowering $M(Q_\theta)$. Because each update shifts P_{ϕ_t}, the optimum is a moving target: a fixed Q_θ drifts, whereas joint training Huang et al. (2023) keeps Q_θ matched to P_ϕ throughout, a positive feedback loop in which distillation supplies low-variance gradients and the improved target classifier yields a finer P_ϕ.

Our variance analysis also makes precise the common intuition that harder negatives help: the variance-optimal P_ϕ up-weights hard (high-scoring) classes, yet remains fully supported, so truncating to the top-κ classes drives the importance weights, and hence $M(Q_\theta)$, out of control; mixing in random samples Huang et al. (2023) restores full support and lowers the variance (Appendix D).

5 Generalization through Distribution Alignment

The candidate generator shapes both target-classifier training and the set of classes available at inference. The variance analysis describes its role during training; candidate coverage determines which classes the target classifier can predict. Distillation transfers information from the target classifier to the candidate generator, with teacher quality and the generator's approximation capacity governing the resulting coverage bound. We derive conditions under which weighted distillation tightens the coverage-loss upper bound relative to independent training, then combine coverage loss and target-classifier error in an end-to-end generalization bound.

5.1 Coverage-Loss Bounds under Weighted Distillation

For any class distribution Q, let $S_\kappa(Q) \subseteq \mathcal{I}$ be its κ most probable classes, and using a reference distribution P define

$$\Gamma_\kappa(Q;P) = P(S_\kappa(Q)), \qquad \Delta_\kappa(Q;P) = \Gamma_\kappa(P;P) - \Gamma_\kappa(Q;P),$$

i.e. the mass under P of the top-κ set retrieved according to Q, and the coverage lost relative to retrieving according to P itself; by maximality of the top-κ set, $\Delta_\kappa(Q;P) \in [0,1]$. This coverage loss is controlled by distributional distance: for any P, Q and $\kappa \le |\mathcal{I}|$, $\Delta_\kappa(Q;P) \le 2\,\mathrm{TV}(P,Q) \le \sqrt{2\,D_{\mathrm{KL}}(P\|Q)}$, with equality at $Q = P$ (Lemma B.5, proof in Appendix B). This reduces top-κ candidate-set coverage loss to a D_{KL}-divergence control—the foundational tool for the analysis below.

We next analyze a weighted distillation objective against the true class distribution P^*. Write $c = (u, \mathbf{z})$, where the candidate generator observes u and the target classifier additionally observes $\mathbf{z}$; let $P_c^* = P^*(\cdot|u, \mathbf{z})$ be the full-context class distribution and $\bar{P}_u^* = P^*(\cdot|u) = \mathbb{E}_{\mathbf{z}|u}[P_c^*]$ its $\mathbf{z} \mid u$ marginal. For a candidate-generator class $\mathcal{H}_R$ of u-measurable distributions, define the independent-fit error $e_{\text{ind}}(Q) = \mathbb{E}_u[D_{\text{KL}}(\bar{P}_u^* \| Q(\cdot|u))]$ (fitting the marginal truth) and the distillation error $e_{\text{dist}}(Q) = \mathbb{E}_{u,\mathbf{z}}[D_{\text{KL}}(P_\phi(\cdot|u, \mathbf{z}) \| Q(\cdot|u))]$ (imitating the target classifier), and set

$$Q_\lambda \in \arg\min_{Q \in \mathcal{H}_R} \big\{ e_{\text{ind}}(Q) + \lambda e_{\text{dist}}(Q) \big\}, \tag{13}$$

where $\lambda = 0$ recovers independent training. Two λ-independent reference errors are the teacher error $e_{\text{teacher}} = \mathbb{E}_{u,\mathbf{z}}[D_{\text{KL}}(P_c^* \| P_\phi(\cdot|u, \mathbf{z}))]$ and the hidden information $e_{\text{hidden}} = \mathbb{E}_{u,\mathbf{z}}[D_{\text{KL}}(P_c^* \| \bar{P}_u^*)]$, the latter unavailable to any u-measurable candidate generator Lopez-Paz et al. (2016). Its true-distribution coverage loss is

$$\Delta\tilde{\Gamma}_\kappa(Q) = \mathbb{E}_{u,\mathbf{z}}[\Delta_\kappa(Q(\cdot|u); P_c^*)]. \tag{14}$$

We bound this loss by the smaller of two routes: a direct route through $e_{\text{hidden}} + e_{\text{ind}}(\lambda)$ and a teacher route through $e_{\text{teacher}} + e_{\text{dist}}(\lambda)$ (proof in Appendix B).

Theorem 5.1 (True-class-distribution coverage upper bound for the weighted-distillation candidate generator)**.** *Write $e_{\text{ind}}(\lambda) = e_{\text{ind}}(Q_\lambda)$ and $e_{\text{dist}}(\lambda) = e_{\text{dist}}(Q_\lambda)$. For any $\lambda \geq 0$,*

$$\Delta\tilde{\Gamma}_\kappa(Q_\lambda) \leq B_\lambda := \min\left\{ \sqrt{2\big(e_{\text{hidden}} + e_{\text{ind}}(\lambda)\big)},\ \sqrt{2e_{\text{teacher}}} + \sqrt{2e_{\text{dist}}(\lambda)} \right\}. \tag{15}$$

A positive weight strictly improves this bound under an explicit condition (proof in Appendix B).

Corollary 5.2 (Sufficient condition for improvement under a positive weight)**.** *Let B_0 be the upper bound* (15) *at $\lambda = 0$.*

- *(i)* ***Trivial case: the teacher is no better than the independent baseline.*** *If $\sqrt{2e_{\text{teacher}}} \geq B_0$, then the teacher route of* (15)*, $\sqrt{2e_{\text{teacher}}} + \sqrt{2e_{\text{dist}}(\lambda)} \geq \sqrt{2e_{\text{teacher}}}$, is already at least B_0, so distillation through the teacher route cannot improve the upper bound.*
- *(ii)* ***Effective case: the teacher is better than the independent baseline.*** *If $\sqrt{2e_{\text{teacher}}} < B_0$, define the* distillation-route budget *$m_0 = (B_0 - \sqrt{2e_{\text{teacher}}})^2/2 > 0$. If there exists $\widetilde{Q} \in \mathcal{H}_R$ with $e_{\text{dist}}(\widetilde{Q}) < m_0$, then for any $\lambda > e_{\text{ind}}(\widetilde{Q})/(m_0 - e_{\text{dist}}(\widetilde{Q}))$ we have $B_\lambda < B_0$.*

The two conditions couple teacher quality and candidate-generator capacity through m_0. When both hold and λ lies in the specified interval, $B_\lambda < B_0$. Appendix C details B_0 and the thresholds (Tables 4 and 5).

5.2 End-to-End Generalization of the Two-Stage System

Recall the notation of Section 5.1: $c = (u, \mathbf{z})$, the candidate generator sees only u, the target classifier all of c, and $P_c^* = P^*(\cdot|u, \mathbf{z})$. Given budget κ, let $S(u) = S_\kappa(Q_\lambda(\cdot|u))$ and $\hat{y}(c) = \arg\max_{j \in S(u)} P_\phi(j|c)$. The end-to-end top-1 error is $\mathcal{R}_\kappa(\lambda) = \Pr_{c,\, y \sim P_c^*}[\hat{y}(c) \neq y] = 1 - \mathbb{E}_{u,\mathbf{z}}[P_c^*(\hat{y}(c))]$, and the Bayes error is $\mathcal{R}_{\text{Bayes}}^* = \mathbb{E}_{u,\mathbf{z}}[1 - \max_{j \in \mathcal{I}} P_c^*(j)]$.

Combining the retrieval-side and training-side bounds decomposes the end-to-end error (proof in Appendix B).

Theorem 5.3 (Two-sided decomposition of the end-to-end generalization error)**.** *For any $\lambda \geq 0$ and any $1 \leq \kappa \leq |\mathcal{I}|$, the two-stage system's end-to-end generalization error satisfies*

$$\mathcal{R}_\kappa(\lambda) \ \leq\ \mathcal{R}_{\text{Bayes}}^* \ + \ \underbrace{\Delta\tilde{\Gamma}_\kappa(Q_\lambda)}_{\textit{retrieval side: coverage loss}} \ + \ \underbrace{\sqrt{2\, e_{\text{teacher}}}}_{\textit{training side: target-classifier error}}, \tag{16}$$

where $\Delta\tilde{\Gamma}_\kappa(Q_\lambda)$ is the candidate generator's true-class-distribution coverage loss (Eq. (14)*) and e_{teacher} is the extent to which the target classifier deviates from the true class distribution.*

The bound splits the top-1 error into the irreducible Bayes error plus exactly the two controllable quantities of the two sides—the retrieval-side coverage loss $\Delta\tilde{\Gamma}_\kappa(Q_\lambda)$ and the training-side teacher error e_{teacher}. When no information is hidden from the candidate generator and $Q_\lambda = P_\phi = P_c^*$, both vanish and the bound reduces to the Bayes error.

Under the conditions of Corollary 5.2, distillation tightens the coverage-loss bound B_λ. On the training side, if positives follow P_c^*, e_{teacher} decomposes into optimization and approximation errors. Under PL-type gradient domination Polyak (1963); Karimi et al. (2016); Liu et al. (2022), the optimization gap is bounded by $\mathcal{E}_T(Q)/(2\mu)$ (Appendix C.3). Its proposal-dependent dominant term, $\bar{\chi}^2(Q)$, is minimized at $Q_\theta = P_\phi$ (Proposition 4.4).

6 Experiments

We study recommendation as an instance of large-scale classification: RQ1 measures gradient variance and RQ2 convergence and final accuracy. Appendices E.3 and E.4 examine standalone retrieval and the choice of distillation divergence.

6.1 Experimental Setup

We evaluate on MovieLens-10M and Amazon Electronics using MF Koren et al. (2009)+DeepFM Guo et al. (2017), SASRec Kang & McAuley (2018)+DIN Zhou et al. (2018), and XSimGCL Yu et al. (2024)+FinalMLP Mao et al. (2023). We compare Oracle, Distill, Retriever, Uniform, and the full-support Top-M mixture. Distill uses candidate-set KL with $\lambda = 0.1$. At inference, the retriever selects 100 candidates for re-ranking. RQ2 checkpoints are selected by validation end-to-end NDCG@10 before test evaluation. Data splits, architectures, proposal definitions, training protocols, and evaluation details appear in Appendix E.1. In Table 2, * denotes paired-test significance against the strongest baseline ($p < 0.05$).

6.2 Gradient Variance across Proposals (RQ1)

Distill and Retriever start from the same retriever checkpoint and receive the same number of training steps, differing only in the distillation term. With the pretrained ranker fixed, we compare self-normalized gradient variance and empirical χ^2 divergence using $N = 200$ examples, $R = 50$ repetitions, and $K = 20$ samples per repetition (Appendix E.1).

Table 1 shows that Oracle has the lowest measured variance and Distill the lowest non-Oracle variance and empirical χ^2 divergence in all six settings. For XSimGCL+FinalMLP, Distill reduces variance relative to Retriever by 8.51% on ML-10M and 26.10% on Amazon, even though the retriever already uses a contrastive objective.

Table 1: Sampling variance and χ^2 divergence.

Dataset	Model	Metric	Oracle	Distill	Retriever	Uniform	Top-M
ML-10M	MF+DeepFM	$\widehat{V}_{\text{IS}}$	12.32	**17.21**	21.29	30.17	41.98
		$\widehat{\chi}^2$	—	**0.68**	2.35	16.05	16.53
	XSimGCL+FinalMLP	$\widehat{V}_{\text{IS}}$	10.69	**19.67**	21.50	24.02	30.45
		$\widehat{\chi}^2$	—	**0.76**	2.32	19.72	23.09
	SASRec+DIN	$\widehat{V}_{\text{IS}}$	13.58	**19.75**	21.26	75.12	47.08
		$\widehat{\chi}^2$	—	**1.04**	1.57	6.30	23.97
Amazon	MF+DeepFM	$\widehat{V}_{\text{IS}}$	104.93	**177.89**	274.05	298.14	329.01
		$\widehat{\chi}^2$	—	**1.06**	4.79	4.92	4.70
	XSimGCL+FinalMLP	$\widehat{V}_{\text{IS}}$	95.24	**196.54**	265.97	347.49	417.30
		$\widehat{\chi}^2$	—	**3.79**	4.82	6.52	6.91
	SASRec+DIN	$\widehat{V}_{\text{IS}}$	15.42	**34.09**	35.10	36.39	48.52
		$\widehat{\chi}^2$	—	**1.44**	2.74	3.15	5.91

Table 2: Final ranking performance.

