LEARNING MULTI-CLASS CLASSIFIERS FROM INCOMPLETELY REVEALED SAME-CLASS RELATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study multi-class classification from unlabeled instances, revealed same-class pairs, and an aggregate class prior, without instance-level class labels. The supervision consists of revealed same-class pairs, where each pair indicates that its two instances belong to the same class. We propose PRIME (Prior-Regularized Incomplete-Similarity Multi-Class Estimation), which constructs training pairs by sampling unlabeled instances and models the probability that a pair is revealed through the same-class similarity induced by the latent classifier under a class-independent one-sided reveal process. A permutation-invariant class-prior regularizer matches the predicted class marginal to the class prior up to permutation and constrains marginal latent-class usage. We prove that, when the ground-truth same-class relation is realizable by the induced pairwise function class, every population minimizer recovers that relation almost surely. When the latent output dimension equals the number of ground-truth classes, relation recovery determines the classifier up to a global class permutation. We further derive a uniform finite-sample generalization bound and establish permutation invariance and marginal-collapse exclusion. Experiments on five image benchmarks show that PRIME improves instance-level classification and pairwise relation recovery over observation-naive training, with larger gains when fewer same-class relations are revealed.

1 INTRODUCTION

We consider instance-level multi-class classification from an unlabeled instance pool, a set of revealed same-class pairs, and an aggregate class prior, without access to instance-level class labels. Each revealed pair indicates that its two instances belong to the same semantic class, but does not reveal the class identity. The goal is to recover an instance-level latent classifier from this relational and aggregate supervision.

The central difficulty is that the revealed pairs constitute only a subset of the underlying same-class relations. We model this observation process by assuming that each same-class relation is revealed independently with a fixed probability, whereas different-class pairs are never revealed. Consequently, an unrevealed pair cannot be interpreted as a different-class pair. We therefore distinguish the latent same-class relation from its reveal status and explicitly model how latent relations generate the observed pairwise supervision. Figure 1 summarizes how the revealed pairs are modeled as observations of a latent same-class relation and used to recover the instance-level classifier.

To learn under this observation model, we propose PRIME (*Prior-Regularized Incomplete-Similarity Multi-Class Estimation*). PRIME maps each instance to a probability vector over latent classes and uses the inner product between two output vectors as their same-class similarity. This similarity parameterizes the probability that a sampled pair is revealed, yielding an incomplete-similarity likelihood consistent with the one-sided reveal process. PRIME further incorporates the aggregate class prior by matching the predicted class marginal to the prior up to permutation, without imposing a predefined correspondence between latent coordinates and semantic classes.

Our analysis shows that, when the ground-truth same-class relation is realizable by the classifier-induced pairwise function class, every population minimizer of the incomplete-similarity risk recovers that relation almost surely. When the latent output dimension matches the number of ground-truth classes, relation recovery further determines the classifier up to a global class permutation. We

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

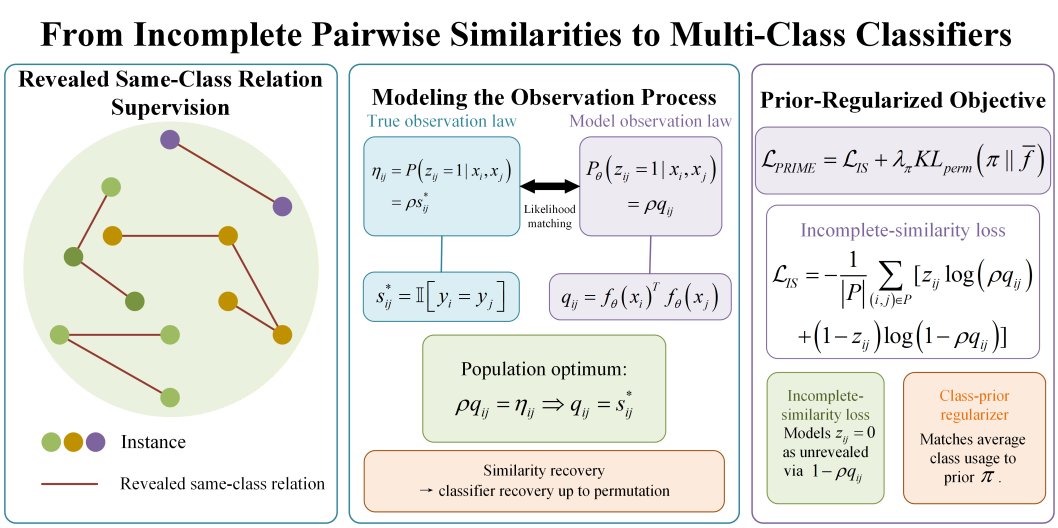


Figure 1: Overview of PRIME. Revealed same-class pairs are treated as stochastic observations of a latent same-class relation under a one-sided reveal process. PRIME models reveal probabilities using classifier-induced similarities and combines the resulting incomplete-similarity likelihood with a permutation-invariant class-prior regularizer. Recovering the latent relation yields instance-level classification up to a global class permutation.

also derive a uniform finite-sample generalization bound and show that the class-prior regularizer is permutation invariant and rules out solutions that leave latent coordinates unused when the target prior has full support.

The scale of the incomplete-similarity excess risk also varies with the reveal rate, which becomes particularly relevant in sparse-reveal regimes. We therefore parameterize the prior coefficient in proportion to the reveal rate, so that its scale remains comparable to that of the relational excess risk as the reveal rate changes. Experiments on five image benchmarks evaluate both instance-level classification and same-class relation recovery across different reveal rates, including sparse-relation regimes and misspecified observation models.

2 Related Work

Pairwise and relational supervision. Pairwise supervision has been studied for classification from similar pairs (Bao et al., 2018), similarity and dissimilarity labels (Hsu et al., 2019; Shimada et al., 2021), and similarity-confidence information (Cao et al., 2021). Related formulations also use pairwise confidence differences as weak supervision (Dai et al., 2025). In constrained clustering, must-link and cannot-link constraints are used to guide the learned partition (Wagstaff et al., 2001; Basu et al., 2004). In these settings, the observed pairwise information directly specifies or quantifies the relation between the two instances. In PRIME, by contrast, the observed variable indicates whether an underlying same-class relation has been revealed: only a subset of same-class relations is observed, while the complete relation structure remains latent. This observation model differs from noisy pairwise learning, where pairwise relation labels are observed but may be corrupted (Wu et al., 2022).

One-sided observation models. Positive-unlabeled (PU) learning studies binary classification from observed positive and unlabeled instances under a one-sided labeling process (Elkan & Noto, 2008; Bekker & Davis, 2020). A central line of work constructs unbiased or non-negative risks from this observation mechanism (du Plessis et al., 2014; Kiryo et al., 2017; Li et al., 2025). PRIME adopts a related one-sided observation model at the pair level: the observed variable records whether a latent same-class relation is revealed, rather than whether an individual instance carries a binary label. The resulting learning problem is therefore to recover an instance-level multi-class classifier from incomplete relational observations.

Aggregate class supervision. Learning from label proportions (LLP) recovers instance-level predictors from bags annotated with aggregate class proportions (Scott & Zhang, 2020; Busa-Fekete et al., 2023; Havaldar et al., 2024). Recent variants further study robust and proportional-value formulations (Chen et al., 2025; Ma et al., 2026). Aggregate supervision has also been considered beyond the standard LLP setting, including more general aggregate observations (Wei et al., 2023) and multiple unlabeled datasets with known class priors (Lu et al., 2021; Wei et al., 2024). PRIME instead combines a single aggregate class prior with incompletely revealed same-class relations. Since the latent output coordinates have no predefined correspondence with the semantic classes, the predicted class marginal is matched to the prior only up to permutation.

3 Proposed Method

We present **PRIME**, a **P**rior-**R**egularized **I**ncomplete-Similarity **M**ulti-Class **E**stimation approach for learning an instance-level latent classifier from unlabeled instances with incompletely revealed same-class relations.

3.1 Problem Formulation and Observation Model

Let $\mathcal{U} = \{x_i\}_{i=1}^n$ be an unlabeled training set, where each $x_i \in \mathcal{X}$ has an unknown class label $y_i \in \{1, \dots, K\}$. We define the latent same-class relation between two instances as

$$s_{ij}^\star = s^\star(x_i, x_j) = \mathbb{I}[y_i = y_j].$$

The class labels and the complete relation structure are both unavailable during training.

Let $\mathcal{S}^+$ denote the set of revealed same-class pairs. For $i < j$, the observation indicator is defined as $z_{ij} = \mathbb{I}[(i, j) \in \mathcal{S}^+]$. Therefore, z_{ij} indicates whether the latent same-class relation is revealed rather than directly observing $s_{ij}^\star$.

The objective is to recover an instance-level latent classifier $f_\theta : \mathcal{X} \to \Delta^{K-1}$, where $f_i = f_\theta(x_i)$. PRIME represents the latent relation through the classifier-induced similarity $q_{ij} = f_i^\top f_j \in [0, 1]$.

We assume a class-independent one-sided reveal process, where each underlying same-class relation is revealed independently with probability $\rho \in (0, 1]$, while different-class pairs are never revealed. The observation mechanism is therefore characterized by

$$P(Z = 1 \mid X = x, X' = x') = \rho s^\star(x, x'). \tag{1}$$

Replacing the latent relation with the classifier-induced similarity gives the model probability

$$P_\theta(z_{ij} = 1 \mid x_i, x_j) = \rho q_{ij} = \rho f_i^\top f_j.$$

The complementary probability is $P_\theta(z_{ij} = 0 \mid x_i, x_j) = 1 - \rho q_{ij}$. Hence, $z_{ij} = 0$ represents an unrevealed relation rather than an observed different-class relation.

Let $\pi = [\pi_1, \dots, \pi_K]^\top$ denote the aggregate class prior, where $\pi_k = P(Y = k)$. When ρ is unavailable, it can be estimated from the marginal reveal frequency using π; the corresponding estimator is provided in Appendix A.2.

3.2 PRIME Objective

Given a set of sampled training pairs $\mathcal{P}$, PRIME maximizes the likelihood induced by the one-sided reveal process. Specifically, the incomplete-similarity negative log-likelihood is defined as

$$\mathcal{L}_{\text{IS}} = -\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[z_{ij} \log(\rho q_{ij}) + (1 - z_{ij}) \log(1 - \rho q_{ij}) \right].$$

To incorporate the aggregate class prior, we estimate the predicted class marginal from an instance mini-batch $\mathcal{B}_{\text{prior}}$ as

$$\bar{f} = \frac{1}{|\mathcal{B}_{\text{prior}}|} \sum_{x\in\mathcal{B}_{\text{prior}}} f_\theta(x).$$

Since the latent output coordinates do not have predefined semantic correspondence with the ground-truth classes, the prior is matched only up to a permutation. Let $\mathcal{P}_K$ denote the set of $K \times K$ permutation matrices. We define the permutation-invariant prior regularizer as

$$\mathcal{L}_{\text{prior}} = \min_{P \in \mathcal{P}_K} D_{\text{KL}}\big(\pi \| P\bar{f}\big)\,. \tag{2}$$

The PRIME objective is

$$\mathcal{L}_{\text{PRIME}} = \mathcal{L}_{\text{IS}} + \lambda_\pi \mathcal{L}_{\text{prior}}, \qquad \lambda_\pi \geq 0.$$

The first term corresponds to the likelihood induced by the one-sided reveal process, whereas the second term aligns the predicted class marginal with the given prior without fixing the semantic identity of latent output coordinates.