Dataset	Model	Metric	Distill	Retriever	Uniform	Top-M
ML-10M	MF+DeepFM	NDCG@10	**0.2495***	0.2454	0.2383	0.2432
		Recall@10	**0.2346***	0.2298	0.2200	0.2280
		MRR@10	**0.4037***	0.3976	0.3925	0.3970
	XSimGCL+FinalMLP	NDCG@10	**0.2854***	0.2529	0.2763	0.2554
		Recall@10	**0.2773***	0.2473	0.2501	0.2285
		MRR@10	**0.4710***	0.4012	0.4554	0.4102
	SASRec+DIN	NDCG@10	**0.1020***	0.0997	0.0930	0.0998
		Recall@10	**0.1913***	0.1829	0.1775	0.1860
		MRR@10	**0.1012***	0.1001	0.0957	0.0987
Amazon	MF+DeepFM	NDCG@10	**0.0220***	0.0211	0.0180	0.0209
		Recall@10	**0.0377***	0.0344	0.0300	0.0358
		MRR@10	**0.0274***	0.0261	0.0217	0.0229
	XSimGCL+FinalMLP	NDCG@10	**0.0260***	0.0219	0.0213	0.0250
		Recall@10	**0.0429***	0.0411	0.0316	0.0390
		MRR@10	**0.0293***	0.0268	0.0251	0.0238
	SASRec+DIN	NDCG@10	**0.0235***	0.0222	0.0219	0.0212
		Recall@10	**0.0389***	0.0367	0.0370	0.0378
		MRR@10	**0.0263***	0.0246	0.0243	0.0252

6.3 Convergence and Final Ranking Performance (RQ2)

We train with the four non-Oracle proposals under the same hyperparameters; Distill updates both stages, while the other methods freeze the retriever. Oracle is excluded because it requires full-catalog probabilities at every step.

Table 2 shows that Distill attains the highest NDCG@10, Recall@10, and MRR@10 in all six settings. On XSimGCL+FinalMLP, its NDCG@10 exceeds the strongest non-distillation baseline by 3.29% on ML-10M and 4.00% on Amazon. On ML-10M, Distill reaches NDCG@10 targets of 0.24 and 0.0979 in 5 and 6 epochs for MF+DeepFM and SASRec+DIN, respectively, and is fastest at medium-to-high target accuracies (Appendix E.2).

7 Conclusion

This work explains cross-stage knowledge distillation in large-scale classification through importance-sampling variance reduction. The candidate-generator distribution determines the sampling variance of target-classifier training. In the context of cross-stage information asymmetry, the output distribution of the target classifier is the minimax-optimal proposal for the unbiased estimator. For the practical self-normalized estimator, distribution alignment controls variance and bias bounds, linking proposal quality to convergence. Regarding retrieval, weighted distillation strictly tightens the coverage-loss upper bound compared to independent training, provided that the teacher quality, model capacity, and weight conditions are met. An end-to-end error decomposition links candidate coverage to target-classifier error, unifying the training and retrieval analyses. Experiments using real-world datasets demonstrate the practical value of this alignment. Among the compared proposals, Distill achieves the lowest non-Oracle sampling variance and the highest final ranking metrics across six dataset-model combinations. Further experiments demonstrate faster convergence and enhanced standalone retrieval coverage. Together, these results motivate training the candidate generator to align with the target classifier, both as a sampling proposal during training and as a candidate selector at inference.

AI USE STATEMENT

We used AI tools to assist with polishing the manuscript, including improving readability, refining the organization of the text, and polishing the wording. The authors take responsibility for the final content of the paper.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Yoshua Bengio and Jean-Sébastien Sénécal. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4):713–722, 2008.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 3163–3171, 2020.

Gang Chen, Jiawei Chen, Fuli Feng, Sheng Zhou, and Xiangnan He. Unbiased knowledge distillation for recommendation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pp. 976–984, 2023.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, pp. 191–198, 2016.

Luke Gallagher, Ruey-Cheng Chen, Roi Blanco, and J. Shane Culpepper. Joint optimization of cascade ranking models. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM'19)*, pp. 15–23, 2019. doi: 10.1145/3289600.3290986.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Edouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. Efficient softmax approximation for GPUs. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1302–1310, 2017.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of IJCAI'17*, pp. 1725–1731, 2017.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304, 2010.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. 2020. arXiv:2010.02666.

Jiri Hron, Karl Krauth, Michael I. Jordan, and Niki Kilbertus. On component interactions in two-stage recommender systems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2744–2757, 2021.

Xu Huang, Defu Lian, Jin Chen, Zheng Liu, Xing Xie, and Enhong Chen. Cooperative retriever and ranker in deep recommenders. In *Proceedings of the ACM Web Conference 2023 (WWW'23)*, Austin, TX, USA, 2023. doi: 10.1145/3543507.3583422.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1–10, 2015.

Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. LightXML: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 7987–7994, 2021.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pp. 315–323, 2013.

SeongKu Kang, Junyoung Hwang, Wonbin Kweon, and Hwanjo Yu. DE-RRD: A knowledge distillation framework for recommender system. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 605–614, 2020.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2525–2534, 2018.

Siddhant Kharbanda, Atmadeep Banerjee, Erik Schultheis, and Rohit Babbar. CascadeXML: Rethinking transformers for end-to-end multi-resolution training in extreme multi-label classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

Defu Lian, Qi Liu, and Enhong Chen. Personalized ranking with importance sampling. In *Proceedings of The Web Conference 2020*, pp. 1093–1103, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*, 2016.

Kelong Mao, Jieming Zhu, Liangcai Su, Guohao Cai, Yuru Li, and Zhenhua Dong. FinalMLP: An enhanced two-stream MLP model for CTR prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4552–4560, 2023.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424*, 2019.

Art B. Owen. *Monte Carlo theory, methods and examples*. Self-published, 2013. Available at `https://artowen.su.domains/mc/`.

Mary Phuong and Christoph H. Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning (ICML)*, 2019.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Yashoteja Prabhu and Manik Varma. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 263–272, 2014.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, pp. 993–1002, 2018.

Jiarui Qin, Jiachen Zhu, Bo Chen, Zhirong Liu, Weiwen Liu, Ruiming Tang, Rui Zhang, Yong Yu, and Weinan Zhang. Rankflow: Joint optimization of multi-stage cascade ranking systems as flows. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 814–824, 2022.

Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method*. John Wiley & Sons, 3rd edition, 2016.

Mher Safaryan, Alexandra Peste, and Dan Alistarh. Knowledge distillation performs partial variance reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiaxi Tang and Ke Wang. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2289–2298, 2018.

Yunli Wang, Zhen Zhang, Zhiqiang Wang, Zixuan Yang, Yu Li, Jian Yang, Shiyang Wen, Peng Jiang, and Kun Gai. Learning cascade ranking as one network. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Zhe Wang, Liqin Zhao, Biye Jiang, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. COLD: Towards the next generation of pre-ranking system. In *DLP-KDD 2020: Workshop on Deep Learning Practice for High-Dimensional Sparse Data with KDD*, 2020.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5812–5822, 2019.

Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Quoc Viet Hung Nguyen, and Hongzhi Yin. XSimGCL: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):913–926, 2024.

Jin Zhang, Ze Liu, Defu Lian, and Enhong Chen. Generalization error bounds for two-stage recommender systems with tree structure. *Advances in Neural Information Processing Systems*, 37:37070–37099, 2024.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1–9, 2015.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1059–1068, 2018.

A NOTATION

Table 3 summarizes the main notation.

Table 3: Summary of notation.

Symbol	Meaning
Models and training	
$\mathcal{I}$	Full class set.
c, i^+	Input and its positive class.
ϕ, θ	Target-classifier and candidate-generator parameters.
P_ϕ, Q_θ	Target-classifier distribution and candidate-generator proposal distribution.
$s_\phi(j,c)$, $s_\theta(j,c)$	Scores assigned to class j by the two models.
$Z_\phi(c)$	Target-classifier softmax partition function.
K, $\mathcal{S}$	Number of proposal draws and the sampled collection used by the surrogate loss.
$\mathcal{L}$, $\hat{\mathcal{L}}$, $\bar{\mathcal{L}}$	Full-softmax loss, sampled-softmax surrogate loss, and population loss.
$\hat{g}$, $\hat{g}^{\mathrm{SN}}$	Unbiased importance-sampling and self-normalized gradient estimators.
η, T	Learning rate and number of SGD iterations.
L, G	Loss smoothness constant and score-gradient norm bound.
Variance and proposal distributions	
V_{IS}, V_{data}	Sampling variance and variance due to training-example sampling.
σ^2	Uniform bound on the total gradient-estimator variance.
$w_{\max}$	Maximum importance ratio $P_\phi(j\|c)/Q_\theta(j\|c)$ over inputs and classes.
$M(Q_\theta)$	Proposal-dependent second-moment objective.
Q^*_θ, Q^{mm}_θ	Variance-optimal and minimax-optimal proposals.
χ^2_c, $\bar{\chi}^2$	Conditional χ^2 divergence between the two distributions and its average over training examples.
Coverage and generalization	
u, $\mathbf{z}$	Information observed by the candidate generator and additional information observed by the target classifier.
P^*_c, $\bar{P}^*_u$	True full-context class distribution and its marginal conditional on u.
$\mathcal{H}_R$	Candidate-generator model class.
κ, $S_\kappa(Q)$	Candidate budget and the top-κ class set under Q.
$\Delta_\kappa(Q;P)$	Coverage loss under reference distribution P.
$\Delta\tilde{\Gamma}_\kappa(Q)$	Expected coverage loss under the true class distribution.
λ, Q_λ	Distillation weight and a minimizer of the weighted training objective.
e_{ind}, e_{dist}	Expected KL errors for fitting the true marginal and imitating the teacher.
e_{teacher}, e_{hidden}	Teacher KL error and the KL discrepancy due to information unavailable to the candidate generator.
B_λ, B_0	Coverage-loss upper bound and its independent-training reference value.
m_0	Distillation-route error budget in the sufficient condition for bound improvement.
$\mathcal{R}_\kappa(\lambda)$, $\mathcal{R}^*_{\mathrm{Bayes}}$	End-to-end top-1 error and Bayes error.

B PROOFS

The regularity conditions (Assumption 1) are stated in Section 3.3.

Lemma B.1 (L-smoothness of the global loss)**.** *Under Assumption 1, $\mathcal{L}(\phi)$ is L-smooth, i.e., $\|\nabla^2\mathcal{L}(\phi)\|_{\mathrm{op}} \le L := 2\beta + G^2$, so that $\mathcal{L}(\phi') \le \mathcal{L}(\phi) + \nabla\mathcal{L}(\phi)^\top(\phi' - \phi) + \frac{L}{2}\|\phi' - \phi\|^2$ holds for all ϕ, ϕ'.*

PROOF OF LEMMA B.1

Proof. The following proof is carried out under the assumption that $s_\phi(j, c)$ is twice continuously differentiable in ϕ; in this case Assumption 1(b) is equivalent to a uniformly bounded Hessian, $\|\nabla^2_\phi s_\phi(j, c)\|_{\mathrm{op}} \le \beta$, so we may operate directly with the Hessian below. The conclusion holds equally in the general Lipschitz-gradient case, with a slightly different argument. For brevity, we omit the conditioning on c and write $\boldsymbol{g}_j = \nabla_\phi s_\phi(j, c)$ and $\boldsymbol{H}_j = \nabla^2_\phi s_\phi(j, c)$.

From $Z_\phi = \sum_{k\in\mathcal{I}} \exp(s_\phi(k, c))$, taking the first-order derivative of $\log Z_\phi$ gives

$$\nabla_\phi \log Z_\phi = \frac{1}{Z_\phi}\sum_{j\in\mathcal{I}} \exp(s_\phi(j, c))\, \boldsymbol{g}_j = \sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{g}_j = \mathbb{E}_{P_\phi}[\boldsymbol{g}_j].$$

We differentiate the above once more. Note that $P_\phi(j) = \exp(s_\phi(j, c))/Z_\phi$; by the quotient rule,

$$\nabla_\phi P_\phi(j) = P_\phi(j)\big(\boldsymbol{g}_j - \mathbb{E}_{P_\phi}[\boldsymbol{g}_k]\big).$$

Applying the product rule term by term to $\nabla_\phi \log Z_\phi = \sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{g}_j$ and substituting the above:

$$\begin{aligned}
\nabla^2 \log Z_\phi &= \sum_{j\in\mathcal{I}} \boldsymbol{g}_j \big[\nabla_\phi P_\phi(j)\big]^\top + \sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{H}_j \\
&= \sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{g}_j \big(\boldsymbol{g}_j - \mathbb{E}_{P_\phi}[\boldsymbol{g}_k]\big)^\top + \sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{H}_j \\
&= \underbrace{\sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{g}_j\boldsymbol{g}_j^\top - \mathbb{E}_{P_\phi}[\boldsymbol{g}_j]\, \mathbb{E}_{P_\phi}[\boldsymbol{g}_j]^\top}_{=\mathrm{Cov}_{P_\phi}[\boldsymbol{g}_j]} + \sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{H}_j.
\end{aligned}$$

From $\mathcal{L}(\phi) = -s_\phi(i^+, c) + \log Z_\phi(c)$, taking the second-order derivative directly gives

$$\nabla^2\mathcal{L} = -\boldsymbol{H}_{i^+} + \sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{H}_j + \mathrm{Cov}_{P_\phi}[\boldsymbol{g}_j].$$

We now bound the operator norm of each term. By Assumption 1(b), $\|\boldsymbol{H}_{i^+}\|_{\mathrm{op}} \le \beta$. For the second term, by the triangle inequality and the property of convex combinations,

$$\Big\|\sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{H}_j\Big\|_{\mathrm{op}} \le \sum_{j\in\mathcal{I}} P_\phi(j)\, \|\boldsymbol{H}_j\|_{\mathrm{op}} \le \beta.$$

For the covariance term, expanding its trace by definition:

$$\begin{aligned}
\mathrm{Tr}\big(\mathrm{Cov}_{P_\phi}[\boldsymbol{g}_j]\big) &= \mathrm{Tr}\Big(\sum_{j\in\mathcal{I}} P_\phi(j)\, \boldsymbol{g}_j\boldsymbol{g}_j^\top - \mathbb{E}_{P_\phi}[\boldsymbol{g}_j]\, \mathbb{E}_{P_\phi}[\boldsymbol{g}_j]^\top\Big) \\
&= \sum_{j\in\mathcal{I}} P_\phi(j)\, \mathrm{Tr}(\boldsymbol{g}_j\boldsymbol{g}_j^\top) - \mathrm{Tr}\big(\mathbb{E}_{P_\phi}[\boldsymbol{g}_j]\, \mathbb{E}_{P_\phi}[\boldsymbol{g}_j]^\top\big) \\
&= \sum_{j\in\mathcal{I}} P_\phi(j)\, \|\boldsymbol{g}_j\|^2 - \|\mathbb{E}_{P_\phi}[\boldsymbol{g}_j]\|^2 \le \mathbb{E}_{P_\phi}\big[\|\boldsymbol{g}_j\|^2\big],
\end{aligned}$$

where the last inequality follows from $\|\mathbb{E}_{P_\phi}[\boldsymbol{g}_j]\|^2 \ge 0$. The covariance matrix is positive semidefinite, so its operator norm does not exceed the trace, hence

$$\|\mathrm{Cov}_{P_\phi}[\boldsymbol{g}_j]\|_{\mathrm{op}} \le \mathrm{Tr}\big(\mathrm{Cov}_{P_\phi}[\boldsymbol{g}_j]\big) \le \mathbb{E}_{P_\phi}\big[\|\boldsymbol{g}_j\|^2\big] \le G^2,$$

where the last step uses Assumption 1(c). Combining the three terms by the triangle inequality,

$$\|\nabla^2\mathcal{L}\|_{\mathrm{op}} \le \beta + \beta + G^2 = 2\beta + G^2 =: L.$$

To derive the quadratic upper bound, apply the integral-remainder form of the Taylor expansion along the segment $\phi(t) = \phi + t(\phi' - \phi)$:

$$\mathcal{L}(\phi') = \mathcal{L}(\phi) + \nabla\mathcal{L}(\phi)^\top(\phi' - \phi) + \int_0^1 (1-t)\,(\phi' - \phi)^\top \nabla^2\mathcal{L}(\phi(t))\,(\phi' - \phi)\,\mathrm{d}t.$$

By $\|\nabla^2\mathcal{L}(\phi(t))\|_{\mathrm{op}} \le L$, bounding the remainder and integrating gives

$$\mathcal{L}(\phi') \le \mathcal{L}(\phi) + \nabla\mathcal{L}(\phi)^\top(\phi' - \phi) + \frac{L}{2}\,\|\phi' - \phi\|^2.$$

□

Proofs for Section 4

B.0.1 Proof of Theorem 4.1

Proof. At each step we sample $(c_t, i_t^+) \sim \mathcal{D}$, draw K negatives from $Q_\theta(\cdot|c_t)$, and form the estimator $\hat{g}_t$; write $\bar{\mathcal{L}}^* = \inf_\phi \bar{\mathcal{L}}(\phi)$. $\bar{\mathcal{L}}$ is the expectation of $\mathcal{L}(\phi; c, i^+)$, and by Lemma B.1, $\mathcal{L}(\phi; c, i^+)$ is L-smooth, so $\bar{\mathcal{L}}$ is L-smooth as well. Applying the smoothness of $\bar{\mathcal{L}}$ to the SGD update $\phi_{t+1} = \phi_t - \eta\,\hat{g}_t$:

$$\bar{\mathcal{L}}(\phi_{t+1}) \le \bar{\mathcal{L}}(\phi_t) - \eta\,\nabla\bar{\mathcal{L}}(\phi_t)^\top \hat{g}_t + \frac{\eta^2 L}{2}\|\hat{g}_t\|^2.$$

Take the full expectation over (c_t, i_t^+) and the negative samples (conditional on ϕ_t). Since $\hat{g}_t$ is unbiased for $\bar{\mathcal{L}}$ ($\mathbb{E}[\hat{g}_t \mid \phi_t] = \nabla\bar{\mathcal{L}}(\phi_t)$),

$$\mathbb{E}[\nabla\bar{\mathcal{L}}(\phi_t)^\top \hat{g}_t \mid \phi_t] = \|\nabla\bar{\mathcal{L}}(\phi_t)\|^2.$$

For the second-moment term, by the variance decomposition $\mathbb{E}[\|\hat{g}_t\|^2 \mid \phi_t] = \|\nabla\bar{\mathcal{L}}(\phi_t)\|^2 + \mathrm{Tr}(\mathbb{V}[\hat{g}_t \mid \phi_t])$. Substituting and using $\mathrm{Tr}(\mathbb{V}[\hat{g}_t \mid \phi_t]) \le \sigma^2$:

$$\mathbb{E}[\bar{\mathcal{L}}(\phi_{t+1}) \mid \phi_t] \le \bar{\mathcal{L}}(\phi_t) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla\bar{\mathcal{L}}(\phi_t)\|^2 + \frac{\eta^2 L}{2}\,\sigma^2.$$

Taking the full expectation and summing from $t = 0$ to $T - 1$:

$$\mathbb{E}[\bar{\mathcal{L}}(\phi_T)] - \bar{\mathcal{L}}(\phi_0) \le -\alpha \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\bar{\mathcal{L}}(\phi_t)\|^2] + \frac{\eta^2 L}{2}\,T\,\sigma^2,$$

where $\alpha = \eta(1 - \eta L/2) > 0$. Since $\mathbb{E}[\bar{\mathcal{L}}(\phi_T)] \ge \bar{\mathcal{L}}^*$, rearranging and dividing by αT yields Eq. (8). □

B.0.2 Decomposition of the unnormalized gradient variance

The following decomposition, stated in one sentence in Section 4.1, defines the sampling variance V_{IS} and the data variance V_{data} used throughout Section 4.

Proposition B.2 (Full decomposition of the unnormalized gradient variance)**.** *Let $Y = (c, i^+) \sim \mathcal{D}$ be a training example, and $j_1, \ldots, j_K \sim Q_\theta(\cdot|c)$ be negative samples. For the unnormalized estimator $\hat{g}_t$ (Eq. (5)), $\mathbb{E}[\hat{g}_t] = \nabla\bar{\mathcal{L}}(\phi_t)$ holds exactly, and the total variance satisfies*

$$\mathrm{Tr}(\mathbb{V}[\hat{g}_t]) = \underbrace{\mathbb{E}_{Y\sim\mathcal{D}}\big[\mathrm{Tr}(\mathbb{V}_{Q_\theta}[\hat{g}_t \mid Y])\big]}_{V_{\mathrm{IS}}(\phi_t, Q_\theta)\ \textit{(sampling variance)}} + \underbrace{\mathrm{Tr}\big(\mathbb{V}_{Y\sim\mathcal{D}}\big[\nabla\mathcal{L}(\phi_t; c, i^+)\big]\big)}_{V_{\mathrm{data}}(\phi_t)\ \textit{(data variance)}}. \tag{17}$$

This decomposition holds exactly for any K. V_{IS} depends on Q_θ, whereas V_{data} does not.

Proof. Write $f(j) = \nabla_\phi s_\phi(j, c)$. Applying the law of total variance to the random vector $\hat{g}_t$:

$$\mathbb{V}[\hat{g}_t] = \mathbb{E}_Y\big[\mathbb{V}[\hat{g}_t \mid Y]\big] + \mathbb{V}_Y\big[\mathbb{E}[\hat{g}_t \mid Y]\big]. \tag{18}$$

For the unnormalized estimator, the conditional expectation given $Y = (c, i^+)$ is

$$\begin{aligned}\mathbb{E}_{Q_\theta}[\hat{g}_t \mid c, i^+] &= -f(i^+) + \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{j_k\sim Q_\theta}\left[\frac{P_\phi(j_k|c)}{Q_\theta(j_k|c)}\,f(j_k)\right] \\ &= -f(i^+) + \sum_{j\in\mathcal{I}} P_\phi(j|c)\,f(j) = \nabla\mathcal{L}(\phi_t; c, i^+).\end{aligned}$$

This identity holds exactly for any K and any fully supported Q_θ. Taking the expectation over Y gives $\mathbb{E}[\hat{g}_t] = \nabla\mathcal{L}(\phi_t)$.

Substituting into Eq. (18): in the first term, $\mathbb{V}[\hat{g}_t \mid Y] = \mathbb{V}_{Q_\theta}[\hat{g}_t \mid c, i^+]$ is the conditional variance with respect to negative-sample sampling; taking its trace and the expectation over Y gives V_{IS}; in the second term, $\mathbb{E}[\hat{g}_t \mid Y] = \nabla\mathcal{L}(\phi_t; c, i^+)$, and taking the trace of the outer variance gives V_{data}. Since Q_θ cancels in the conditional unbiasedness, V_{data} does not contain Q_θ. □

The statement of Theorem 4.2 appears in Section 4.2; we use the full-support convention that a term $P_\phi(j|c)^2\|\nabla_\phi s_\phi(j,c)\|^2/Q_\theta(j|c)$ with $Q_\theta(j|c) = 0$ is taken to be $+\infty$, consistent with the full-support requirement of importance sampling. For completeness, the conditional variance used in Section 4.2 is

$$\mathrm{Tr}\big(\mathbb{V}_{Q_\theta}[\hat{g} \mid c, i^+]\big) = \frac{1}{K}\left[\sum_{j\in\mathcal{I}} \frac{P_\phi(j|c)^2}{Q_\theta(j|c)}\|\nabla_\phi s_\phi(j,c)\|^2 - \big\|\nabla_\phi\mathcal{L}(\phi; c, i^+) + \nabla_\phi s_\phi(i^+, c)\big\|^2\right]. \tag{19}$$

B.0.3 Proof of Theorem 4.2

Proof. Let $f_j = P_\phi(j|c)\|\nabla_\phi s_\phi(j,c)\| > 0$ (by the full support of P_ϕ and the strict positivity of the gradient norm), so that the problem becomes

$$\min_{Q\in\Delta_N^\circ} \sum_j f_j^2/Q_j.$$

By the Cauchy–Schwarz inequality,

$$\sum_{j\in\mathcal{I}} \frac{f_j^2}{Q_j} \geq \frac{\big(\sum_{j\in\mathcal{I}} f_j\big)^2}{\sum_{j\in\mathcal{I}} Q_j} = \Big(\sum_{j\in\mathcal{I}} f_j\Big)^2,$$

where the inequality takes $a_j = f_j/\sqrt{Q_j}$ and $b_j = \sqrt{Q_j}$, using $\sum_{j\in\mathcal{I}} Q_j = 1$. Equality holds if and only if $f_j/\sqrt{Q_j} \propto \sqrt{Q_j}$, i.e., $Q_j \propto f_j$, whence $Q_j^* = f_j/\sum_i f_i \in \Delta_N^\circ$; uniqueness follows from the equality condition of Cauchy–Schwarz. □

B.0.4 Proof of Theorem 4.3

Proof. The proof proceeds in the order "first solve the inner maximization, then solve the outer minimization."

For fixed $Q_\theta \in \Delta_N^\circ$, the inner problem is

$$\max_{g_j\geq 0,\ \sum_{j\in\mathcal{I}} P_\phi(j)\, g_j^2 \leq C} \sum_{j\in\mathcal{I}} \frac{P_\phi(j)^2}{Q_\theta(j)}\, g_j^2.$$

Let $h_j = g_j^2 \geq 0$; the above becomes

$$\max_{h_j\geq 0} \sum_{j\in\mathcal{I}} c_j\, h_j, \quad \text{s.t.} \quad \sum_{j\in\mathcal{I}} a_j\, h_j \leq C,$$

where $c_j = P_\phi(j)^2/Q_\theta(j) > 0$ and $a_j = P_\phi(j) > 0$. This is a linear program. For any feasible solution h, let its support be $S = \{j : h_j > 0\}$. If $|S| \geq 2$, take $j_1, j_2 \in S$ with $c_{j_1}/a_{j_1} \geq c_{j_2}/a_{j_2}$, and transfer the entire budget of h_{j_2} to j_1 (set $h_{j_1} \leftarrow h_{j_1} + a_{j_2}h_{j_2}/a_{j_1}$, $h_{j_2} \leftarrow 0$); the constraint remains satisfied while the objective does not decrease. Repeating this, the entire budget concentrates on $j^* = \arg\max_j P_\phi(j)/Q_\theta(j)$, and the maximum value is $C \cdot \max_{j\in\mathcal{I}} P_\phi(j)/Q_\theta(j)$.