The minimizing permutation in (2) is obtained by linear assignment. PRIME is optimized by stochastic gradient descent and predicts $\widehat{y}(x) = \arg\max_k f_\theta(x)_k$ at inference. Numerical stabilization, reveal-rate estimation, pair-sampling strategies, optimization details, and the entropy-regularized variant PRIME+Ent are described in Appendix A.

4 Theoretical Analysis

We first characterize the population behavior of the incomplete-similarity objective. We then provide a finite-sample generalization bound, analyze the permutation-invariant class-prior regularizer, and study how the reveal rate affects the relative scale of the two objective terms. Proofs and additional results are provided in appendices.

4.1 Population Recovery

Let $X, X' \overset{\text{i.i.d.}}{\sim} P_X$, let $y : \mathcal{X} \to \{1, \ldots, K\}$ denote the ground-truth labeling function, and let $\mathcal{F}$ be a class of classifiers $f : \mathcal{X} \to \Delta^{K-1}$. For $f \in \mathcal{F}$, denote the induced similarity by $g_f(x, x') = f(x)^\top f(x')$. Under the one-sided reveal process defined in (1), the classifier predicts the reveal probability through $\rho g_f(x, x')$. The corresponding population incomplete-similarity risk is

$$R^0_{\text{IS}}(f) = \mathbb{E}\left[-Z \log(\rho g_f(X, X')) - (1 - Z) \log(1 - \rho g_f(X, X'))\right].$$

Proposition 1 (Relation recovery)**.** *Assume that the ground-truth same-class relation is realizable by the classifier-induced function class*

$$\mathcal{G} = \left\{(x, x') \mapsto f(x)^\top f(x') : f \in \mathcal{F}\right\},$$

i.e., $s^\star \in \mathcal{G}$. Then every population minimizer $f^\star$ of R^0_{IS} satisfies

$$g_{f^\star}(x, x') = s^\star(x, x')$$

almost surely.

Corollary 1 (Classification recovery up to permutation)**.** *Assume that every ground-truth class has positive probability and that the number of latent output dimensions equals the number of ground-truth classes. If*

$$g_f(x, x') = s^\star(x, x')$$

almost surely, then there exists a permutation σ of $\{1, \ldots, K\}$ such that

$$f(x) = e_{\sigma(y(x))}$$

almost surely.

Therefore, the incomplete-similarity objective identifies the latent same-class relation and recovers the instance-level classifier up to a global permutation of class coordinates.

4.2 Finite-Sample Generalization

For $0 < \epsilon < 1/2$, let

$$\mathrm{clip}_\epsilon(t) = \min\{\max\{t, \epsilon\}, 1-\epsilon\},$$

and define the clipped pairwise loss

$$\ell_{\rho,\epsilon}(g, z) = -z \log \mathrm{clip}_\epsilon(\rho g) - (1-z) \log\left[1 - \mathrm{clip}_\epsilon(\rho g)\right].$$

For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_X$, define the empirical risk

$$\widehat{R}^{\epsilon}_{\mathrm{IS},n}(f) = \binom{n}{2}^{-1} \sum_{i<j} \ell_{\rho,\epsilon}(g_f(X_i, X_j), Z_{ij}),$$

and let $R^{\epsilon}_{\mathrm{IS}}(f)$ denote the corresponding population risk.

Theorem 1 (Finite-sample generalization). *Suppose that, conditional on $X_1, \ldots, X_n$, the reveal variables $\{Z_{ij}\}_{i<j}$ are generated independently according to the one-sided reveal process. Let*

$$m = \left\lfloor \frac{n}{2} \right\rfloor, \qquad M_\epsilon = \log \frac{1}{\epsilon}.$$

Then, with probability at least $1-\delta$,

$$\sup_{f \in \mathcal{F}} \left| R^{\epsilon}_{\mathrm{IS}}(f) - \widehat{R}^{\epsilon}_{\mathrm{IS},n}(f) \right| \le 2\frac{\rho}{\epsilon} \mathfrak{R}_m(\mathcal{G}) + M_\epsilon \sqrt{\frac{3\log(1/\delta)}{n}},$$

where $\mathfrak{R}_m(\mathcal{G})$ is evaluated on m independent draws from $P_X \otimes P_X$.

Generalization analysis for pairwise learning requires accounting for the dependence induced by shared instances, and prior work has studied this structure through stability-based analyses (Lei et al., 2020; Yang et al., 2021). Theorem 1 instead establishes a uniform bound for the incomplete-similarity risk by exploiting the pairwise structure directly. The corresponding loss-class formulation and excess-risk bound are given in Appendix B.6.

4.3 Properties of the Class-Prior Regularizer

Let $\mu_f = \mathbb{E}_{X \sim P_X}[f(X)]$ denote the population predicted marginal. Define the permutation-invariant prior discrepancy as

$$\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_f) = \min_{P \in \mathcal{P}_K} D_{\mathrm{KL}}(\pi \| P\mu_f).$$

Proposition 2 (Permutation invariance). *For any $Q \in \mathcal{P}_K$, let $\widetilde{f}(x) = Qf(x)$. Then*

$$g_{\widetilde{f}}(x, x') = g_f(x, x')$$

and

$$\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_{\widetilde{f}}) = \mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_f).$$

Hence the population PRIME objective is invariant under global permutations of the latent output coordinates.

Proposition 3 (Marginal-support property). *Assume that the target prior equals the true class prior and has full support. If f recovers the ground-truth classes up to permutation, then*

$$\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_f) = 0.$$

Conversely, if $\mu_{f,r} = 0$ for any output coordinate r, then

$$\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_f) = +\infty.$$

In particular, a classifier assigning all instances to a single latent coordinate, $f(x) = e_c$ almost surely, has infinite prior discrepancy when $K > 1$.

The population PRIME objective is

$$R_{\mathrm{PRIME}}(f) = R^0_{\mathrm{IS}}(f) + \lambda_\pi \,\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_f).$$

Theorem 2 (Population recovery of PRIME). *Assume the conditions of Proposition 1 and Corollary 1, and suppose that the target prior equals the true class prior and has full support. Then, for any $\lambda_\pi \geq 0$, every population minimizer $f^\star$ of R_{PRIME} satisfies*

$$g_{f^\star}(x, x') = s^\star(x, x')$$

almost surely, and there exists a permutation σ such that

$$f^\star(x) = e_{\sigma(y(x))}$$

almost surely. Moreover,

$$\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_{f^\star}) = 0.$$

The prior term therefore preserves the population solutions identified by the incomplete-similarity objective while matching the aggregate class prior.

4.4 Reveal-Rate Normalization

We finally study how the reveal rate affects the relative scale between the incomplete-similarity term and the class-prior regularizer. Although the population minimizer is unchanged for any nonnegative regularization weight, the relative magnitude of the two terms influences their contribution during finite-sample optimization.

Denote the population incomplete-similarity risk with reveal rate ρ by $R^0_{\mathrm{IS},\rho}$, and define the optimal relational risk and the corresponding excess risk as

$$R^{0,\star}_{\mathrm{IS},\rho} = \inf_{f \in \mathcal{F}} R^0_{\mathrm{IS},\rho}(f), \qquad \Delta_\rho(f) = R^0_{\mathrm{IS},\rho}(f) - R^{0,\star}_{\mathrm{IS},\rho}.$$

For $s \in \{0, 1\}$ and $g \in [0, 1]$, define the conditional excess risk

$$d_\rho(s, g) = D_{\mathrm{KL}}\left(\mathrm{Bern}(\rho s) \| \,\mathrm{Bern}(\rho g)\right).$$

Under the realizability condition of Proposition 1, the excess relational risk can be written as

$$\Delta_\rho(f) = \mathbb{E}\left[d_\rho\big(s^\star(X, X'), g_f(X, X')\big)\right].$$

Proposition 4 (Reveal-rate normalization). *For every $g \in [0, 1]$,*

$$\lim_{\rho \downarrow 0} \frac{d_\rho(0, g)}{\rho} = g.$$

For every $g \in (0, 1]$,

$$\lim_{\rho \downarrow 0} \frac{d_\rho(1, g)}{\rho} = -\log g - (1 - g),$$

with the limit interpreted as $+\infty$ when $g = 0$. Consequently, whenever

$$\mathbb{E}\left[s^\star(X, X')\big(-\log g_f(X, X')\big) + \big(1 - s^\star(X, X')\big) g_f(X, X')\right] < \infty,$$

the normalized excess risk $\Delta_\rho(f)/\rho$ admits a finite limit as $\rho \downarrow 0$.

Proposition 4 shows that the model-dependent part of the relational objective has a first-order dependence on the reveal rate. Therefore, to maintain a consistent relative scale between the relational excess risk and the prior discrepancy, we parameterize the prior coefficient as $\lambda_\pi = \beta\rho, \beta \geq 0$. Here, β controls the relative strength of the class-prior regularizer. This parameterization provides a reveal-rate-aware scaling rule rather than an optimal choice of λ_π for all possible settings.

Under this parameterization, up to the additive constant $R^{0,\star}_{\mathrm{IS},\rho}$,

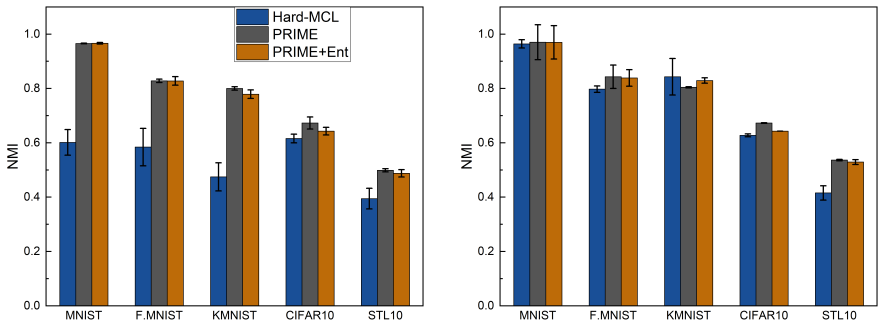
$$\frac{R_{\mathrm{PRIME}}(f) - R^{0,\star}_{\mathrm{IS},\rho}}{\rho} = \frac{\Delta_\rho(f)}{\rho} + \beta \,\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_f).$$

Hence, fixing β preserves the relative scale between the normalized relational objective and the class-prior regularizer as the reveal rate changes.

If the reveal rate is unknown, ρ can be replaced by its estimate $\hat{\rho}$ obtained from the marginal reveal frequency. In particular, for a reference configuration $(\rho_{\mathrm{ref}}, \lambda_{\mathrm{ref}})$, the corresponding reveal-rate-aware scaling is $\lambda_\pi(\rho) = \lambda_{\mathrm{ref}} \frac{\rho}{\rho_{\mathrm{ref}}}$.

Table 1: ACC-Hungarian (%) under varying reveal rates ρ. Results are reported as mean $\pm$ standard deviation over repeated runs.