The outer problem simplifies to $\min_{Q_\theta\in\Delta_N^\circ} \max_{j\in\mathcal{I}} P_\phi(j)/Q_\theta(j)$. Note that

$$\max_j \frac{P_\phi(j)}{Q_\theta(j)} \geq \sum_j Q_\theta(j) \cdot \frac{P_\phi(j)}{Q_\theta(j)} = \sum_j P_\phi(j) = 1,$$

with equality if and only if $P_\phi(j)/Q_\theta(j)$ is constant for all j; combined with $\sum_j Q_\theta(j) = \sum_j P_\phi(j) = 1$, this constant equals 1, hence $Q_\theta^{\mathrm{mm}} = P_\phi$, with minimax value C.

Case of the pointwise uncertainty set. Under the stricter constraint $\{g_j \le G, \forall j\}$, the inner maximization is attained at $g_j = G$, giving $\max V = G^2 \sum_j P_\phi(j)^2/Q_\theta(j)$; the outer $\min_{Q_\theta \in \Delta_N^\circ} \sum_j P_\phi(j)^2/Q_\theta(j)$ is bounded below by 1 via the Cauchy–Schwarz inequality, with equality if and only if $Q_\theta = P_\phi$. Under both uncertainty sets the minimax value is G^2 and the optimal solution is P_ϕ. □

PROOFS FOR SECTION 4.3

B.0.5 CONVERGENCE OF THE SELF-NORMALIZED ESTIMATOR (STATEMENT AND PROOF)

Under the same SGD setting as Theorem 4.1, let the full-expectation bias be $\bar{\boldsymbol{b}}_t = \mathbb{E}[\hat{g}_t^{\mathrm{SN}} \mid \phi_t] - \nabla\bar{\mathcal{L}}(\phi_t)$, $b = \sup_t \|\bar{\boldsymbol{b}}_t\|$, and the total variance bound $\sigma_{\mathrm{SN}}^2 = \sup_t \mathrm{Tr}(\mathbb{V}[\hat{g}_t^{\mathrm{SN}} \mid \phi_t])$.

Theorem B.3 (Convergence rate of the self-normalized estimator)**.** *Under Assumption 1 (with $L = 2\beta + G^2$), suppose the target classifier runs T steps of SGD on $\bar{\mathcal{L}}(\phi)$ with learning rate $\eta \in (0, 2/L)$, using the self-normalized gradient $\hat{g}_t^{\mathrm{SN}}$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla\bar{\mathcal{L}}(\phi_t)\|^2\big] \le \underbrace{\frac{2(\bar{\mathcal{L}}(\phi_0) - \bar{\mathcal{L}}^*)}{\eta(1-\eta L/2)\,T}}_{\to 0 \text{ as } T\to\infty} + \underbrace{\frac{\eta L}{1-\eta L/2}\,\sigma_{\mathrm{SN}}^2 + \left[\frac{(1+\eta L)^2}{(1-\eta L/2)^2} + \frac{\eta L}{1-\eta L/2}\right] b^2}_{\text{asymptotic residual of variance and bias}}. \tag{20}$$

Proof. By Lemma B.1, each $\mathcal{L}(\phi; c, i^+)$ is L-smooth, hence $\bar{\mathcal{L}}$ is L-smooth as well. Applying the quadratic upper bound of $\bar{\mathcal{L}}$ to the SGD update $\phi_{t+1} = \phi_t - \eta\,\hat{g}_t^{\mathrm{SN}}$:

$$\bar{\mathcal{L}}(\phi_{t+1}) \le \bar{\mathcal{L}}(\phi_t) - \eta\,\nabla\bar{\mathcal{L}}(\phi_t)^\top \hat{g}_t^{\mathrm{SN}} + \frac{\eta^2 L}{2}\|\hat{g}_t^{\mathrm{SN}}\|^2.$$

Decompose $\hat{g}_t^{\mathrm{SN}}$ into a deterministic part and a stochastic part: $\hat{g}_t^{\mathrm{SN}} = \nabla\bar{\mathcal{L}}(\phi_t) + \bar{\boldsymbol{b}}_t + \boldsymbol{\epsilon}_t$, where $\bar{\boldsymbol{b}}_t = \mathbb{E}[\hat{g}_t^{\mathrm{SN}} \mid \phi_t] - \nabla\bar{\mathcal{L}}(\phi_t)$ is the bias and $\boldsymbol{\epsilon}_t = \hat{g}_t^{\mathrm{SN}} - \mathbb{E}[\hat{g}_t^{\mathrm{SN}} \mid \phi_t]$ is zero-mean noise. Taking the full expectation (conditioned on ϕ_t), we handle each term in turn.

For the cross term,

$$\begin{aligned}\mathbb{E}[\nabla\bar{\mathcal{L}}(\phi_t)^\top \hat{g}_t^{\mathrm{SN}} \mid \phi_t] &= \nabla\bar{\mathcal{L}}(\phi_t)^\top(\nabla\bar{\mathcal{L}}(\phi_t) + \bar{\boldsymbol{b}}_t)\\ &= \|\nabla\bar{\mathcal{L}}(\phi_t)\|^2 + \nabla\bar{\mathcal{L}}(\phi_t)^\top\bar{\boldsymbol{b}}_t.\end{aligned}$$

By Cauchy–Schwarz, $\nabla\bar{\mathcal{L}}(\phi_t)^\top\bar{\boldsymbol{b}}_t \ge -\|\nabla\bar{\mathcal{L}}(\phi_t)\|\cdot b$, hence

$$\mathbb{E}[\nabla\bar{\mathcal{L}}(\phi_t)^\top \hat{g}_t^{\mathrm{SN}} \mid \phi_t] \ge \|\nabla\bar{\mathcal{L}}(\phi_t)\|^2 - \|\nabla\bar{\mathcal{L}}(\phi_t)\|\cdot b. \tag{21}$$

For the second-moment term, using $\mathbb{E}[\|\boldsymbol{X}\|^2] = \|\mathbb{E}[\boldsymbol{X}]\|^2 + \mathrm{Tr}(\mathbb{V}[\boldsymbol{X}])$,

$$\begin{aligned}\mathbb{E}[\|\hat{g}_t^{\mathrm{SN}}\|^2 \mid \phi_t] &= \|\mathbb{E}[\hat{g}_t^{\mathrm{SN}} \mid \phi_t]\|^2 + \mathrm{Tr}(\mathbb{V}[\hat{g}_t^{\mathrm{SN}} \mid \phi_t])\\ &= \|\nabla\bar{\mathcal{L}}(\phi_t) + \bar{\boldsymbol{b}}_t\|^2 + \mathrm{Tr}(\mathbb{V}[\hat{g}_t^{\mathrm{SN}} \mid \phi_t]).\end{aligned}$$

By the triangle inequality and $\mathrm{Tr}(\mathbb{V}[\hat{g}_t^{\mathrm{SN}} \mid \phi_t]) \le \sigma_{\mathrm{SN}}^2$,

$$\mathbb{E}[\|\hat{g}_t^{\mathrm{SN}}\|^2 \mid \phi_t] \le \|\nabla\bar{\mathcal{L}}(\phi_t)\|^2 + 2\|\nabla\bar{\mathcal{L}}(\phi_t)\|\cdot b + b^2 + \sigma_{\mathrm{SN}}^2. \tag{22}$$

Substituting Eqs. (21) and (22) back into the smoothness inequality and taking expectations, with $\alpha = \eta(1-\eta L/2) > 0$ and $r_t = \|\nabla\bar{\mathcal{L}}(\phi_t)\|$,

$$\begin{aligned}\mathbb{E}[\bar{\mathcal{L}}(\phi_{t+1}) \mid \phi_t] &\le \bar{\mathcal{L}}(\phi_t) - \eta(r_t^2 - r_t b) + \frac{\eta^2 L}{2}(r_t^2 + 2r_t b + b^2 + \sigma_{\mathrm{SN}}^2)\\ &= \bar{\mathcal{L}}(\phi_t) - \alpha\, r_t^2 + \eta(1+\eta L)\, r_t b + \frac{\eta^2 L}{2}(b^2 + \sigma_{\mathrm{SN}}^2).\end{aligned}$$

Applying Young's inequality $r_t b \le \frac{\alpha}{2\eta(1+\eta L)}\, r_t^2 + \frac{\eta(1+\eta L)}{2\alpha}\, b^2$ to the cross term,

$$\mathbb{E}[\bar{\mathcal{L}}(\phi_{t+1}) \mid \phi_t] \le \bar{\mathcal{L}}(\phi_t) - \frac{\alpha}{2}\, r_t^2 + \frac{\eta^2(1+\eta L)^2}{2\alpha}\, b^2 + \frac{\eta^2 L}{2}(b^2 + \sigma_{\mathrm{SN}}^2).$$

Taking the full expectation and summing from $t = 0$ to $T-1$,

$$\mathbb{E}[\bar{\mathcal{L}}(\phi_T)] - \bar{\mathcal{L}}(\phi_0) \le -\frac{\alpha}{2}\sum_{t=0}^{T-1}\mathbb{E}[r_t^2] + T\left[\frac{\eta^2(1+\eta L)^2}{2\alpha} + \frac{\eta^2 L}{2}\right] b^2 + \frac{\eta^2 L\, T}{2}\,\sigma_{\mathrm{SN}}^2.$$

Since $\mathbb{E}[\bar{\mathcal{L}}(\phi_T)] \geq \bar{\mathcal{L}}^*$, rearranging and dividing by $\frac{\alpha T}{2}$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[r_t^2] \leq \frac{2(\bar{\mathcal{L}}(\phi_0)-\bar{\mathcal{L}}^*)}{\alpha\, T} + \frac{\eta^2 L}{\alpha}\,\sigma_{\mathrm{SN}}^2 + \left[\frac{\eta^2(1+\eta L)^2}{\alpha^2} + \frac{\eta^2 L}{\alpha}\right] b^2.$$

Substituting $\alpha = \eta(1-\eta L/2)$ back and simplifying the coefficients yields Eq. (20). □

B.0.6 Full decomposition of the self-normalized gradient variance (statement and proof)

The total variance of the self-normalized estimator admits the same data/sampling decomposition as the unbiased case (Proposition B.2), plus an extra term induced by the bias.

Proposition B.4 (Full decomposition of the self-normalized gradient variance)**.** *We follow the notation of Proposition B.2, and further assume the importance ratio is uniformly bounded,* $w_{\max} := \sup_{c,j} P_\phi(j|c)/Q_\theta(j|c) < \infty$.

(i) Exact decomposition. *For the self-normalized estimator* $\hat{g}_t^{\mathrm{SN}}$ *(Eq.* (7)*), the law of total variance holds exactly for any finite* $K \geq 1$*:*

$$\mathrm{Tr}(\mathbb{V}[\hat{g}_t^{\mathrm{SN}}]) = \underbrace{\mathbb{E}_{Y\sim\mathcal{D}}\big[\mathrm{Tr}(\mathbb{V}_{Q_\theta}[\hat{g}_t^{\mathrm{SN}} \mid Y])\big]}_{\text{conditional variance (depends on } Q_\theta)} + \underbrace{V_{\mathrm{data}}(\phi_t) + \delta_K(\phi_t, Q_\theta)}_{\text{data variance + bias-correction term}}, \tag{23}$$

where $\delta_K(\phi_t, Q_\theta) = \mathrm{Tr}(\mathbb{V}_Y[\boldsymbol{b}_t(c,i^+)]) + 2\,\mathrm{Tr}(\mathrm{Cov}_Y[\nabla\mathcal{L}(\phi_t;c,i^+),\, \boldsymbol{b}_t(c,i^+)])$*, and* $\boldsymbol{b}_t(c,i^+) = \mathbb{E}_{Q_\theta}[\hat{g}_t^{\mathrm{SN}} \mid Y] - \nabla\mathcal{L}(\phi_t;c,i^+)$ *is the conditional bias.*

(ii) Asymptotic bias bound. *In the first-order asymptotic regime as* $K \to \infty$*, writing* $w(j) = P_\phi(j|c)/Q_\theta(j|c)$*, the conditional bias satisfies*

$$\|\boldsymbol{b}_t(c,i^+)\| \leq \frac{2G\, w_{\max}}{K} + \frac{G\sqrt{w_{\max}}}{K}\sqrt{\chi^2(P_\phi(\cdot|c)\|Q_\theta(\cdot|c))} + O(K^{-2}), \tag{24}$$

whose first term arises from the positive example participating in self-normalization and whose second term is the standard SNIS bias. The bias-correction term is in turn controlled by $w_{\max}$ *and the average* χ^2 *divergence: there exists a nonnegative remainder* $R_K = O(K^{-3})$ *such that*

$$|\delta_K(\phi_t, Q_\theta)| \leq A_K + R_K + 2\sqrt{V_{\mathrm{data}}(\phi_t)}\sqrt{A_K + R_K}, \qquad A_K := \frac{8G^2 w_{\max}^2}{K^2} + \frac{2G^2 w_{\max}}{K^2}\,\bar{\chi}^2, \tag{25}$$

where $\bar{\chi}^2 = \mathbb{E}_{(c,i^+)\sim\mathcal{D}}[\chi^2(P_\phi(\cdot|c)\|Q_\theta(\cdot|c))]$.

Proof. Let $f(j) = \nabla_\phi s_\phi(j,c)$ and $\mu = \mathbb{E}_{P_\phi}[f(j)]$. The law of total variance (18) holds exactly for any random vector and applies directly to $\hat{g}_t^{\mathrm{SN}}$.

For the first term $\mathbb{E}_Y[\mathbb{V}[\hat{g}_t^{\mathrm{SN}} \mid Y]]$, the quantity $\mathbb{V}[\hat{g}_t^{\mathrm{SN}} \mid Y]$ is the conditional covariance over the negative sampling alone, given the training example (c, i^+), and is well defined for any finite K. Taking the trace and the expectation over $Y \sim \mathcal{D}$ yields the first term of Eq. (23).

For the second term $\mathbb{V}_Y[\mathbb{E}[\hat{g}_t^{\mathrm{SN}} \mid Y]]$, we need to compute the conditional expectation $\mathbb{E}_{Q_\theta}[\hat{g}_t^{\mathrm{SN}} \mid Y]$. Unlike in Proposition B.2, the self-normalized estimator is biased for the single-example loss; denoting the bias by

$$\boldsymbol{b}_t(c,i^+) = \mathbb{E}_{Q_\theta}[\hat{g}_t^{\mathrm{SN}} \mid Y] - \nabla\mathcal{L}(\phi_t;c,i^+),$$

the outer variance expands as

$$\mathbb{V}_Y[\nabla\mathcal{L}(\phi_t;c,i^+) + \boldsymbol{b}_t(c,i^+)] = \mathbb{V}_Y[\nabla\mathcal{L}] + \mathbb{V}_Y[\boldsymbol{b}_t(c,i^+)] + 2\,\mathrm{Cov}_Y[\nabla\mathcal{L},\, \boldsymbol{b}_t(c,i^+)].$$

Taking the trace: the first term is $\mathrm{Tr}(\mathbb{V}_Y[\nabla\mathcal{L}]) = V_{\mathrm{data}}(\phi_t)$, and the latter two together are defined as $\delta_K(\phi_t, Q_\theta)$. Substituting into (18) and taking the trace yields (23).

Derivation of the asymptotic bias bounds (24) *and* (25). We now derive, in the first-order asymptotic sense as $K \to \infty$, the conditional bias bound. For the corrected $\hat{g}^{\mathrm{SN}}$ (Eq. (7)), performing a first-order delta-method expansion of the ratio estimator, the conditional bias decomposes as:

$$\boldsymbol{b}_t(c,i^+) = \underbrace{\frac{w(i^+)(f(i^+)-\mu)}{K}}_{\text{positive-example normalization bias}} - \underbrace{\frac{1}{K}\,\mathrm{Cov}_{P_\phi}[w(j),\, f(j)]}_{\text{standard SNIS bias}} + O(1/K^2). \tag{26}$$

For the first term, by $\|f(i^+) - \mu\| \le 2G$ (Assumption 1(c)) and $w(i^+) \le w_{\max}$,

$$\left\| \frac{w(i^+)(f(i^+) - \mu)}{K} \right\| \le \frac{2G\, w_{\max}}{K}.$$

For the second term, applying Cauchy–Schwarz under P_ϕ:

$$\|\mathrm{Cov}_{P_\phi}[w(j),\, f(j)]\| \le \sqrt{\mathbb{E}_{P_\phi}[(w(j)-1)^2]} \cdot \sqrt{\mathbb{E}_{P_\phi}[\|f(j) - \mu\|^2]}.$$

By the variance identity, $\sqrt{\mathbb{E}_{P_\phi}[\|f(j) - \mu\|^2]} \le G$. For the first factor, $\mathbb{E}_{P_\phi}[(w-1)^2] \le w_{\max}\, \chi^2(P_\phi \| Q_\theta)$. Combining yields the conditional bias bound (24).

We then bound the bias-correction term δ_K, writing $\bar{\chi}^2 = \mathbb{E}_{(c,i^+)\sim\mathcal{D}}[\chi^2(P_\phi(\cdot|c)\|Q_\theta(\cdot|c))]$ for the average χ^2 divergence. For the bias-variance term, by $(a+b)^2 \le 2a^2+2b^2$, there exists a nonnegative remainder $R_K = O(K^{-3})$ such that

$$\mathrm{Tr}(\mathbb{V}_Y[\boldsymbol{b}_t(c,i^+)]) \le \mathbb{E}_Y[\|\boldsymbol{b}_t(c,i^+)\|^2] \le \frac{8G^2 w_{\max}^2}{K^2} + \frac{2G^2 w_{\max}}{K^2}\,\bar{\chi}^2 + R_K = A_K + R_K.$$

For the covariance term, by the matrix Cauchy–Schwarz inequality,

$$|\mathrm{Tr}(\mathrm{Cov}_Y[\nabla\mathcal{L}, \boldsymbol{b}_t(c,i^+)])| \le \sqrt{V_{\mathrm{data}}(\phi_t)} \cdot \sqrt{A_K + R_K}.$$

Combining the two terms yields the bias-correction bound (25). □

Compared with the unnormalized case (Proposition B.2), the decomposition (23) carries an extra bias-correction term δ_K. When $Q_\theta = P_\phi$, the importance weights satisfy $w(j) \equiv 1$ and the standard SNIS bias vanishes; the positive-example normalization bias $w(i^+)(f(i^+)-\mu)/K = (f(i^+)-\mu)/K$ (with $f(i^+) = \nabla_\phi s_\phi(i^+, c)$ and $\mu = \mathbb{E}_{P_\phi}[\nabla_\phi s_\phi(j,c)]$) is the structural contribution from positive-example normalization and decays at rate $O(1/K)$.

B.0.7 Proof of Proposition 4.4

Proof. Let $f(j) = \nabla_\phi s_\phi(j,c)$, $\mu = \mathbb{E}_{P_\phi}[f(j)]$, $\alpha = w(i^+)$, and $\boldsymbol{a} = w(i^+)(f(i^+) - \mu)$. By Eq. (7), $\hat{g}^{\mathrm{SN}} = -f(i^+) + \hat{\mu}_{\mathrm{SN}}$, where

$$\hat{\mu}_{\mathrm{SN}} = \frac{\alpha\, f(i^+) + \sum_{k=1}^K w(j_k)\, f(j_k)}{\alpha + \sum_{k=1}^K w(j_k)}.$$

Given (c, i^+), $f(i^+)$ is constant, hence $\mathbb{V}_{Q_\theta}[\hat{g}^{\mathrm{SN}} \mid c, i^+] = \mathbb{V}_{Q_\theta}[\hat{\mu}_{\mathrm{SN}} \mid c, i^+]$. By $\mathbb{V}[\boldsymbol{X}] \le \mathbb{E}[\|\boldsymbol{X} - \mu\|^2]$,

$$\mathrm{Tr}(\mathbb{V}_{Q_\theta}[\hat{g}^{\mathrm{SN}} \mid c, i^+]) \le \mathbb{E}_{Q_\theta}\left[\left\| \frac{\boldsymbol{a} + KD_K}{\alpha + KB_K} \right\|^2 \right],$$

where $D_K = \frac{1}{K}\sum_{k=1}^K w(j_k)(f(j_k) - \mu)$ and $B_K = \frac{1}{K}\sum_{k=1}^K w(j_k)$. Note that $\|\hat{\mu}_{\mathrm{SN}} - \mu\| \le 2G$ holds deterministically.

We truncate according to the event $\mathcal{E} = \{B_K \ge 1/2\}$. Dividing numerator and denominator by K: on $\mathcal{E}$ we have $B_K + \alpha/K \ge 1/2$, hence

$$\left\| \frac{D_K + \boldsymbol{a}/K}{B_K + \alpha/K} \right\|^2 \le 4\, \|D_K + \boldsymbol{a}/K\|^2.$$

Using $\mathbb{E}_{Q_\theta}[D_K] = 0$ (each term is zero-mean by importance sampling), the cross term vanishes:

$$\mathbb{E}_{Q_\theta}\|D_K + \boldsymbol{a}/K\|^2 = \frac{M_{\mathrm{SN}}(Q_\theta)}{K} + \frac{\|\boldsymbol{a}\|^2}{K^2},$$

where $M_{\mathrm{SN}}(Q_\theta) = \mathbb{E}_{Q_\theta}[w(j)^2\|f(j)-\mu\|^2] \le 4G^2(1+\chi_c^2)$. On $\mathcal{E}^c$, by the deterministic bound $\|\hat{\mu}_{\mathrm{SN}}-\mu\|^2 \le 4G^2$ and Chebyshev's inequality $\Pr(\mathcal{E}^c) \le 4\chi_c^2/K$. Substituting $\|\boldsymbol{a}\| \le 2G\, w_{\max}$:

$$\begin{aligned}\mathbb{E}_{Q_\theta}\left[\left\|\frac{\boldsymbol{a}+KD_K}{\alpha+KB_K}\right\|^2\right] &\le 4\,\mathbb{E}_{Q_\theta}\|D_K+\boldsymbol{a}/K\|^2 + 4G^2\,\Pr(\mathcal{E}^c)\\ &\le \frac{4M_{\mathrm{SN}}(Q_\theta)}{K} + \frac{4\|\boldsymbol{a}\|^2}{K^2} + \frac{16G^2\chi_c^2}{K}\\ &\le \frac{16G^2(1+\chi_c^2)}{K} + \frac{16G^2 w_{\max}^2}{K^2} + \frac{16G^2\chi_c^2}{K}.\end{aligned}$$

Combining the last two terms yields Eq. (12). □

PROOFS FOR SECTION 5

Lemma B.5 (Distributional-distance upper bound on the candidate-set coverage loss)**.** *For any distributions P, Q and any $\kappa \le |\mathcal{I}|$,*

$$\Delta_\kappa(Q;P) \le 2\,\mathrm{TV}(P,Q) \le \sqrt{2\,D_{\mathrm{KL}}(P\|Q)}. \tag{27}$$

B.0.8 PROOF OF LEMMA B.5

Proof. Write $S^* = S_\kappa(P)$ and $S = S_\kappa(Q)$. The coverage loss admits the identity decomposition

$$\Delta_\kappa(Q;P) = P(S^*) - P(S) = \underbrace{[P(S^*)-Q(S^*)]}_{\text{(I)}} + \underbrace{[Q(S^*)-Q(S)]}_{\text{(II)}} + \underbrace{[Q(S)-P(S)]}_{\text{(III)}}.$$

By the maximality of $S = S_\kappa(Q)$, we have $Q(S) \ge Q(S^*)$, hence (II) ≤ 0. Dropping (II) and taking absolute values of the remaining two terms:

$$\Delta_\kappa(Q;P) \le |P(S^*)-Q(S^*)| + |P(S)-Q(S)|.$$

By the definition of the TV divergence $\mathrm{TV}(P,Q) = \sup_{A\subseteq\mathcal{I}}|P(A)-Q(A)|$, each term is at most $\mathrm{TV}(P,Q)$, hence $\Delta_\kappa(Q;P) \le 2\,\mathrm{TV}(P,Q)$. The second inequality follows from Pinsker's inequality $\mathrm{TV}(P,Q) \le \sqrt{D_{\mathrm{KL}}(P\|Q)/2}$. □

B.0.9 PROOF OF THEOREM 5.1

The statement of Theorem 5.1 (with B_λ and Eq. (15)) appears in Section 5.1.

Proof. We reduce the coverage loss to the TV distance. By Lemma B.5, taking $P = P_c^*$ and $Q = Q_\lambda(\cdot|u)$ for each $(u,\mathbf{z})$,

$$\Delta_\kappa(Q_\lambda(\cdot|u);P_c^*) \le 2\,\mathrm{TV}(P_c^*, Q_\lambda(\cdot|u)).$$

Taking expectation over $(u,\mathbf{z})$ yields $\Delta\tilde{\Gamma}_\kappa(Q_\lambda) \le \mathbb{E}_{u,\mathbf{z}}[2\,\mathrm{TV}(P_c^*, Q_\lambda(\cdot|u))]$. We now derive an upper bound on this TV expectation along two routes.