Dataset	ρ	Hard-Neg	PRIME	PRIME+Ent	KCL	DCC	CC
MNIST	0.2	57.50 ± 9.31	98.79 ± 0.05	$\mathbf{98.84 \pm 0.11}$	54.73 ± 6.71	11.35 ± 0.00	98.74 ± 0.06
MNIST	0.5	98.67 ± 0.24	98.99 ± 0.05	98.97 ± 0.01	$\mathbf{99.13 \pm 0.11}$	11.46 ± 0.20	98.74 ± 0.06
FashionMNIST	0.2	50.23 ± 1.30	$\mathbf{90.25 \pm 0.40}$	90.15 ± 1.04	45.61 ± 2.92	10.00 ± 0.00	57.40 ± 6.83
FashionMNIST	0.5	76.63 ± 6.95	$\mathbf{91.32 \pm 0.29}$	90.86 ± 0.52	91.01 ± 0.14	10.00 ± 0.00	57.40 ± 6.83
KMNIST	0.2	46.18 ± 4.50	$\mathbf{90.75 \pm 1.53}$	89.53 ± 0.82	61.57 ± 8.57	10.66 ± 1.14	48.64 ± 1.06
KMNIST	0.5	62.88 ± 0.56	90.94 ± 2.63	92.49 ± 1.66	$\mathbf{95.71 \pm 0.16}$	10.00 ± 0.00	48.64 ± 1.06
CIFAR10	0.2	42.31 ± 0.14	$\mathbf{81.22 \pm 0.28}$	79.32 ± 1.08	16.09 ± 0.56	10.00 ± 0.00	56.40 ± 2.84
CIFAR10	0.5	52.25 ± 3.15	$\mathbf{89.33 \pm 0.27}$	89.29 ± 0.48	22.32 ± 3.14	10.00 ± 0.00	56.40 ± 2.84
STL10	0.2	58.73 ± 0.93	$\mathbf{89.08 \pm 0.12}$	84.32 ± 0.57	18.08 ± 2.28	10.00 ± 0.00	35.48 ± 0.88
STL10	0.5	69.65 ± 2.84	$\mathbf{98.50 \pm 0.71}$	96.35 ± 0.35	22.81 ± 1.71	10.00 ± 0.00	35.48 ± 0.88



(a) NMI, $\rho = 0.2$ (b) NMI, $\rho = 0.5$

Figure 2: Clustering structure recovery under incomplete-similarity supervision with varying reveal rates ρ. NMI is reported as mean $\pm$ standard deviation over repeated runs.

5 Experiments

5.1 Experimental Setup

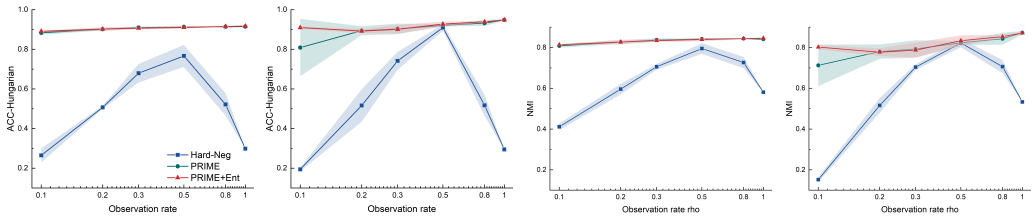
Datasets. We evaluate PRIME on five image classification benchmarks: MNIST (LeCun et al., 1998), FashionMNIST (Xiao et al., 2017), KMNIST (Clanuwat et al., 2018), CIFAR10 (Krizhevsky, 2009), and STL10 (Coates et al., 2011). For MNIST, FashionMNIST, and KMNIST, we use a convolutional backbone. For CIFAR10 and STL10, we use ResNet-18 (He et al., 2016) with a modified first convolutional layer and without the initial max-pooling operation. STL10 images are resized to 32×32.

Baselines. We compare PRIME with an observation-naive pairwise baseline and three clustering methods. Hard-Neg treats the unrevealed indicator $z_{ij} = 0$ as a different-class observation and optimizes a standard binary pairwise likelihood. PRIME+Ent is an entropy-regularized variant of PRIME, following the general principle of entropy minimization (Grandvalet & Bengio, 2004).

KCL (Hsu & Kira, 2016) learns cluster assignments from pairwise constraints, while DCC (Zhang et al., 2021) incorporates pairwise constraints into deep representation learning and clustering. For both methods, we provide revealed same-class relations as similarity or must-link constraints and do not construct dissimilarity or cannot-link constraints from unrevealed pairs. CC (Li et al., 2021) is an unsupervised contrastive clustering method and uses only the unlabeled instances.

Incomplete-similarity setting. Training data consist of unlabeled instances, revealed same-class relations, and aggregate class-prior information. Ground-truth labels are used only to construct the controlled experimental setting and are never provided as instance-level training targets. Unless otherwise specified, we use reveal rates $\rho \in \{0.2, 0.5\}$.

Evaluation metrics. Instance-level recovery is evaluated using ACC-Hungarian and normalized mutual information (NMI), following standard deep-clustering evaluation protocols (Xie et al., 2016). ACC-Hungarian aligns the latent output coordinates with the ground-truth classes using the Hungarian algorithm. Pairwise relation recovery is evaluated using Pair AUC, Pair AP (Davis



FashionMNIST KMNIST FashionMNIST KMNIST

Figure 3: Performance under varying incomplete-similarity observation rates ρ on FashionMNIST and KMNIST. Panels (a) and (b) report ACC-Hungarian, while panels (c) and (d) report NMI.

Table 2: Same-class relation recovery on FashionMNIST under incomplete-similarity supervision. Results are reported as mean $\pm$ standard deviation over three seeds. Pair AUC and Pair AP evaluate the learned similarity score $q_{ij} = f_\theta(x_i)^\top f_\theta(x_j)$, and the similarity gap is defined as $q_{\text{same}} - q_{\text{diff}}$.

ρ	Method	Pair AUC	Pair AP	$q_{\text{same}} - q_{\text{diff}}$
0.2	Hard-Neg	0.934 ± 0.002	0.569 ± 0.018	0.122 ± 0.005
	PRIME	0.987 ± 0.002	$\mathbf{0.917 \pm 0.009}$	0.702 ± 0.025
	PRIME+Ent	$\mathbf{0.987 \pm 0.003}$	0.917 ± 0.013	$\mathbf{0.750 \pm 0.027}$
0.5	Hard-Neg	0.976 ± 0.003	0.806 ± 0.046	0.400 ± 0.015
	PRIME	$\mathbf{0.989 \pm 0.000}$	$\mathbf{0.930 \pm 0.003}$	0.764 ± 0.011
	PRIME+Ent	0.989 ± 0.001	0.927 ± 0.005	$\mathbf{0.768 \pm 0.019}$

& Goadrich, 2006), and the similarity gap $q_{\text{same}} - q_{\text{diff}}$, where q_{same} and q_{diff} denote the mean similarities $q_{ij} = f_\theta(x_i)^\top f_\theta(x_j)$ over same-class and different-class pairs, respectively.

Implementation. All methods use the same backbone and training protocol for controlled comparison. Unless otherwise specified, results are averaged over three random seeds and reported as mean $\pm$ standard deviation. Additional experimental and implementation details are provided in Appendix A.6.

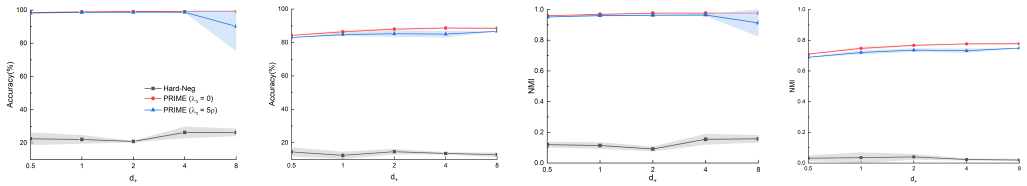
5.2 Results

Overall Recovery Performance. We first evaluate recovery of the instance-level classes and latent same-class relations from incomplete-similarity supervision. Table 1 reports ACC-Hungarian on five datasets under $\rho \in \{0.2, 0.5\}$. PRIME achieves strong performance across the evaluated datasets, with particularly large gains over Hard-Neg in the lower-reveal regime. For example, at $\rho = 0.2$, PRIME improves ACC-Hungarian from 42.31% to 81.22% on CIFAR10 and from 58.73% to 89.08% on STL10. When the reveal rate increases to 0.5, PRIME further reaches 89.33% and 98.50% on the two datasets, respectively.

Among the clustering baselines, KCL is competitive on MNIST and KMNIST when $\rho = 0.5$, but its performance decreases on several datasets when fewer same-class relations are revealed. DCC performs poorly when provided only with the revealed same-class relations as must-link constraints. CC, which uses only the unlabeled instances, provides an unsupervised reference and performs substantially below PRIME on CIFAR10 and STL10.

Figure 2 reports NMI of the recovered latent partitions. The results follow the same overall pattern as ACC-Hungarian: PRIME and PRIME+Ent perform strongly under both reveal rates, with larger advantages over Hard-Neg when $\rho = 0.2$. Thus, the improvement is also reflected in the recovered partition structure rather than only in the Hungarian-aligned class assignments.

Table 2 directly evaluates recovery of the latent same-class relation on FashionMNIST. At $\rho = 0.2$, PRIME increases Pair AP from 0.569 to 0.917 and enlarges the similarity gap $q_{\text{same}} - q_{\text{diff}}$ from 0.122 to 0.702. The improvements remain at $\rho = 0.5$, showing that PRIME recovers more discriminative pairwise similarities in addition to improving instance-level classification. Extended relation-recovery results on MNIST and KMNIST are reported in Appendix A.7.



(a) MNIST ACC (b) CIFAR-10 ACC (c) MNIST NMI (d) CIFAR-10 NMI

Figure 4: Performance under sparse same-class relation supervision. The relation budget is measured by the average revealed degree $d_+ = 2|\mathcal{S}^+|/n$. For each target d_+, the effective reveal rate is $\rho_{\text{eff}} = M/N_{\text{same}}$, where $M = \text{round}(nd_+/2)$ same-class relations are revealed. We compare Hard-Neg, PRIME with $\lambda_\pi = 0$, and PRIME with $\lambda_\pi = 5\rho_{\text{eff}}$. Results are reported as mean $\pm$ standard deviation over three random seeds.

Effect of Reveal Modeling. We next examine how performance changes with the reveal rate. Under the one-sided observation model, an unrevealed pair can be either a different-class pair or a same-class pair whose relation was not revealed. Hard-Neg instead interprets every unrevealed pair as a different-class observation.

Figure 3 reports ACC-Hungarian and NMI over a range of reveal rates on FashionMNIST and KMNIST. PRIME and PRIME+Ent degrade much less as the reveal rate decreases, whereas Hard-Neg becomes substantially less accurate in the low-reveal regime. As ρ decreases, a larger fraction of the underlying same-class relations remain unrevealed, making the hard-negative interpretation increasingly inconsistent with the observation process. Results under misspecified and estimated reveal rates are provided in Appendix A.7.

Sparse-Relation Recovery. We finally consider a sparse-relation regime in which the amount of revealed supervision is controlled directly by the average revealed degree d_+. Figure 4 reports ACC-Hungarian and NMI as the relation budget decreases on MNIST and CIFAR10.

Hard-Neg deteriorates sharply under sparse supervision because it treats every unrevealed pair as a different-class observation, including same-class pairs whose relations were not revealed. In contrast, both PRIME objectives remain effective over substantially smaller relation budgets.

At the lowest relation budget $d_+ = 0.5$, PRIME with reveal-rate-scaled prior regularization achieves 98.22% ACC-Hungarian on MNIST and 83.01% on CIFAR10, compared with 22.50% and 14.47%, respectively, for Hard-Neg. The corresponding NMI results show the same pattern.