Route one (not through the teacher). For each $(u,\mathbf{z})$, applying Pinsker's inequality and then Jensen's inequality:

$$\mathbb{E}_{u,\mathbf{z}}[2\,\mathrm{TV}(P_c^*, Q_\lambda(\cdot|u))] \le \sqrt{2\,\mathbb{E}_{u,\mathbf{z}}[D_{\mathrm{KL}}(P_c^*\|Q_\lambda(\cdot|u))]}.$$

Since $Q_\lambda(\cdot|u)$ depends on u alone, for fixed u,

$$\mathbb{E}_{\mathbf{z}|u}[D_{\mathrm{KL}}(P_c^*\|Q_\lambda(\cdot|u))] = -\mathbb{E}_{\mathbf{z}|u}[H(P_c^*)] + H(\bar{P}_u^*, Q_\lambda(\cdot|u)).$$

By the cross-entropy identity and the conditional-entropy identity, rearranging gives

$$\mathbb{E}_{\mathbf{z}|u}[D_{\mathrm{KL}}(P_c^*\|Q_\lambda(\cdot|u))] = D_{\mathrm{KL}}(\bar{P}_u^*\|Q_\lambda(\cdot|u)) + \mathbb{E}_{\mathbf{z}|u}[D_{\mathrm{KL}}(P_c^*\|\bar{P}_u^*)].$$

Taking expectation over u gives

$$\mathbb{E}_{u,\mathbf{z}}[D_{\mathrm{KL}}(P_c^*\|Q_\lambda(\cdot|u))] = e_{\mathrm{ind}}(\lambda) + e_{\mathrm{hidden}}. \tag{28}$$

Substituting back gives route one's upper bound $\sqrt{2(e_{\mathrm{ind}}(\lambda)+e_{\mathrm{hidden}})}$.

Route two (through the teacher). For each $(u,\mathbf{z})$, applying the TV triangle inequality:

$$2\,\mathrm{TV}(P_c^*, Q_\lambda(\cdot|u)) \le 2\,\mathrm{TV}(P_c^*, P_\phi(\cdot|u,\mathbf{z})) + 2\,\mathrm{TV}(P_\phi(\cdot|u,\mathbf{z}), Q_\lambda(\cdot|u)).$$

Taking expectation of the two terms separately and applying Pinsker's and Jensen's inequalities:

$$\mathbb{E}_{u,\mathbf{z}}[2\,\mathrm{TV}(P_c^*, P_\phi(\cdot|u,\mathbf{z}))] \le \sqrt{2e_{\text{teacher}}}, \qquad \mathbb{E}_{u,\mathbf{z}}[2\,\mathrm{TV}(P_\phi(\cdot|u,\mathbf{z}), Q_\lambda(\cdot|u))] \le \sqrt{2e_{\text{dist}}(\lambda)}.$$

Taking the smaller of the two routes yields (15).

Derivation of Eq. (29). Since Q_λ minimizes (13) over $\mathcal{H}_R$, for any $\widetilde{Q} \in \mathcal{H}_R$, $e_{\text{ind}}(\lambda) + \lambda\, e_{\text{dist}}(\lambda) \le e_{\text{ind}}(\widetilde{Q}) + \lambda\, e_{\text{dist}}(\widetilde{Q})$. Using $e_{\text{ind}}(\lambda) \ge 0$ and dividing both sides by $\lambda > 0$ gives

$$e_{\text{dist}}(\lambda) \le e_{\text{dist}}(\widetilde{Q}) + \frac{e_{\text{ind}}(\widetilde{Q})}{\lambda}. \tag{29}$$

Substituting back into (15) gives

$$B_\lambda \le \min\left\{ \sqrt{2\big(e_{\text{hidden}} + e_{\text{ind}}(\lambda)\big)},\ \sqrt{2e_{\text{teacher}}} + \sqrt{2e_{\text{dist}}(\widetilde{Q}) + \frac{2e_{\text{ind}}(\widetilde{Q})}{\lambda}} \right\}. \tag{30}$$

□

B.0.10 Proof of Corollary 5.2

Proof. (i) Since $\sqrt{2e_{\text{teacher}}} \ge B_0$, the second term of (30) is $\ge \sqrt{2e_{\text{teacher}}} \ge B_0$, hence the distillation-route upper bound cannot be strictly smaller than B_0.

(ii) Since $\sqrt{2e_{\text{teacher}}} < B_0$, we have $m_0 > 0$ and $\sqrt{2e_{\text{teacher}}} + \sqrt{2m_0} = B_0$. The condition $\lambda > e_{\text{ind}}(\widetilde{Q})/(m_0 - e_{\text{dist}}(\widetilde{Q}))$ implies

$$e_{\text{dist}}(\widetilde{Q}) + \frac{e_{\text{ind}}(\widetilde{Q})}{\lambda} < m_0,$$

and substituting into the second term of (30):

$$\sqrt{2e_{\text{teacher}}} + \sqrt{2e_{\text{dist}}(\widetilde{Q}) + \frac{2e_{\text{ind}}(\widetilde{Q})}{\lambda}} < \sqrt{2e_{\text{teacher}}} + \sqrt{2m_0} = B_0.$$

By Eq. (30), B_λ is at most the smaller of the two terms on its right-hand side; since the second term is already strictly smaller than B_0, we have $B_\lambda < B_0$. □

B.0.11 Proof of Theorem 5.3

Proof. For each c, $\hat{y}(c) = \arg\max_{j \in S_\kappa(Q_\lambda(\cdot|u))} P_\phi(j|c)$ is fixed, so $\Pr_{y\sim P_c^*}[\hat{y}(c) = y] = P_c^*(\hat{y}(c))$. Let $y^* = \arg\max_{j\in\mathcal{I}} P_c^*(j)$ be the Bayes class and $\tilde{y} = \arg\max_{j\in S(u)} P_c^*(j)$ the class optimal under the true distribution within the candidate set. We decompose the pointwise error:

$$1 - P_c^*(\hat{y}) = \underbrace{\big[1 - P_c^*(y^*)\big]}_{\text{Bayes}} + \underbrace{\big[P_c^*(y^*) - P_c^*(\tilde{y})\big]}_{=:A_c\ \text{(retrieval)}} + \underbrace{\big[P_c^*(\tilde{y}) - P_c^*(\hat{y})\big]}_{=:B_c\ \text{(ranking)}}.$$

Ranking term B_c. We have $\tilde{y}, \hat{y} \in S(u)$, and by the definition of $\hat{y}$ we have $P_\phi(\hat{y}|c) \ge P_\phi(\tilde{y}|c)$, hence

$$B_c = \big[P_c^*(\tilde{y}) - P_\phi(\tilde{y}|c)\big] + \underbrace{\big[P_\phi(\tilde{y}|c) - P_\phi(\hat{y}|c)\big]}_{\le 0} + \big[P_\phi(\hat{y}|c) - P_c^*(\hat{y})\big] \le 2\,\mathrm{TV}\big(P_c^*, P_\phi(\cdot|c)\big).$$

Taking expectation over $(u, \mathbf{z})$, by Pinsker's and Jensen's inequalities:

$$\mathbb{E}_{u,\mathbf{z}}[B_c] \le \mathbb{E}_{u,\mathbf{z}}\big[2\,\mathrm{TV}(P_c^*, P_\phi)\big] \le \sqrt{2\,e_{\text{teacher}}}.$$

Retrieval term A_c. Write $S^* = S_\kappa(P_c^*)$ and $S = S_\kappa(Q_\lambda(\cdot|u))$. Since $\kappa \ge 1$, the Bayes class y^* must lie in S^*. If $y^* \in S$, then $A_c = 0 \le \Delta_\kappa(Q_\lambda(\cdot|u); P_c^*)$. If $y^* \notin S$, let $S' = (S \setminus \{\tilde{y}\}) \cup \{y^*\}$. Since S^* is the set of the κ classes with the largest mass under P_c^*,

$$\Delta_\kappa(Q_\lambda(\cdot|u); P_c^*) = P_c^*(S^*) - P_c^*(S) \ge P_c^*(S') - P_c^*(S) = P_c^*(y^*) - P_c^*(\tilde{y}) = A_c.$$

In both cases $A_c \le \Delta_\kappa(Q_\lambda(\cdot|u); P_c^*)$. Taking expectation over $(u, \mathbf{z})$ and using (14) gives $\mathbb{E}_{u,\mathbf{z}}[A_c] \le \Delta\tilde{\Gamma}_\kappa(Q_\lambda)$.

Bayes term. $\mathbb{E}_{u,\mathbf{z}}[1 - P_c^*(y^*)] = \mathcal{R}^*_{\text{Bayes}}$. Summing the three terms gives (16). □

C Deferred Material for Section 5

C.1 The reference value B_0 and its constituent quantities

All the decisions in Corollary 5.2 depend on the specific value of B_0. Substituting $\lambda = 0$ into (15) and noting that Q_0 is the independent-training solution (at $\lambda = 0$ the objective reduces to minimizing $e_{\mathrm{ind}}(Q)$ alone), B_0 is the smaller of the two routes at Q_0:

$$B_0 = \min\Big\{ \underbrace{\sqrt{2\big(e_{\mathrm{hidden}} + e^\star_{\mathrm{ind}}\big)}}_{B_0^{(\mathrm{I})}\ :\ \text{not through the teacher}} \ , \ \underbrace{\sqrt{2e_{\mathrm{teacher}}} + \sqrt{2e_{\mathrm{dist}}(Q_0)}}_{B_0^{(\mathrm{II})}\ :\ \text{through the teacher}} \Big\}, \tag{31}$$

where $e^\star_{\mathrm{ind}} = \min_{Q\in\mathcal{H}_R} e_{\mathrm{ind}}(Q) = e_{\mathrm{ind}}(Q_0)$ and $e_{\mathrm{dist}}(Q_0) = \mathbb{E}_{u,\mathbf{z}}[D_{\mathrm{KL}}(P_\phi(\cdot|u,\mathbf{z})\|Q_0(\cdot|u))]$; Q_0 minimizes e_{ind}, while $e_{\mathrm{dist}}(Q_0)$ measures its discrepancy from the teacher.

Which route attains B_0 determines the specific forms of the teacher-quality threshold and of the budget m_0. When $B_0^{(\mathrm{I})} \le B_0^{(\mathrm{II})}$, route one dominates, $B_0 = \sqrt{2(e_{\mathrm{hidden}} + e^\star_{\mathrm{ind}})}$, and the threshold $\sqrt{2e_{\mathrm{teacher}}} < B_0$ is equivalent to

$$e_{\mathrm{teacher}} < e_{\mathrm{hidden}} + e^\star_{\mathrm{ind}},$$

i.e., the teacher error is strictly smaller than the sum of the "information unobservable in $\mathbf{z}$" and the "hypothesis-class fitting error"; the budget $\sqrt{2m_0} = \sqrt{2(e_{\mathrm{hidden}} + e^\star_{\mathrm{ind}})} - \sqrt{2e_{\mathrm{teacher}}}$ is route one's margin relative to the teacher error. When $B_0^{(\mathrm{II})} < B_0^{(\mathrm{I})}$, route two dominates, $B_0 = \sqrt{2e_{\mathrm{teacher}}} + \sqrt{2e_{\mathrm{dist}}(Q_0)}$, and the threshold simplifies to $e_{\mathrm{dist}}(Q_0) > 0$, which holds automatically except in the degenerate case where P_ϕ does not depend on $\mathbf{z}$; the budget $\sqrt{2m_0} = \sqrt{2e_{\mathrm{dist}}(Q_0)}$ equals the residual of the independent-training solution imitating the teacher, and any $\widetilde{Q}$ strictly better than Q_0 under the distillation objective satisfies the second threshold.

B_0 is determined by four quantities, whose definitions and dependencies are listed in Table 4; among them, $e_{\mathrm{dist}}(Q_0)$ is the only quantity that simultaneously involves both the candidate generator and the target classifier. When $\mathcal{H}_R$ is expressive enough that $\bar{P}^*_u \in \mathcal{H}_R$, we have $Q_0 = \bar{P}^*_u$, and $e_{\mathrm{dist}}(Q_0)$ further decomposes as

$$\mathbb{E}_u[D_{\mathrm{KL}}(\tilde{P}_u\|\bar{P}^*_u)] + \mathbb{E}_u\mathbb{E}_{\mathbf{z}|u}[D_{\mathrm{KL}}(P_\phi(\cdot|u,\mathbf{z})\|\tilde{P}_u)], \tag{32}$$

where $\tilde{P}_u(\cdot) = \mathbb{E}_{\mathbf{z}|u}[P_\phi(\cdot|u,\mathbf{z})]$ is the marginal of the teacher's conditional distribution over $\mathbf{z}$: the first term measures how far the teacher marginal deviates from the true marginal, and the second measures the teacher conditional distribution's dependence on $\mathbf{z}$, both being nonnegative.

Table 4: The four constituent quantities of B_0 and their dependencies. $Q_0 = \arg\min_{Q\in\mathcal{H}_R} e_{\mathrm{ind}}(Q)$.