Comparing PRIME with $\lambda_\pi = 0$ and with $\lambda_\pi = 5\rho_{\text{eff}}$ further separates the contributions of the two objective terms. The incomplete-similarity likelihood already provides strong recovery in the sparse regime, while reveal-rate-scaled prior regularization further improves performance at the lowest relation budgets. This behavior is consistent with the reveal-rate normalization analysis in Section 4.4.

6 Conclusion

We presented PRIME for learning an instance-level multi-class classifier from unlabeled instances with incompletely revealed same-class relations. PRIME models the one-sided reveal process through an incomplete-similarity likelihood and incorporates aggregate class-prior information through a permutation-invariant regularizer. Our analysis establishes recovery of the latent same-class relation, classification recovery up to a global class permutation, and properties of the class-prior regularizer. Experiments on five image benchmarks show that PRIME improves both instance-level classification and latent relation recovery over observation-naive learning, while maintaining competitive performance under reveal-rate and class-prior misspecification. These results demonstrate the potential of combining relational supervision with aggregate class information for learning from incomplete supervision.

Reproducibility Statement

We provide detailed experimental protocols and theoretical derivations to facilitate reproducibility. Section 5.1 specifies the datasets, model architectures, evaluation metrics, and comparison settings. Appendix A describes pair construction, reveal-rate estimation, numerical stabilization, optimization, implementation details, random-seed averaging, and the protocols used in the additional experiments. Appendix B provides complete proofs and auxiliary derivations for the theoretical results in Section 4.

AI Use Statement

In this work, we did not use generative AI tools for tasks requiring mandatory disclosure. Generative AI tools were used to assist with writing and language polishing, literature retrieval and discovery, and software code editing and debugging. All AI-assisted text, retrieved references, and code were reviewed and independently verified by the authors. In particular, retrieved references were checked against their original sources, and AI-assisted code was inspected and tested for correctness. We take responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

References

Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 452–461. PMLR, 2018.

Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 59–68. ACM, 2004. doi: 10.1145/1014052.1014062.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020. doi: 10.1007/s10994-020-05877-5.

Róbert Busa-Fekete, Heejin Choi, Travis Dick, Claudio Gentile, and Andrés Muñoz Medina. Easy learning from label proportions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14957–14968, 2023. doi: 10.52202/075280-0656.

Yuzhou Cao, Lei Feng, Yitian Xu, Bo An, Gang Niu, and Masashi Sugiyama. Learning from similarity-confidence data. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1272–1282. PMLR, 2021.

Jueyu Chen, Wantao Wen, Yeqiang Wang, Erliang Lin, Yemin Wang, and Yuheng Jia. Robust label proportions learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 8719–8744, 2025. doi: 10.52202/085713-0267.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223. PMLR, 2011.

Yuanchao Dai, Ximing Li, and Changchun Li. Confidence difference reflects various supervised signals in confidence-difference classification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12106–12120. PMLR, 2025.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. ACM, 2006. doi: 10.1145/1143844.1143874.

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, volume 27, pp. 703–711, 2014.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220. ACM, 2008. doi: 10.1145/1401890.1401920.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, volume 17, pp. 529–536, 2004.

Shreyas Havaldar, Navodita Sharma, Shubhi Sareen, Karthikeyan Shanmugam, and Aravindan Raghuveer. Learning from label proportions: Bootstrapping supervised learners via belief propagation. In *International Conference on Learning Representations*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. IEEE, 2016. doi: 10.1109/CVPR.2016.90.

Yen-Chang Hsu and Zsolt Kira. Neural network-based clustering using pairwise constraints. In *International Conference on Learning Representations Workshop Track*, 2016.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *International Conference on Learning Representations*, 2019.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1675–1685, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21236–21246, 2020.

Ximing Li, Yuanchao Dai, Bing Wang, Changchun Li, Jianfeng Qu, and Renchu Guan. Balancing positive and negative classification error rates in positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 81467–81490, 2025. doi: 10.52202/085713-2456.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8547–8555, 2021. doi: 10.1609/aaai.v35i10.17037.

Nan Lu, Shida Lei, Gang Niu, Issei Sato, and Masashi Sugiyama. Binary classification from multiple unlabeled datasets via surrogate set classification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7134–7144. PMLR, 2021.

Tianhao Ma, Wei Wang, Ximing Li, Gang Niu, and Masashi Sugiyama. Learning from label proportions via proportional value classification. In *International Conference on Learning Representations*, 2026.

Clayton Scott and Jianxin Zhang. Learning from label proportions: A mutual contamination framework. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22256–22267, 2020.

Takuya Shimada, Han Bao, Issei Sato, and Masashi Sugiyama. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. *Neural Computation*, 33(5):1234–1268, 2021. doi: 10.1162/neco_a_01373.

Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 577–584, 2001.

Zixi Wei, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Xiaofeng Zhu, and Heng Tao Shen. A universal unbiased method for classification from aggregate observations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36804–36820. PMLR, 2023.

Zixi Wei, Senlin Shu, Yuzhou Cao, Hongxin Wei, Bo An, and Lei Feng. Consistent multi-class classification from multiple unlabeled datasets. In *International Conference on Learning Representations*, 2024.

Songhua Wu, Tongliang Liu, Bo Han, Jun Yu, Gang Niu, and Masashi Sugiyama. Learning from noisy pairwise similarity and unlabeled data. *Journal of Machine Learning Research*, 23(307): 1–34, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 478–487. PMLR, 2016.

Zhenhuan Yang, Yunwen Lei, Puyu Wang, Tianbao Yang, and Yiming Ying. Simple stochastic and online gradient descent algorithms for pairwise learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20160–20171, 2021.

Hongjing Zhang, Tianyang Zhan, Sugato Basu, and Ian Davidson. A framework for deep constrained clustering. *Data Mining and Knowledge Discovery*, 35:593–620, 2021. doi: 10.1007/s10618-020-00734-4.

A Additional Method Details

This section provides implementation details omitted from the main text, including the construction of training pairs, estimation of the reveal rate, numerical stabilization of the objective, the entropy-regularized variant, and the optimization and inference procedures.

A.1 Pair Construction

PRIME learns from revealed same-class relations rather than from complete pairwise similarity labels. Therefore, the sampled training pairs should preserve the original observation mechanism, where the reveal indicator is determined by membership in the observed relation set.

Let $\Omega_n = \{(i,j) : 1 \le i < j \le n\}$ denote the set of unordered pairs among the training instances. Since pairwise relations are symmetric, for any $i \neq j$, we use the canonical index representation $c(i,j) = (\min\{i,j\}, \max\{i,j\}) \in \Omega_n$. The observed same-class relations form a subset $\mathcal{S}^+ \subseteq \Omega_n$. By construction, every revealed pair satisfies

$$(i,j) \in \mathcal{S}^+ \Longrightarrow s^\star_{ij} = 1,$$

and the corresponding reveal indicator is defined as $z_{ij} = \mathbb{I}[c(i,j) \in \mathcal{S}^+]$. During training, an instance mini-batch $\mathcal{B}_{\mathrm{pair}} = \{x_b\}_{b=1}^{B_{\mathrm{pair}}}$ is sampled from the unlabeled training set. The pair set used for optimization is constructed as

$$\mathcal{P} \subseteq \{c(i,j) : x_i, x_j \in \mathcal{B}_{\mathrm{pair}},\ i \neq j\},$$

where the construction is independent of $\mathcal{S}^+$. For each sampled pair, the reveal indicator is subsequently obtained by checking whether the pair belongs to the observed relation set.

This procedure follows the incomplete-similarity setting: unrevealed pairs are sampled as ordinary unlabeled pairs and are not explicitly generated as different-class examples.

A.2 Reveal-Rate Estimation

When the reveal rate ρ is unavailable, PRIME estimates it from the marginal frequency of revealed relations. Let $\pi = [\pi_1, \ldots, \pi_K]^\top, \pi_k = P(Y = k)$, denote the class prior. For two independent samples $X, X' \sim P_X$, the probability that they belong to the same class is $P(s^\star(X, X') = 1) = \sum_{k=1}^K \pi_k^2$. Under the one-sided reveal mechanism, a pair is observed only when it is a same-class pair and the relation is revealed. Therefore, $P(Z = 1) = \rho \sum_{k=1}^K \pi_k^2$. This relation provides a population-level estimator of ρ when the class prior is known. In finite samples, however, PRIME constructs pairs without self-pairs. Let n_k denote the number of training instances from class k, where $\sum_{k=1}^K n_k = n$. The probability that two uniformly sampled distinct training instances belong to the same class is then

$$p_{\text{same}}^{(n)} = \frac{\sum_{k=1}^K n_k(n_k - 1)}{n(n-1)}.$$

When the supplied prior corresponds to the empirical training distribution, $\pi_k = n_k/n$, the above quantity can also be written as

$$p_{\text{same}}^{(n)} = \frac{n \sum_{k=1}^K \pi_k^2 - 1}{n - 1}.$$

This finite-sample correction accounts for sampling without replacement and converges to $\sum_{k=1}^K \pi_k^2$ as n increases.

Given N randomly constructed training pairs with reveal indicators $\{z_m\}_{m=1}^N$, the empirical reveal frequency is $\widehat{p}_z = \frac{1}{N} \sum_{m=1}^N z_m$. Since the conditional marginal reveal probability satisfies $P(Z = 1 \mid U) = \rho p_{\text{same}}^{(n)}$, we estimate the reveal rate by $\widehat{\rho}_{\text{clip}} = \min\left\{1, \frac{\widehat{p}_z}{p_{\text{same}}^{(n)}}\right\}$. When only an estimated class prior $\widehat{\pi}$ is available, we use the plug-in estimate

$$\widehat{p}_{\text{same}}^{(n)} = \frac{n \sum_{k=1}^K \widehat{\pi}_k^2 - 1}{n - 1},$$

and substitute it into the above estimator.

A.3 Numerical Stabilization

The objective in Section 3 is defined according to the ideal observation model. In practical optimization, however, the predicted observation probability may approach the boundary values of the Bernoulli likelihood, which can lead to numerical instability. We therefore introduce clipping and smoothing operations for the empirical objective.

The modeled probability of observing a revealed relation is $P_\theta(z_{ij} = 1 \mid x_i, x_j) = \rho f_i^\top f_j$. For a small constant $0 < \epsilon < 1/2$, we define the clipping operator $\text{clip}_\epsilon(t) = \min\{\max\{t, \epsilon\}, 1 - \epsilon\}$. The stabilized incomplete-similarity loss for a sampled pair is

$$\ell_{\text{IS}}^\epsilon(x_i, x_j, z_{ij}) = -z_{ij} \log \text{clip}_\epsilon\left(\rho f_i^\top f_j\right) - (1 - z_{ij}) \log\left(1 - \text{clip}_\epsilon\left(\rho f_i^\top f_j\right)\right).$$

Accordingly, the empirical relational loss used during training is

$$\mathcal{L}_{\text{IS}}^\epsilon = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \ell_{\text{IS}}^\epsilon(x_i, x_j, z_{ij}).$$

When the reveal rate is unavailable, the estimated value $\widehat{\rho}_{\text{clip}}$ from Appendix A.2 is used in place of ρ.

The class-prior regularizer also requires numerical stabilization because a mini-batch prediction marginal may contain zero entries. Given a prior mini-batch $\mathcal{B}_{\text{prior}}$, the predicted class marginal is $\bar{f} = \frac{1}{B_{\text{prior}}} \sum_{x_b \in \mathcal{B}_{\text{prior}}} f_\theta(x_b)$. We apply additive smoothing to each coordinate:

$$\bar{f}_k^{\epsilon} = \frac{\bar{f}_k + \epsilon}{1 + K\epsilon}, \qquad k = 1, \ldots, K. \tag{3}$$

The implemented prior regularizer is then $\mathcal{L}^{\epsilon}_{\text{prior}} = \min_{P \in \mathcal{P}_K} D_{\text{KL}}\left(\pi \| P\bar{f}^{\epsilon}\right)$. Combining the two stabilized terms gives the optimization objective

$$\mathcal{L}^{\epsilon}_{\text{PRIME}} = \mathcal{L}^{\epsilon}_{\text{IS}} + \lambda_{\pi}\mathcal{L}^{\epsilon}_{\text{prior}}.$$

The above stabilization operations are used only for numerical optimization. The population risks and theoretical results in Section 4 are defined according to the original, unstabilized objective.

A.4 Entropy-Regularized Variant

In addition to PRIME, we consider an entropy-regularized variant, denoted PRIME+Ent. This variant introduces an additional instance-level regularization term to encourage lower-entropy predictions, while the basic PRIME objective remains unchanged.

For a prior mini-batch $\mathcal{B}_{\text{prior}}$, the entropy regularizer is defined as

$$\mathcal{L}_{\text{ent}} = \frac{1}{B_{\text{prior}}} \sum_{x_b \in \mathcal{B}_{\text{prior}}} H(f_\theta(x_b)), \qquad H(p) = -\sum_{k=1}^{K} p_k \log p_k.$$

PRIME+Ent optimizes

$$\mathcal{L}^{\epsilon}_{\text{PRIME+Ent}} = \mathcal{L}^{\epsilon}_{\text{IS}} + \lambda_{\pi}\mathcal{L}^{\epsilon}_{\text{prior}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}, \qquad \lambda_{\text{ent}} \geq 0.$$

The entropy term is introduced only as an additional experimental variant. It is not part of the basic PRIME formulation or the theoretical analysis.

A.5 Optimization and Inference

PRIME is optimized using mini-batch stochastic optimization. At each iteration, an instance batch $\mathcal{B}_{\text{pair}}$ is sampled to construct the pair set $\mathcal{P}$ and evaluate $\mathcal{L}^{\epsilon}_{\text{IS}}$. An independent batch $\mathcal{B}_{\text{prior}}$ is sampled to compute $\mathcal{L}^{\epsilon}_{\text{prior}}$ and, for PRIME+Ent, $\mathcal{L}_{\text{ent}}$.

The permutation in the class-prior regularizer is obtained by solving the linear assignment problem $\min_{P \in \mathcal{P}_K} D_{\text{KL}}\left(\pi \| P\bar{f}^{\epsilon}\right)$, and the resulting permutation is fixed when computing gradients of the prior term.

When the reveal rate is unknown, PRIME uses $\widehat{\rho}_{\text{clip}}$ estimated from Appendix A.2. The model parameters are then updated using the stabilized PRIME objective, or the entropy-regularized objective for PRIME+Ent.

At inference time, the predicted class is obtained as $\widehat{y}(x) = \arg\max_{k \in \{1,\ldots,K\}} f_\theta(x)_k$. Since latent output coordinates are identifiable only up to a global class permutation, class indices are aligned with ground-truth labels only for permutation-dependent evaluation metrics. Permutation-invariant metrics, such as NMI and ARI, do not require this alignment.

A.6 Additional Experimental Details

This section provides additional details of the experimental protocols, including reveal-rate estimation, prior perturbation, long-tailed data construction, sparse-relation sampling, and implementation settings.

Additional evaluation metrics. For the supplementary analyses, we additionally report adjusted Rand index (ARI), the number of active predicted clusters, the maximum predicted cluster fraction (MaxFrac), and prediction entropy. ARI evaluates agreement between the predicted and ground-truth partitions without requiring a label permutation. The active-cluster count and MaxFrac characterize the usage of the latent output coordinates, while prediction entropy measures the concentration of instance-level predictions.

Reveal-rate estimation. When the reveal rate is not provided, we estimate it from the empirical frequency of revealed relations. For a training set with class counts $\{n_k\}_{k=1}^K$, the probability that two uniformly sampled distinct instances form a same-class pair is

$$p_{\text{same}}^{(n)} = \frac{\sum_{k=1}^K n_k(n_k-1)}{n(n-1)}.$$

Given N randomly constructed training pairs, the empirical reveal frequency is $\widehat{p}_z = \frac{1}{N}\sum_{m=1}^N z_m$. The estimated reveal rate is obtained as

$$\widehat{\rho}_{\text{clip}} = \min\left\{1, \frac{\widehat{p}_z}{p_{\text{same}}^{(n)}}\right\}.$$

When only an estimated class prior $\widehat{\pi}$ is available, we replace $p_{\text{same}}^{(n)}$ with

$$\widehat{p}_{\text{same}}^{(n)} = \frac{n\sum_{k=1}^K \widehat{\pi}_k^2 - 1}{n-1}.$$

Prior perturbation protocol. To evaluate the influence of prior misspecification, we perturb the empirical training prior using a multiplicative log-normal model. Let

$$\widehat{\pi}_k = \frac{n_k^{\text{train}}}{\sum_{c=1}^K n_c^{\text{train}}}$$

denote the empirical prior. For each random seed s, we sample $\xi^{(s)} \sim \mathcal{N}(0, I_K)$ and construct

$$\widetilde{\pi}_{\alpha,k}^{(s)} = \frac{\widehat{\pi}_k \exp(\alpha\xi_k^{(s)})}{\sum_{c=1}^K \widehat{\pi}_c \exp(\alpha\xi_c^{(s)})}, \qquad \alpha \geq 0. \tag{4}$$

We evaluate $\alpha \in \{0, 0.1, 0.2, 0.4, 0.8\}$. The same perturbation direction is used across methods for each random seed.

Long-tailed MNIST protocol. We construct a long-tailed MNIST training set with imbalance factor $\text{IF} = 10$. For class k, the retained number of samples is $n_k = \left\lceil n_k^{\text{orig}} \, \text{IF}^{-\frac{k}{K-1}} \right\rceil, k = 0, \ldots, K-1$. The resulting class distribution is $(5923, 5221, 3572, 2846, 2100, 1509, 1275, 1046, 756, 595)$. The corresponding empirical prior is

$$\widehat{\pi}_k^{\text{LT}} = \frac{n_k}{\sum_{c=0}^{K-1} n_c}.$$

We compare the empirical prior, its perturbation with $\alpha = 0.2$, and the uniform prior $\pi_k = 1/K$.

Sparse-relation protocol. We evaluate PRIME under sparse revealed same-class relations by controlling the average number of revealed neighbors per instance. Specifically, we define the relation budget as $d_+ = \frac{2|\mathcal{S}^+|}{n}$, and consider $d_+ \in \{0.5, 1, 2, 4, 8\}$. For each dataset and random seed, we reveal exactly $M = \text{round}(nd_+/2)$ same-class relations sampled uniformly from all ground-truth same-class pairs. The effective reveal rate used by the incomplete-similarity likelihood is $\rho_{\text{eff}} = \frac{M}{N_{\text{same}}}$, where $N_{\text{same}} = \sum_{k=1}^K \binom{n_k}{2}$. Under these sparse budgets, uniformly sampling from all unordered training pairs would yield very few revealed relations in a mini-batch. We therefore adopt a stratified pair sampler.

Each pair mini-batch contains 512 pairs, with 256 revealed pairs sampled from $\mathcal{S}^+$ and 256 unrevealed pairs sampled from $\Omega_n \setminus \mathcal{S}^+$. The unrevealed stratum is constructed only according to membership in $\mathcal{S}^+$, without using ground-truth class labels. Hence, an unrevealed pair may be either a different-class pair or a same-class pair whose relation was not revealed.

Because stratified sampling changes the pair-sampling distribution, we use an importance-corrected estimator to preserve the original empirical pairwise objective. Let $w_+ = \frac{|\mathcal{S}^+|}{\binom{n}{2}}$ denote the probability that a uniformly sampled unordered pair is revealed. For a generic pairwise loss ℓ, the corrected estimator is

$$\widehat{L}_{\text{pair}} = w_+ \frac{1}{B_+} \sum_{(i,j)\in\mathcal{B}_+} \ell(q_{ij}, 1) + (1-w_+)\frac{1}{B_0} \sum_{(i,j)\in\mathcal{B}_0} \ell(q_{ij}, 0),$$

where $B_+ = B_0 = 256$. This estimator is unbiased for the empirical pairwise objective under uniform sampling over all unordered training pairs.

For the sparse-relation comparison, we evaluate three objectives under the same sampling protocol. Hard-Neg directly uses the reveal indicator as the pairwise similarity label. PRIME with $\lambda_\pi = 0$ removes the prior regularizer and optimizes only the incomplete-similarity likelihood. The full PRIME model uses the reveal-rate-scaled coefficient $\lambda_\pi = 5\rho_{\text{eff}}$, which follows from the reference configuration $(\rho_{\text{ref}}, \lambda_{\text{ref}}) = (0.2, 1)$. The scaling coefficient is fixed before the experiment and is not tuned for individual datasets or relation budgets.

Implementation details. All experiments use the same backbone and training protocol for controlled comparison. Models are trained with Adam for 40 epochs and 200 optimization steps per epoch, using a learning rate of 10^{-3} and a pair batch size of 512. For PRIME and PRIME+Ent, an additional instance batch of size 1024 is sampled to evaluate the class-prior regularizer. The numerical clipping and marginal smoothing operations described in Appendix A.3 are used during optimization.

Unless otherwise specified, we use the reveal-rate-aware parameterization $\lambda_\pi = \beta\rho$ introduced in Section 4.4, with $\beta = 5$. This corresponds to the reference configuration $(\rho_{\text{ref}}, \lambda_{\text{ref}}) = (0.2, 1.0)$. When the reveal rate is estimated rather than given, ρ is replaced by its estimate $\hat{\rho}$. PRIME+Ent additionally uses $\lambda_{\text{ent}} = 0.05$. The target prior is set to the empirical class distribution of the training set.

For the clustering baselines, KCL, DCC, and CC use the same training/test splits and the same number of target clusters as PRIME. KCL receives revealed same-class pairs as similarity constraints and does not construct dissimilarity constraints from unrevealed pairs. DCC uses the revealed relations as must-link constraints, while CC uses only the unlabeled instances.

Unless otherwise specified, results are averaged over three random seeds and reported as mean $\pm$ standard deviation.

A.7 Additional Experimental Results

Extended Relation-Recovery Results. Table A.1 extends the relation-recovery evaluation in the main text to MNIST and KMNIST. The improvements observed on FashionMNIST also appear on these datasets. At $\rho = 0.2$, PRIME increases Pair AP from 0.615 to 0.898 on KMNIST, while on MNIST the similarity gap increases from 0.126 for Hard-Neg to 0.919 for PRIME. PRIME+Ent further increases the MNIST similarity gap to 0.964. Across the three datasets, the differences are generally larger in the lower-reveal regime, consistent with the main classification results.