Quantity	Definition	Dependencies
e_{hidden}	$I(Y;\mathbf{z}\mid u)$	data
$e^\star_{\mathrm{ind}}$	$\min_{Q\in\mathcal{H}_R}\mathbb{E}_u[D_{\mathrm{KL}}(\bar{P}^*_u\|Q(\cdot\|u))]$	$\mathcal{H}_R,\ \bar{P}^*_u$
e_{teacher}	$\mathbb{E}_{u,\mathbf{z}}[D_{\mathrm{KL}}(P^*_c\|P_\phi)]$	data, P_ϕ
$e_{\mathrm{dist}}(Q_0)$	$\mathbb{E}_{u,\mathbf{z}}[D_{\mathrm{KL}}(P_\phi\|Q_0)]$	$\mathcal{H}_R,\ \bar{P}^*_u,\ P_\phi$

C.2 Simplified forms of the two thresholds

The effective case of Corollary 5.2 requires that two thresholds be passed simultaneously: the teacher-quality threshold $\sqrt{2e_{\mathrm{teacher}}} < B_0$ and the model-class-capacity threshold $\exists\,\widetilde{Q} \in \mathcal{H}_R$ with $e_{\mathrm{dist}}(\widetilde{Q}) < m_0$, where the budget $m_0 = (B_0 - \sqrt{2e_{\mathrm{teacher}}})^2/2$ equals the square of the first threshold's margin. The simplified forms of the two thresholds under the two dominant cases of B_0 are summarized in Table 5. The teacher-quality margin determines m_0: a small margin requires a correspondingly small distillation error, while a larger margin allows a larger approximation error. The margin of the first threshold directly determines the feasible region of the second, so B_0 enters both decisions.

The teacher-quality threshold compares the teacher error with the independent-training reference bound. The capacity threshold concerns how closely the candidate-generator class can approximate the teacher using u. The decomposition (32) separates teacher marginal mismatch from variation due

Table 5: Simplified forms of the two thresholds under the two dominant cases of B_0. $m_0 = (B_0 - \sqrt{2e_{\text{teacher}}})^2/2$.

Threshold	Route one dominates	Route two dominates	Meaning
Teacher quality	$e_{\text{teacher}} < e_{\text{hidden}} + e^{\star}_{\text{ind}}$	$e_{\text{dist}}(Q_0) > 0$	teacher closer to P^* than the independent baseline
Model-class capacity	$\inf_{\widetilde{Q}\in\mathcal{H}_R} e_{\text{dist}}(\widetilde{Q}) < m_0$	$\inf_{\widetilde{Q}\in\mathcal{H}_R} e_{\text{dist}}(\widetilde{Q}) < e_{\text{dist}}(Q_0)$	candidate-generator architecture has enough capacity to imitate the teacher

to the additional information $\mathbf{z}$. When $\mathcal{H}_R$ is sufficiently expressive, this threshold is determined by the "teacher's dependence on $\mathbf{z}$" term in the decomposition (32) (the "teacher marginal deviation" term can be eliminated by the choice of $\widetilde{Q}$).

C.3 Training-side bound

This subsection gives the training-side derivation deferred from Section 5.2, showing how aligning $Q_\theta \to P_\phi$ tightens the training-side term e_{teacher} of Theorem 5.3. On the training side, suppose the training positive examples follow the true class distribution ($i^+ \mid c \sim P_c^*$); by $\mathcal{L}(\phi; c, i^+) = -\log P_\phi(i^+|c)$ (Eq. (2)), we have

$$\bar{\mathcal{L}}(\phi) = \mathbb{E}_{u,\mathbf{z}}\big[H(P_c^*, P_\phi(\cdot|c))\big] = \mathbb{E}_{u,\mathbf{z}}[H(P_c^*)] + \mathbb{E}_{u,\mathbf{z}}[D_{\text{KL}}(P_c^* \| P_\phi(\cdot|c))]\,, \tag{33}$$

denoting the training-side term by $e_{\text{teacher}}(\phi) = \mathbb{E}_{u,\mathbf{z}}[D_{\text{KL}}(P_c^* \| P_\phi(\cdot|c))]$ and the hypothesis-class approximation error by $e_{\text{app}} = \bar{\mathcal{L}}^* - \mathbb{E}_{u,\mathbf{z}}[H(P_c^*)] \geq 0$ (with $\bar{\mathcal{L}}^* = \inf_\phi \bar{\mathcal{L}}(\phi)$). We obtain

$$e_{\text{teacher}}(\phi) = \big(\bar{\mathcal{L}}(\phi) - \bar{\mathcal{L}}^*\big) + e_{\text{app}}, \tag{34}$$

that is, the training-side term equals the sum of the optimization gap $\bar{\mathcal{L}}(\phi) - \bar{\mathcal{L}}^*$ and the irreducible approximation error e_{app}; the latter is determined by the hypothesis-class capacity and is independent of the proposal distribution, so Q influences the training-side term through the optimization gap. This gap is controlled by the convergence analysis: we introduce a standard PL-type gradient-domination assumption Polyak (1963); Karimi et al. (2016); Liu et al. (2022), under which there exists $\mu > 0$ in the region the training trajectory occupies such that

$$\|\nabla\bar{\mathcal{L}}(\phi)\|^2 \geq 2\mu\big(\bar{\mathcal{L}}(\phi) - \bar{\mathcal{L}}^*\big), \tag{35}$$

i.e., $\bar{\mathcal{L}}(\phi) - \bar{\mathcal{L}}^* \leq \frac{1}{2\mu}\|\nabla\bar{\mathcal{L}}(\phi)\|^2$; this condition does not require convexity and only rules out the situation near the trajectory where "the loss is still high but the gradient has already vanished." Taking the expectation over the training randomness $\mathbb{E}_{\text{alg}}$ of Eq. (34) and successively substituting Eq. (35) and Theorem B.3:

$$\begin{aligned} e_{\text{teacher}}(Q) := \mathbb{E}_{\text{alg}}\, e_{\text{teacher}}(\phi_Q) &= e_{\text{app}} + \frac{1}{T}\sum_{t<T}\mathbb{E}_{\text{alg}}\big[\bar{\mathcal{L}}(\phi_t) - \bar{\mathcal{L}}^*\big] \\ &\leq e_{\text{app}} + \frac{1}{2\mu T}\sum_{t<T}\mathbb{E}_{\text{alg}}\|\nabla\bar{\mathcal{L}}(\phi_t)\|^2 \;\leq\; e_{\text{app}} + \frac{1}{2\mu}\,\mathcal{E}_T(Q), \end{aligned} \tag{36}$$

where $\mathcal{E}_T(Q)$ is the average gradient-norm upper bound given by Theorem B.3 under the proposal distribution Q (the right-hand side of Eq. (20)). By Section 4.3 (Proposition 4.4 and Eq. (25)), the candidate-generator-controllable dominant term in $\mathcal{E}_T(Q)$ is controlled by the average divergence $\bar{\chi}^2(Q) = \mathbb{E}_{(c,i^+)}[\chi^2(P_\phi(\cdot|c)\|Q_\theta(\cdot|c))]$ and attains its minimum at $Q_\theta = P_\phi$. Therefore, aligning $Q_\theta \to P_\phi$ lowers the candidate-generator-controllable part of the right-hand side of Eq. (36), thereby providing a guarantee for the training-side term e_{teacher}.

D The Variance-Optimal Proposal and Hard Negatives

A common intuition in the negative-sampling literature is that harder negatives help training. Our variance analysis offers a corresponding characterization: the variance-optimal proposal distribution P_ϕ (Theorem 4.3) assigns larger probability to hard classes (high-scoring classes), so a sampling strategy targeting P_ϕ naturally tends to select harder negatives, in agreement with the empirical observation that "hard negatives are effective." However, P_ϕ is a smooth distribution covering the full class set and is not truncated to the hardest classes. Using only the top-κ classes as negatives

produces a sharp truncation: most of the support of P_ϕ is assigned zero probability, the importance weights tend to infinity, and the second moment $M(Q_\theta)$ is thereby pushed extremely large, so that the variance goes out of control. Mixing in random samples in practice Huang et al. (2023) is precisely a mitigation against truncation—it smooths the proposal distribution, restores full support, and lowers $M(Q_\theta)$.

E Supplementary Experiments

E.1 Additional Setup Details

Table 6: Dataset statistics.

Dataset	#Users	#Items	#Interactions (general)	#Interactions (sequential)
ML-10M	69,584	9,175	8,233,570	1,501,225
Amazon	13,455	8,360	234,524	196,229

E.1.1 Data splits

The general recommendation tasks (MF+DeepFM and XSimGCL+FinalMLP) randomly splits the data into training/validation/test sets in an 80%/10%/10% ratio. The sequential recommendation task (SASRec+DIN) adopts a leave-one-out strategy: for each user, the last interaction is used for testing, the second-to-last for validation, and the rest for training, with the maximum sequence length truncated to 20.

E.1.2 Model architectures

For general recommendation, the retriever is matrix factorization (MF) Koren et al. (2009) with one 64-dimensional embedding each for users and items and inner-product scoring, and the ranker is DeepFM Guo et al. (2017) (an FM component plus a three-layer MLP $[128, 64, 32]$, with LayerNorm + ReLU + Dropout, output clamped to $[-30, 30]$). For sequential recommendation, the retriever is SASRec Kang & McAuley (2018) (a single-layer Transformer with 2-head attention, $d = 64$, $d_{\text{ff}} = 256$, Dropout $= 0.3$, causal mask) and the ranker is the Deep Interest Network (DIN) Zhou et al. (2018) (an attention MLP that aggregates the user's historical behaviors, plus a prediction MLP $[200, 80, 1]$, with LayerNorm + Sigmoid + Dropout).

XSimGCL Yu et al. (2024) uses a user–item bipartite graph built only from training interactions. It applies four LightGCN layers with degree-normalized aggregation and 64-dimensional embeddings; the final representation averages layers 1–4 and is scored by inner product. During training, uniform random vectors are normalized per node, aligned with the signs of the current representation, and scaled by $\epsilon = 0.1$ before being added to each layer's output. Layer 3 and the final representation form the cross-layer contrastive pair. InfoNCE is computed separately for users and positive items after L_2 normalization, using the same node as the positive and other deduplicated nodes of the same type in the batch as negatives, with temperature $\tau = 0.2$ and loss weight $\lambda_{\text{CL}} = 0.2$. Inference uses noise-free propagation.

FinalMLP Mao et al. (2023) concatenates 64-dimensional user and item ID embeddings. Two feature gates use separately learned user and item conditioning embeddings, respectively, with $64 \to 64 \to 128$ MLPs, ReLU hidden activations, and $2 \times \text{sigmoid}$ outputs that weight the concatenated input. The two main MLPs have widths $[400, 400, 400]$ and $[400, 400]$, with ReLU, no normalization, and zero dropout. Each output is split into ten 40-dimensional blocks; the score sums the corresponding bilinear block interactions, the two linear terms, and a bias.

E.1.3 Training hyperparameters

All experiments use Adam, embedding dimension $d = 64$, and $K = 20$ ranker samples per positive. Batch size is 256 for MF+DeepFM and SASRec+DIN and 2048 for XSimGCL+FinalMLP. The latter uses L_2 weight decay 10^{-4}; pretraining and each experimental training stage run for 100 epochs with validation after every epoch. The learning rate is grid-searched uniformly over $\{10^{-2}, 10^{-3}, 10^{-4}\}$, and the best learning rate is used for the final evaluation.

E.1.4 PROPOSAL DISTRIBUTIONS

Table 7 gives the full definitions of the five proposal distributions used in the experiments.

Table 7: The five proposal distributions used in the experiments.

Name	Description
Oracle	The ranker softmax distribution. Requires full computation of Z_ϕ, and is used only in RQ1.
Distill	The distribution of the retriever trained with an additional distillation loss, sampled exactly from Q_θ.
Retriever	The distribution of a retriever trained independently (without distillation).
Uniform	The uniform distribution.
Top-M	A mixture distribution: with probability 50% it samples uniformly from the retriever's top-M ($M = 100$), and with probability 50% it samples uniformly from the full item set.

E.1.5 TRAINING OBJECTIVES AND UPDATE PROTOCOL

The retriever and ranker are pretrained separately; ranker pretraining uses the uniform proposal. For XSimGCL, the base retriever objective combines Bayesian personalized ranking (BPR) and the cross-layer contrastive loss. BPR uses one uniformly sampled item per positive interaction, excluding the user's training positives; the contrastive term uses the batch users and positive items. Distill adds a distillation term with weight $\lambda = 0.1$ to the same base objective for each model group. FinalMLP uses its fusion score in sampled-softmax training, drawing $K = 20$ items independently with replacement from the full-catalog proposal and using the self-normalized gradient in Eq. (7).

For RQ1 and RQ3, the pretrained ranker is frozen. Distill and Retriever start from the same retriever checkpoint and are trained for the same number of steps, differing only in the distillation term. RQ1 then freezes the models and compares proposals for the same target gradient. For RQ2, Distill updates the retriever and ranker separately in every batch. Retriever, Uniform, and Top-M freeze the pretrained retriever and update only the ranker. The Top-M set is determined by the independently trained retriever.