Sensitivity to Prior Regularization. Figure A.1 evaluates performance and latent-class usage over different values of the prior-regularization coefficient λ_π. PRIME and PRIME+Ent maintain comparatively stable ACC-Hungarian and NMI over the tested range, while the active-cluster count and MaxFrac describe the corresponding changes in latent class usage. Hard-Neg exhibits substantially larger variation in both recovery performance and class usage as λ_π changes.

Figure A.2 further reports prediction entropy on FashionMNIST. PRIME produces more concentrated predictions than Hard-Neg across the tested values of λ_π, while PRIME+Ent yields the lowest prediction entropy because of its additional instance-level entropy regularizer. This separates the marginal constraint imposed by the class-prior term from the instance-level concentration encouraged by the entropy term.

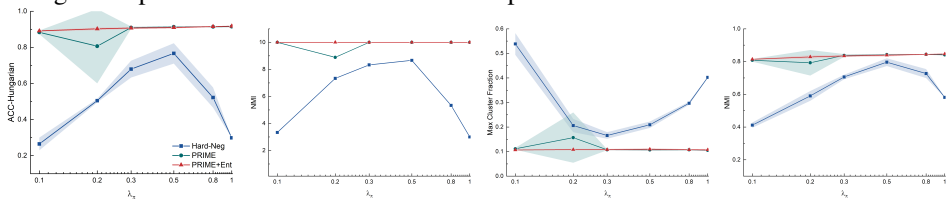
Prior Misspecification and Class Imbalance. Table A.2 examines the effect of perturbing the target class prior on MNIST. Under moderate perturbations, PRIME retains high ACC-Hungarian and uses all or nearly all latent coordinates. As the perturbation becomes larger, both classification performance and the predicted class allocation change more substantially.

The long-tailed experiment further evaluates the effect of the target prior when the training distribution is imbalanced. For PRIME, using the empirical long-tailed prior yields 50.73% ACC-Hungarian with five active clusters, whereas using a uniform target prior yields 69.10% ACC-Hungarian with

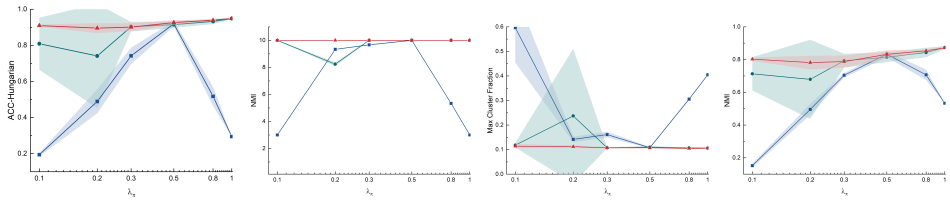
Table A.1: Same-class relation recovery under incomplete-similarity supervision. Results are reported as mean ± standard deviation over three seeds. Pair AUC and Pair AP evaluate the learned similarity score $q_{ij} = f_\theta(x_i)^\top f_\theta(x_j)$ for distinguishing same-class and different-class pairs. The similarity gap is defined as $q_{\text{same}} - q_{\text{diff}}$.

Dataset	ρ	Method	Pair AUC	Pair AP	$q_{\text{same}} - q_{\text{diff}}$
FashionMNIST	0.2	Hard-Neg	0.934 ± 0.002	0.569 ± 0.018	0.122 ± 0.005
		PRIME	0.987 ± 0.002	$\mathbf{0.917 \pm 0.009}$	0.702 ± 0.025
		PRIME+Ent	$\mathbf{0.987 \pm 0.003}$	0.917 ± 0.013	$\mathbf{0.750 \pm 0.027}$
	0.5	Hard-Neg	0.976 ± 0.003	0.806 ± 0.046	0.400 ± 0.015
		PRIME	$\mathbf{0.989 \pm 0.000}$	$\mathbf{0.930 \pm 0.003}$	0.764 ± 0.011
		PRIME+Ent	0.989 ± 0.001	0.927 ± 0.005	$\mathbf{0.768 \pm 0.019}$
KMNIST	0.2	Hard-Neg	0.928 ± 0.005	0.615 ± 0.024	0.103 ± 0.006
		PRIME	$\mathbf{0.977 \pm 0.006}$	$\mathbf{0.898 \pm 0.023}$	0.733 ± 0.032
		PRIME+Ent	0.973 ± 0.006	0.881 ± 0.018	$\mathbf{0.743 \pm 0.023}$
	0.5	Hard-Neg	0.976 ± 0.001	0.895 ± 0.005	0.406 ± 0.004
		PRIME	0.980 ± 0.010	0.906 ± 0.039	0.778 ± 0.070
		PRIME+Ent	$\mathbf{0.984 \pm 0.004}$	$\mathbf{0.927 \pm 0.020}$	$\mathbf{0.826 \pm 0.040}$
MNIST	0.2	Hard-Neg	0.984 ± 0.005	0.888 ± 0.042	0.126 ± 0.002
		PRIME	0.999 ± 0.000	0.995 ± 0.000	0.919 ± 0.007
		PRIME+Ent	$\mathbf{0.999 \pm 0.000}$	$\mathbf{0.996 \pm 0.000}$	$\mathbf{0.964 \pm 0.004}$
	0.5	Hard-Neg	0.998 ± 0.000	0.989 ± 0.002	0.461 ± 0.014
		PRIME	0.999 ± 0.000	0.996 ± 0.001	0.959 ± 0.008
		PRIME+Ent	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.997 \pm 0.000}$	$\mathbf{0.973 \pm 0.002}$

Figure A.1: Performance under different prior-regularization coefficients λ_π on FashionMNIST and KMNIST. Rows correspond to datasets, and columns report ACC-Hungarian, active clusters, MaxFrac, and NMI. All methods use the same permutation-invariant class-prior penalty. Method variants differ in their pairwise likelihoods, and PRIME+Ent additionally includes the entropy term. Shaded regions represent standard deviations over repeated runs.



FashionMNIST



KMNIST

eight active clusters. These results show that the supplied aggregate prior can materially affect the recovered class allocation when the underlying training distribution is strongly imbalanced.

Reveal-Rate Misspecification and Estimation. Table A.3 evaluates PRIME when the reveal rate used by the training objective differs from the rate that generated the observed relations. For both $\rho_{\text{true}} = 0.2$ and 0.5, ACC-Hungarian and NMI remain relatively stable over moderate misspecification of ρ_{model}, whereas the similarity gap becomes more sensitive when the modeled reveal rate is substantially overestimated.

The reveal-rate estimator gives $\widehat{\rho} = 0.201$ when $\rho_{\text{true}} = 0.2$ and $\widehat{\rho} = 0.505$ when $\rho_{\text{true}} = 0.5$. Using these estimates yields performance close to that obtained with the corresponding true reveal rates.

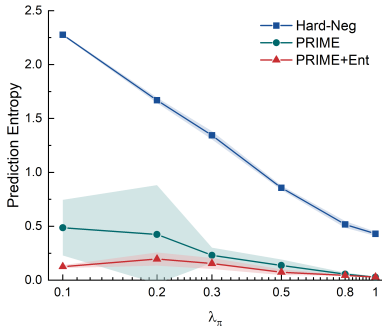


Figure A.2: Prediction entropy on FashionMNIST under different prior-regularization coefficients λ_π. All methods use the same permutation-invariant class-prior penalty; Hard-Neg denotes the standard Hard-Neg likelihood combined with the same penalty. Lower entropy indicates more concentrated latent class predictions. Shaded regions represent standard deviations over repeated runs.

Table A.2: Performance under class-prior misspecification and class imbalance on MNIST at $\rho = 0.2$. The target prior is perturbed according to the log-normal multiplicative model in (4). The long-tailed training split is constructed with IF $= 10$, while the original test split is retained for evaluation. ACC denotes ACC-Hungarian (%), Active denotes the number of active predicted clusters, and MaxFrac denotes the largest predicted cluster fraction.

Setting	Method	ACC (%)	NMI	ARI	Active	MaxFrac	Pair AP	$q_{\text{same}} - q_{\text{diff}}$
Balanced MNIST with perturbed class prior								
$\alpha = 0.0$	PRIME	**99.05**	**0.972**	**0.979**	10	0.113	0.994	0.931
$\alpha = 0.0$	PRIME+Ent	98.75	0.965	0.973	10	**0.112**	**0.997**	**0.969**
$\alpha = 0.1$	PRIME	**98.57**	**0.960**	**0.969**	10	**0.113**	**0.990**	0.878
$\alpha = 0.1$	PRIME+Ent	96.43	0.937	0.924	10	0.114	0.989	**0.913**
$\alpha = 0.2$	PRIME	**98.94**	**0.970**	**0.977**	**10**	**0.113**	**0.990**	0.845
$\alpha = 0.2$	PRIME+Ent	93.45	0.935	0.894	9	0.151	0.975	**0.881**
$\alpha = 0.4$	PRIME	79.38	**0.905**	**0.794**	8	0.201	**0.861**	0.705
$\alpha = 0.4$	PRIME+Ent	**81.46**	0.865	0.764	**9**	**0.176**	0.851	**0.765**
$\alpha = 0.8$	PRIME	59.74	0.808	0.590	6	0.303	**0.620**	0.655
$\alpha = 0.8$	PRIME+Ent	**66.67**	**0.812**	**0.607**	6	**0.255**	0.545	**0.731**
Long-tailed MNIST training set with imbalance factor 10								
Empirical prior	PRIME	**50.73**	**0.729**	**0.427**	**5**	**0.415**	**0.455**	0.555
Empirical prior	PRIME+Ent	38.08	0.662	0.324	4	0.478	0.322	**0.594**
Perturbed empirical prior ($\alpha = 0.2$)	PRIME	**56.38**	**0.760**	**0.510**	**6**	**0.343**	**0.463**	0.572
Perturbed empirical prior ($\alpha = 0.2$)	PRIME+Ent	39.09	0.673	0.382	4	0.405	0.318	**0.617**
Uniform prior	PRIME	**69.10**	**0.827**	**0.687**	**8**	**0.214**	**0.715**	0.640
Uniform prior	PRIME+Ent	68.95	0.813	0.659	7	0.221	0.650	**0.733**

B Proofs and Auxiliary Derivations

This appendix provides the proofs and auxiliary results for the theoretical analysis in the main text.

B.1 Auxiliary Notation

For a classifier $f : \mathcal{X} \to \Delta^{K-1}$, where

$$\Delta^{K-1} = \left\{ u \in \mathbb{R}^K : u_k \geq 0,\ \sum_{k=1}^{K} u_k = 1 \right\},$$

we define the induced pairwise similarity $g_f(x, x') = f(x)^\top f(x')$. Since both $f(x)$ and $f(x')$ belong to the probability simplex, $0 \leq g_f(x, x') \leq 1$. Let $y : \mathcal{X} \to \{1, \ldots, K\}$ denote the ground-truth labeling function. The latent same-class relation is $s^\star(x, x') = \mathbb{I}[y(x) = y(x')]$. When the classifier output is one-hot, $g_f(x, x')$ coincides with the same-class relation induced by the predicted

Table A.3: Performance under reveal-rate misspecification on MNIST. The true reveal rate ρ_{true} determines the observed same-class relations, whereas ρ_{model} is used in the training objective. ACC is reported as ACC-Hungarian (%).

$\rho_{\text{model}}/\rho_{\text{true}}$	ρ_{model}	Method	ACC (%)	NMI	ARI	Pair AP	$q_{\text{same}} - q_{\text{diff}}$
			$\rho_{\text{true}} = 0.2$				
0.5	0.100	PRIME	98.91	0.968	0.976	0.994	0.953
0.5	0.100	PRIME+Ent	98.82	0.966	0.974	0.998	0.972
0.75	0.150	PRIME	98.97	0.971	0.977	0.995	0.941
0.75	0.150	PRIME+Ent	98.89	0.967	0.976	0.997	0.971
1.0	0.200	PRIME	98.84	0.966	0.975	0.996	0.932
1.0	0.200	PRIME+Ent	99.04	0.972	0.979	0.997	0.973
1.25	0.250	PRIME	98.98	0.970	0.978	0.996	0.882
1.25	0.250	PRIME+Ent	98.81	0.965	0.974	0.997	0.963
1.5	0.300	PRIME	98.64	0.961	0.970	0.993	0.671
1.5	0.300	PRIME+Ent	98.57	0.960	0.969	0.995	0.942
Est.	0.201	PRIME	99.01	0.971	0.978	0.995	0.927
Est.	0.201	PRIME+Ent	98.92	0.968	0.976	0.998	0.972
			$\rho_{\text{true}} = 0.5$				
0.5	0.250	PRIME	98.94	0.969	0.977	0.998	0.976
0.5	0.250	PRIME+Ent	98.92	0.968	0.976	0.998	0.976
0.75	0.375	PRIME	99.04	0.971	0.979	0.998	0.977
0.75	0.375	PRIME+Ent	99.18	0.976	0.982	0.997	0.980
1.0	0.500	PRIME	99.01	0.972	0.978	0.994	0.948
1.0	0.500	PRIME+Ent	98.97	0.970	0.977	0.998	0.975
1.25	0.625	PRIME	98.81	0.966	0.974	0.995	0.813
1.25	0.625	PRIME+Ent	99.09	0.973	0.980	0.998	0.965
1.5	0.750	PRIME	98.86	0.968	0.975	0.995	0.656
1.5	0.750	PRIME+Ent	98.60	0.961	0.969	0.990	0.716
Est.	0.505	PRIME	98.83	0.966	0.974	0.993	0.946
Est.	0.505	PRIME+Ent	99.00	0.971	0.978	0.997	0.975

labels. In particular, the population recovery results below characterize when the soft similarity g_f recovers the latent relation $s^\star$.

B.2 Derivation of the Incomplete-Similarity Likelihood

For distinct training indices $i \neq j$, define the latent same-class relation $s^\star(x_i, x_j) = \mathbb{I}[y_i = y_j]$, and write $s^\star_{ij} = s^\star(x_i, x_j)$ for simplicity. Let $z_{ij} \in \{0, 1\}$ indicate whether the pair is included in the revealed same-class relation set. Under the one-sided observation model,

$$P(z_{ij} = 1 \mid x_i, x_j) = \rho s^\star_{ij},$$

where $\rho \in (0, 1]$ is the reveal probability for same-class relations. Equivalently, $P(z_{ij} = 0 \mid x_i, x_j) = 1 - \rho s^\star_{ij}$. PRIME replaces the latent binary relation with the classifier-induced soft similarity $q_{ij} = f_\theta(x_i)^\top f_\theta(x_j) \in [0, 1]$. The corresponding modeled observation probabilities are

$$P_\theta(z_{ij} = 1 \mid x_i, x_j) = \rho q_{ij}, \qquad P_\theta(z_{ij} = 0 \mid x_i, x_j) = 1 - \rho q_{ij}.$$

Therefore, for $z_{ij} \in \{0, 1\}$,

$$P_\theta(z_{ij} \mid x_i, x_j) = (\rho q_{ij})^{z_{ij}} (1 - \rho q_{ij})^{1 - z_{ij}},$$

which gives the ideal incomplete-similarity loss

$$\ell^0_{\text{IS}} = -z_{ij} \log(\rho q_{ij}) - (1 - z_{ij}) \log(1 - \rho q_{ij}).$$

For numerical stability, we define

$$\text{clip}_\epsilon(t) = \min\{\max\{t, \epsilon\}, 1 - \epsilon\}, \qquad 0 < \epsilon < 1/2.$$

The stabilized loss is

$$\ell^\epsilon_{\text{IS}} = -z_{ij} \log \text{clip}_\epsilon(\rho q_{ij}) - (1 - z_{ij}) \log\left(1 - \text{clip}_\epsilon(\rho q_{ij})\right).$$

B.3 Properties of the Stabilized Incomplete-Similarity Loss

For the finite-sample analysis, we consider the scalar form of the stabilized incomplete-similarity loss. For $g \in [0,1]$ and $z \in \{0,1\}$, define $q_\epsilon(g) = \mathrm{clip}_\epsilon(\rho g)$, where

$$\mathrm{clip}_\epsilon(t) = \min\{\max\{t, \epsilon\}, 1-\epsilon\}, \qquad 0 < \epsilon < 1/2.$$

The corresponding loss is

$$\ell_{\rho,\epsilon}(g,z) = -z \log q_\epsilon(g) - (1-z)\log(1-q_\epsilon(g)). \tag{5}$$

Lemma 1 (Boundedness of the stabilized loss)**.** *For every $g \in [0,1]$ and $z \in \{0,1\}$,*

$$0 \le \ell_{\rho,\epsilon}(g,z) \le M_\epsilon,$$

where $M_\epsilon = \log \frac{1}{\epsilon}$.

Proof. By construction, $q_\epsilon(g) \in [\epsilon, 1-\epsilon]$. Therefore,

$$-\log q_\epsilon(g) \le \log \frac{1}{\epsilon},$$

and similarly,

$$-\log(1-q_\epsilon(g)) \le \log \frac{1}{\epsilon}.$$

Since $z \in \{0,1\}$, the loss selects one of these two terms. The non-negativity follows from $q_\epsilon(g) \in (0,1)$. Hence,

$$0 \le \ell_{\rho,\epsilon}(g,z) \le \log \frac{1}{\epsilon}.$$

□

Lemma 2 (Lipschitz continuity of the stabilized loss)**.** *For each fixed $z \in \{0,1\}$, the function $\ell_{\rho,\epsilon}(g,z)$ is $L_{\rho,\epsilon}$-Lipschitz on $[0,1]$, with*

$$L_{\rho,\epsilon} = \frac{\rho}{\epsilon}.$$

Proof. The clipping operator is 1-Lipschitz, and the mapping $g \mapsto \rho g$ is ρ-Lipschitz. Hence,

$$|q_\epsilon(g) - q_\epsilon(g')| \le \rho |g - g'|.$$

On the interval $[\epsilon, 1-\epsilon]$, both functions $u \mapsto -\log u$ and $u \mapsto -\log(1-u)$ are $1/\epsilon$-Lipschitz. Therefore, for fixed z,

$$|\ell_{\rho,\epsilon}(g,z) - \ell_{\rho,\epsilon}(g',z)| \le \frac{1}{\epsilon}|q_\epsilon(g) - q_\epsilon(g')| \le \frac{\rho}{\epsilon}|g - g'|.$$

Thus, $\ell_{\rho,\epsilon}(g,z)$ is $L_{\rho,\epsilon}$-Lipschitz with $L_{\rho,\epsilon} = \rho/\epsilon$. □

B.4 Proof of Proposition 1

Proof. For any fixed pair (x,x'), define $\eta(x,x') = P(Z=1 \mid x,x')$, and let $q_f(x,x') = \rho g_f(x,x')$. The conditional risk associated with this pair is

$$r(q;\eta) = -\eta \log q - (1-\eta)\log(1-q).$$

Using the standard cross-entropy decomposition,

$$r(q;\eta) - r(\eta;\eta) = \mathrm{KL}\left(\mathrm{Bern}(\eta) \| \mathrm{Bern}(q)\right) \ge 0,$$

where boundary cases are defined by continuity. Equality holds if and only if $q = \eta$. Under the one-sided reveal model, $\eta(x,x') = \rho s^\star(x,x')$. Since $s^\star(x,x') \in \{0,1\}$, we have $\eta(x,x') \in \{0,\rho\}$. Moreover, because $g_f(x,x') \in [0,1]$,

$$q_f(x,x') = \rho g_f(x,x') \in [0,\rho],$$

which contains all possible values of $\eta(x,x')$. Hence, the pointwise minimizer of the conditional risk satisfies

$$\rho g_f(x,x') = \rho s^\star(x,x').$$

Because $\rho > 0$, this is equivalent to $g_f(x,x') = s^\star(x,x')$. By the realizability assumption $s^\star \in \mathcal{G}$, there exists a classifier f whose induced pairwise function attains this pointwise optimum. Therefore, every population minimizer $f^\star$ of R^0_{IS} satisfies $g_{f^\star}(x,x') = s^\star(x,x')$ almost surely. □

B.5 Proof of Corollary 1

Proof. Assume that $g_f(x, x') = s^\star(x, x')$ almost surely. Let P_k denote the class-conditional distribution of X given $Y = k$. Since every class has positive probability, P_k is well defined.

For any class k, we have $f(x)^\top f(x') = 1$, for $P_k \otimes P_k$-almost every pair (x, x'). We first note that for any two vectors $a, b \in \Delta^{K-1}$, $a^\top b \le \|a\|_2 \|b\|_2 \le 1$. Equality holds only when $a = b$ and $\|a\|_2 = 1$. Since a probability-simplex vector has unit Euclidean norm only when it is one-hot, it follows that $f(x) = f(x') = e_{\tau(k)}$ for some coordinate $\tau(k)$ and P_k-almost every x. Next, consider two distinct classes $k \neq \ell$. For $P_k \otimes P_\ell$-almost every pair (x, x'), $f(x)^\top f(x') = s^\star(x, x') = 0$. Therefore, $e_{\tau(k)}^\top e_{\tau(\ell)} = 0$, which implies $\tau(k) \neq \tau(\ell)$. Hence, τ is injective. Because the number of ground-truth classes and latent output coordinates are both K, τ is a permutation σ of $\{1, \dots, K\}$. Therefore, $f(x) = e_{\sigma(y(x))}$ almost surely. □

B.6 Proof of Theorem 1

Proof. For $f \in \mathcal{F}$, define the pairwise loss class $h_f(x, x', z) = \ell_{\rho,\epsilon}\big(g_f(x, x'), z\big)$, and let $\mathcal{H}_{\rho,\epsilon} = \{h_f : f \in \mathcal{F}\}$. By Lemma 1,

$$0 \le h_f(x, x', z) \le M_\epsilon.$$

Let

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_X, \qquad m = \left\lfloor \frac{n}{2} \right\rfloor.$$

Conditional on the sampled instances, the reveal variables are generated independently according to the one-sided observation model:

$$P(Z_{ij} = 1 \mid X_i, X_j) = \rho s^\star(X_i, X_j).$$

The empirical risk is

$$\widehat{R}^\epsilon_{\mathrm{IS},n}(f) = \binom{n}{2}^{-1} \sum_{i<j} h_f(X_i, X_j, Z_{ij}).$$

To handle the dependence among pairwise terms, consider a uniformly random matching Π of size m, independent of the data, with matched pairs $(I_1, J_1), \dots, (I_m, J_m)$. Each unordered pair appears in the matching with probability $\frac{m}{\binom{n}{2}}$. Therefore,

$$\widehat{R}^\epsilon_{\mathrm{IS},n}(f) = \mathbb{E}_\Pi \left[\frac{1}{m} \sum_{r=1}^{m} h_f(X_{I_r}, X_{J_r}, Z_{I_r J_r}) \middle| X_1, \dots, X_n, \{Z_{ij}\} \right].$$