At each distillation step, 100 items are sampled without replacement from the current retriever distribution. Both models' scores are separately softmax-normalized over this candidate set, and KL divergence is computed from the teacher probabilities to the retriever probabilities. Teacher probabilities are detached, so this loss updates only the retriever. During a ranker update, proposal probabilities are treated as constants. The implemented distillation objective is thus a KL surrogate computed on the sampled candidate set.

E.1.6 EVALUATION AND CHECKPOINT SELECTION

At inference, the retriever scores all items, masks training interactions, and selects the top 100; the ranker re-ranks these candidates. We report NDCG@10, Recall@10, and MRR@10, computed per user and averaged equally over evaluation users with positive items. RQ3 reports Recall@100 before re-ranking. XSimGCL pretraining and RQ1/RQ3 select the retriever checkpoint by validation Recall@100; FinalMLP pretraining and RQ2 select by validation end-to-end NDCG@10. Test metrics are evaluated after validation-based selection.

For MF+DeepFM and SASRec+DIN, Table 2 reports means over five seeds, with paired t-tests against the strongest baseline at $p < 0.05$. The XSimGCL+FinalMLP entries include the reported significance marks under the same paired-test criterion. The convergence figures and the supplementary RQ3/RQ4 comparisons below cover MF+DeepFM and SASRec+DIN.

E.1.7 RQ1 VARIANCE AND DIVERGENCE ESTIMATORS

We randomly select $N = 200$ training examples and independently repeat sampling $R = 50$ times per example, with $K = 20$ samples from each proposal per repetition and the ranker fixed. For each

sampled (c_n, i_n^+) we estimate the conditional variance

$$\hat{V}_n = \frac{1}{R-1}\sum_{r=1}^{R}\left\|\hat{g}_{n,r}^{\mathrm{SN}} - \bar{g}_n\right\|^2, \quad \bar{g}_n = \frac{1}{R}\sum_{r=1}^{R}\hat{g}_{n,r}^{\mathrm{SN}},$$

and average over the N samples to obtain the estimator $\widehat{V}_{\mathrm{IS}} = \frac{1}{N}\sum_{n=1}^{N}\hat{V}_n$ of the sampling variance V_{IS}. The empirical χ^2 divergence is $\widehat{\chi^2} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{K}\sum_{k=1}^{K} w_n(j_k)^2 - 1\right]$, where $w_n(j) = P_\phi(j|c_n)/Q(j|c_n)$; it is likewise averaged over the R repeated samplings to reduce estimation noise.

E.2 RQ2: Convergence Speed by Target Accuracy

Figure 2 shows the number of epochs needed by each proposal distribution to reach different target NDCG@10 values on ML-10M. Distill is the fastest on both model groups at medium-to-high target accuracies (MF+DeepFM reaches ≥ 0.24 in only 5.0 epochs, and SASRec+DIN reaches ≥ 0.0979 in only 6.0 epochs), whereas Uniform fails to meet the highest target within the epoch budget on both. This is consistent with the prediction of Theorem B.3: lower variance proposals converge faster and reach higher final accuracy. Figure 1 further reports the per-segment training-loss slope $|d\mathcal{L}/dt|$ of each proposal, whose ordering matches the RQ1 variance ranking: Distill reaches a low slope first on both model groups (e.g. ~ 0.04 vs. Uniform's 0.07–0.13 on MF+DeepFM).

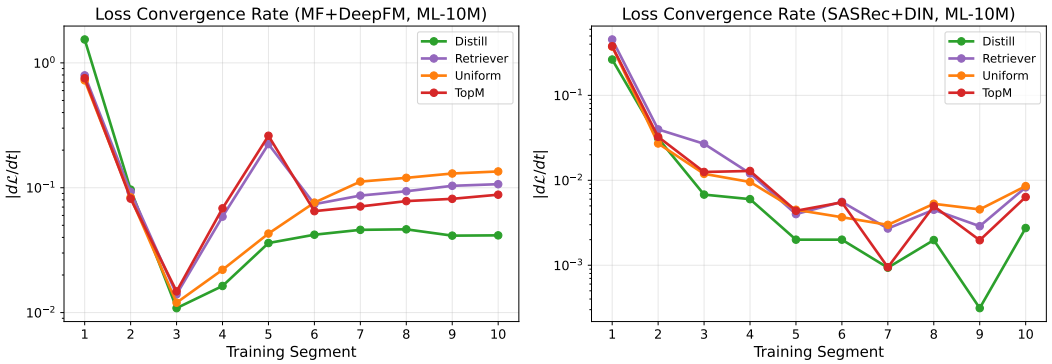
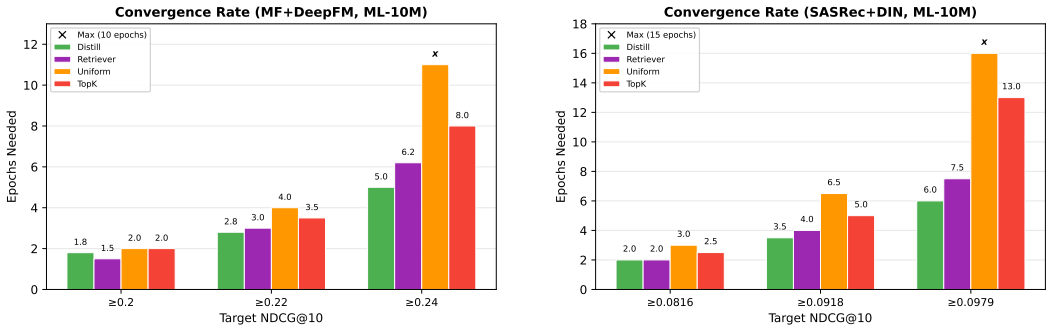


Figure 1: Training-loss convergence rate $|d\mathcal{L}/dt|$ of each proposal on ML-10M: training split into 10 equal segments, the per-segment slope obtained by linear regression on the loss, y on a log scale.



(a) MF+DeepFM (b) SASRec+DIN

Figure 2: Number of epochs needed by each proposal distribution to reach different target NDCG@10 values on ML-10M. $\times$ indicates that the target was not met within the epoch budget.

E.3 The Retriever's Own Retrieval Quality

RQ1 measures sampling variance and divergence, whereas RQ2 measures convergence and end-to-end recommendation accuracy. This section evaluates standalone retrieval quality for MF+DeepFM and SASRec+DIN. The comparison complements the coverage-loss analysis in Section 5.1 with empirical Recall@100 measurements.

E.3.1 METHOD

We take the two retriever groups, Distill and Retriever, trained in Section 6.2 (RQ1); apart from the distillation term, their optimizer, learning rate, number of negatives, and number of training steps are identical. At fixed step intervals on the validation set we compute the retriever's standalone Recall@100—the retriever scores all items in the full set for each user (items already interacted with in the training set are scored $-\infty$), takes the top-100 candidates, and computes the coverage of the true positives among them, without re-ranking. We report the peak Recall@100 reached by each retriever over the course of training (i.e., the best metric used for early stopping).

E.3.2 RESULTS

The Distill retriever attains higher recall in all four sets of experiments. Table 8 summarizes the comparison over the four dataset–model combinations. Across all four sets of experiments, the Distill retriever's peak Recall@100 is strictly higher than that of the Retriever retriever, with relative improvements ranging from $+1.1\%$ to $+8.7\%$.

Table 8: The retriever's standalone peak validation-set Recall@100. Distill is the retriever trained by distillation from a fixed teacher, and Retriever is the retriever obtained from independent training (without distillation).

Dataset	Model	Retriever	Distill	Δ	Rel. improvement
ML-10M	MF+DeepFM	0.4771	**0.4824**	$+0.0053$	$+1.1\%$
	SASRec+DIN	0.2750	**0.2989**	$+0.0239$	$+8.7\%$
Amazon	MF+DeepFM	0.1685	**0.1784**	$+0.0099$	$+5.9\%$
	SASRec+DIN	0.1180	**0.1205**	$+0.0025$	$+2.1\%$

E.3.3 DISCUSSION

With the teacher fixed and the retriever training protocol matched, adding distillation improves peak validation Recall@100 across the four reported combinations. These gains show that the distilled retriever captures more relevant items within the same candidate budget, complementing the conditional coverage-loss bound in Theorem 5.1 and Corollary 5.2.

E.4 CHOICE OF DISTILLATION DIVERGENCE

The theoretical framework of Section 5.1 uses $D_{\mathrm{KL}}(P_\phi \| Q_\theta)$ as the distillation objective that drives $Q_\theta \to P_\phi$, yet the theory itself (Proposition 4.4) controls the variance bound of the self-normalized estimator by $\chi^2(P_\phi \| Q_\theta)$. A natural question is: does using χ^2 directly as the distillation divergence achieve better training results than KL? This section trains two groups of distillation retrievers in parallel for comparison.

E.4.1 METHOD

For each dataset–model combination, we fix the retriever architecture and all training hyperparameters (identical to Section 6.2), varying only the distillation objective: (a) Distill-KL uses $\lambda\, \widehat{D_{\mathrm{KL}}}(P_\phi \| Q_\theta)$, i.e., the Distill of Section 6.2/Section 6.3; (b) Distill-χ^2 replaces the distillation term with $\lambda\, \widehat{\chi^2}(P_\phi \| Q_\theta)$, with the rest of the training procedure and hyperparameters kept exactly the same. The two are compared on sampling variance, χ^2 divergence (RQ1 protocol), and the ranker's final accuracy (RQ2 protocol).

E.4.2 RESULTS

Finding 1: the KL and χ^2 distillation divergences trade wins across the four combinations, with gaps all on the order of ± 0.01. In Table 9, χ^2 distillation attains lower variance on the two Amazon combinations (consistent with its objective of directly minimizing χ^2), but on ML-10M KL distillation has the lower variance instead; the χ^2 divergence column is likewise a mixed distribution—wins are jointly determined by the dataset and the architecture. In Table 10, the gap between KL

Table 9: Comparison of distillation KL and distillation χ^2 on sampling variance and χ^2 divergence. For each dataset–model group, lower variance and χ^2 divergence are better, and the best value within each combination is bolded.

Dataset	Model	Distill. divergence	$\widehat{V}_{\mathrm{IS}}$	$\widehat{\chi^2}(P_\phi \| Q_\theta)$
ML-10M	MF+DeepFM	KL	**17.21**	0.68
		χ^2	18.21	**0.58**
	SASRec+DIN	KL	**19.75**	**1.04**
		χ^2	24.09	1.40
Amazon	MF+DeepFM	KL	177.89	**1.06**
		χ^2	**145.65**	1.35
	SASRec+DIN	KL	34.09	1.44
		χ^2	**30.18**	**1.06**

Table 10: Final test-set metrics of the rankers trained with distillation KL and distillation χ^2. $\Delta = \chi^2 - \mathrm{KL}$.

Dataset	Model	Distill. divergence	NDCG@10	Recall@10	MRR@10
ML-10M	MF+DeepFM	KL	0.2495	0.2346	0.4037
		χ^2	**0.2501**	**0.2357**	**0.4055**
		Δ	$+0.0006$	$+0.0011$	$+0.0018$
	SASRec+DIN	KL	**0.1020**	**0.1913**	**0.1012**
		χ^2	0.0978	0.1830	0.0993
		Δ	-0.0042	-0.0083	-0.0019
Amazon	MF+DeepFM	KL	**0.0220**	**0.0377**	**0.0274**
		χ^2	0.0189	0.0356	0.0216
		Δ	-0.0031	-0.0021	-0.0058
	SASRec+DIN	KL	0.0235	0.0389	0.0263
		χ^2	**0.0269**	**0.0424**	**0.0307**
		Δ	$+0.0034$	$+0.0035$	$+0.0044$

distillation and χ^2 distillation in the ranker's final accuracy is all on the order of ± 0.01, with χ^2 distillation slightly ahead on ML-10M/MF+DeepFM and Amazon/SASRec+DIN, and KL distillation slightly ahead on the other two.

E.4.3 DISCUSSION

The fact that neither distillation divergence dominates the other in performance is consistent with the claim about the optimal proposal distribution in Section 4.2: when the retriever cannot obtain the ranker's per-item gradient norms, $Q_\theta = P_\phi$ is the minimax-optimal proposal distribution, while the specific choice of divergence (KL or χ^2) that drives $Q_\theta \to P_\phi$ is an implementation-level matter that, with the sample budget and number of training steps fixed, does not fundamentally change the final result. The measured gap between the two divergences is in all cases on the order of ± 0.01, on the same scale as the gap between Distill and baselines such as Uniform in RQ2, indicating that the form of the divergence plays only a secondary role relative to the distillation mechanism itself. From the standpoint of computational stability, the gradient estimate of the KL divergence is numerically more robust (it does not involve the square of the importance weights) and is consistent with the default implementation in the existing distillation literature; therefore the main tables of Section 6.2 and Section 6.3 of this work all adopt KL distillation as the default Distill implementation.