By Jensen's inequality,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \widehat{R}^\epsilon_{\mathrm{IS},n}(f) - R^\epsilon_{\mathrm{IS}}(f) \right| \le \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{r=1}^{m} h_f(X_{I_r}, X_{J_r}, Z_{I_r J_r}) - R^\epsilon_{\mathrm{IS}}(f) \right|.$$

Because matched pairs contain disjoint instances and the reveal variables are conditionally independent given the instances, the matched triples are independent and identically distributed according to the population distribution of (X, X', Z). Standard symmetrization yields

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \widehat{R}^\epsilon_{\mathrm{IS},n}(f) - R^\epsilon_{\mathrm{IS}}(f) \right| \le 2\mathfrak{R}_m(\mathcal{H}_{\rho,\epsilon}). \tag{6}$$

Next, we establish concentration around this expectation. Introduce independent auxiliary variables $U_{ij} \sim \mathrm{Unif}(0, 1)$, such that

$$Z_{ij} = \mathbb{I}\left[U_{ij} \le \rho s^\star(X_i, X_j)\right].$$

Changing one instance X_i affects at most $n - 1$ pairwise terms, giving the bounded-difference constant

$$c_i = \frac{(n-1)M_\epsilon}{\binom{n}{2}} = \frac{2M_\epsilon}{n}.$$

Changing one auxiliary variable U_{ij} affects only one term, with

$$d_{ij} = \frac{M_\epsilon}{\binom{n}{2}}.$$

Hence,

$$\sum_i c_i^2 + \sum_{i<j} d_{ij}^2 \le \frac{6M_\epsilon^2}{n}.$$

McDiarmid's inequality combined with (6) gives, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \left| R^\epsilon_{\mathrm{IS}}(f) - \widehat{R}^\epsilon_{\mathrm{IS},n}(f) \right| \le 2\mathfrak{R}_m(\mathcal{H}_{\rho,\epsilon}) + M_\epsilon \sqrt{\frac{3\log(1/\delta)}{n}}. \tag{7}$$

By Lemma 2 and the contraction inequality,

$$\mathfrak{R}_m(\mathcal{H}_{\rho,\epsilon}) \le L_{\rho,\epsilon}\mathfrak{R}_m(\mathcal{G}), \qquad L_{\rho,\epsilon} = \frac{\rho}{\epsilon}.$$

Substituting this bound into (7) yields

$$\sup_{f\in\mathcal{F}} \left| R^\epsilon_{\mathrm{IS}}(f) - \widehat{R}^\epsilon_{\mathrm{IS},n}(f) \right| \le 2\frac{\rho}{\epsilon}\mathfrak{R}_m(\mathcal{G}) + M_\epsilon \sqrt{\frac{3\log(1/\delta)}{n}},$$

which is the stated uniform generalization bound.

For completeness, let

$$\widehat{f}_{n,\epsilon} \in \arg\min_{f\in\mathcal{F}} \widehat{R}^\epsilon_{\mathrm{IS},n}(f), \qquad f^\star_\epsilon \in \arg\min_{f\in\mathcal{F}} R^\epsilon_{\mathrm{IS}}(f).$$

The standard ERM argument then gives

$$R^\epsilon_{\mathrm{IS}}(\widehat{f}_{n,\epsilon}) - R^\epsilon_{\mathrm{IS}}(f^\star_\epsilon) \le 2 \sup_{f\in\mathcal{F}} \left| R^\epsilon_{\mathrm{IS}}(f) - \widehat{R}^\epsilon_{\mathrm{IS},n}(f) \right|.$$

Consequently, with probability at least $1-\delta$,

$$R^\epsilon_{\mathrm{IS}}(\widehat{f}_{n,\epsilon}) - R^\epsilon_{\mathrm{IS}}(f^\star_\epsilon) \le 4\frac{\rho}{\epsilon}\mathfrak{R}_m(\mathcal{G}) + 2M_\epsilon \sqrt{\frac{3\log(1/\delta)}{n}}.$$

□

B.7 Proof of Proposition 2

Proof. Let $Q \in \mathcal{P}_K$ and define $\widetilde{f}(x) = Qf(x)$. Since Q only permutes the latent output coordinates,

$$g_{\widetilde{f}}(x,x') = \widetilde{f}(x)^\top \widetilde{f}(x') = f(x)^\top Q^\top Q f(x') = g_f(x,x').$$

Moreover, the predicted marginal transforms as $\mu_{\widetilde{f}} = Q\mu_f$. Therefore,

$$\mathrm{KL}_{\mathrm{perm}}(\pi\|\mu_{\widetilde{f}}) = \min_{P\in\mathcal{P}_K} D_{\mathrm{KL}}(\pi\|PQ\mu_f) = \min_{R\in\mathcal{P}_K} D_{\mathrm{KL}}(\pi\|R\mu_f) = \mathrm{KL}_{\mathrm{perm}}(\pi\|\mu_f),$$

where the second equality follows from the fact that the mapping $P \mapsto PQ$ is a bijection on the permutation group $\mathcal{P}_K$. Hence, both the pairwise similarity term and the class-prior regularizer are invariant to global permutations of the latent coordinates. □

B.8 Proof of Proposition 3

Proof. First assume that $f^\star$ recovers the ground-truth classes up to a permutation. Then there exists a permutation matrix $Q \in \mathcal{P}_K$ such that $\mu_{f^\star} = Q\pi$. Choosing $P = Q^\top$ gives $P\mu_{f^\star} = \pi$, and therefore

$$\mathrm{KL}_{\mathrm{perm}}(\pi\|\mu_{f^\star}) = \min_{P\in\mathcal{P}_K} D_{\mathrm{KL}}(\pi\|P\mu_{f^\star}) = 0.$$

Next, suppose that for some latent coordinate r, $\mu_{f,r} = 0$. Since permutation matrices only reorder coordinates, every $P\mu_f$ also contains at least one zero entry. Because the target prior has full support, $\pi_k > 0, k = 1, \dots, K$, there exists an index k such that $(P\mu_f)_k = 0$. Hence,

$$D_{\mathrm{KL}}(\pi \| P\mu_f) = \sum_{k=1}^{K} \pi_k \log \frac{\pi_k}{(P\mu_f)_k} = +\infty.$$

Since this holds for every $P \in \mathcal{P}_K$, $\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_f) = +\infty$. Finally, a fully collapsed classifier $f(x) = e_c$ almost surely satisfies $\mu_f = e_c$, which contains zero coordinates whenever $K > 1$. Therefore, its prior discrepancy is infinite. □

B.9 Proof of Theorem 2

Proof. Let $f_0 \in \arg\min_{f \in \mathcal{F}} R^0_{\mathrm{IS}}(f)$ be a population minimizer of the incomplete-similarity risk. By By Proposition 1 and Corollary 1, $g_{f_0}(x, x') = s^\star(x, x')$ almost surely, and f_0 recovers the ground-truth classifier up to a global permutation. Therefore, Proposition 3 implies $\mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_{f_0}) = 0$. Hence, $R_{\mathrm{PRIME}}(f_0) = R^0_{\mathrm{IS}}(f_0)$ for every $\lambda_\pi \geq 0$. For any $f \in \mathcal{F}$,

$$\begin{aligned} R_{\mathrm{PRIME}}(f) &= R^0_{\mathrm{IS}}(f) + \lambda_\pi \, \mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_f) \\ &\geq R^0_{\mathrm{IS}}(f) \\ &\geq R^0_{\mathrm{IS}}(f_0) \\ &= R_{\mathrm{PRIME}}(f_0). \end{aligned}$$

Therefore, f_0 is also a population minimizer of the PRIME objective.

Now let $f^\star_{\mathrm{P}}$ be any population minimizer of R_{PRIME}. Since f_0 is feasible and achieves the minimum,

$$R_{\mathrm{PRIME}}(f^\star_{\mathrm{P}}) = R_{\mathrm{PRIME}}(f_0).$$

Combining this equality with $R^0_{\mathrm{IS}}(f^\star_{\mathrm{P}}) \geq R^0_{\mathrm{IS}}(f_0)$ and $\lambda_\pi \, \mathrm{KL}_{\mathrm{perm}}(\pi \| \mu_{f^\star_{\mathrm{P}}}) \geq 0$, both inequalities must be tight. Consequently, $R^0_{\mathrm{IS}}(f^\star_{\mathrm{P}}) = R^0_{\mathrm{IS}}(f_0)$, and therefore $f^\star_{\mathrm{P}}$ is also a population minimizer of the incomplete-similarity risk.

Applying Proposition 1 and Corollary 1 again gives $g_{f^\star_{\mathrm{P}}}(x, x') = s^\star(x, x')$ almost surely, and there exists a permutation σ such that $f^\star_{\mathrm{P}}(x) = e_{\sigma(y(x))}$ almost surely. Finally, Proposition 3 yields

$$\mathrm{KL}_{\mathrm{perm}}\left(\pi \| \mu_{f^\star_{\mathrm{P}}}\right) = 0.$$

Thus, the stated population consistency properties hold for every $\lambda_\pi \geq 0$. □

B.10 Proof of Proposition 4

Proof. Recall that

$$d_\rho(s, g) = D_{\mathrm{KL}}\big(\mathrm{Bern}(\rho s) \,\|\, \mathrm{Bern}(\rho g)\big), \qquad s \in \{0, 1\}, \quad g \in [0, 1].$$

We consider the two possible values of s separately.

For $s = 0$,

$$d_\rho(0, g) = D_{\mathrm{KL}}\big(\mathrm{Bern}(0) \,\|\, \mathrm{Bern}(\rho g)\big) = -\log(1 - \rho g).$$

Therefore,

$$\frac{d_\rho(0, g)}{\rho} = \frac{-\log(1 - \rho g)}{\rho}.$$

Since

$$\lim_{\rho \downarrow 0} \frac{-\log(1 - \rho g)}{\rho} = g,$$

we obtain

$$\lim_{\rho \downarrow 0} \frac{d_\rho(0, g)}{\rho} = g.$$

For $s = 1$ and $g \in (0, 1]$,

$$\begin{aligned} d_\rho(1, g) &= D_{\mathrm{KL}}\big(\mathrm{Bern}(\rho) \,\|\, \mathrm{Bern}(\rho g)\big) \\ &= \rho \log \frac{\rho}{\rho g} + (1-\rho) \log \frac{1-\rho}{1-\rho g} \\ &= -\rho \log g + (1-\rho)\left[\log(1-\rho) - \log(1-\rho g)\right]. \end{aligned}$$

Dividing by ρ gives

$$\frac{d_\rho(1, g)}{\rho} = -\log g + \frac{1-\rho}{\rho}\left[\log(1-\rho) - \log(1-\rho g)\right].$$

Now,

$$\lim_{\rho \downarrow 0} \frac{\log(1-\rho) - \log(1-\rho g)}{\rho} = -(1-g),$$

which follows, for example, by l'Hospital's rule. Hence

$$\lim_{\rho \downarrow 0} \frac{d_\rho(1, g)}{\rho} = -\log g - (1-g).$$

If $g = 0$, then for every $\rho > 0$,

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(\rho) \,\|\, \mathrm{Bern}(0)\big) = +\infty,$$

so the corresponding limit is $+\infty$.

It remains to establish the consequence for the population excess risk. Under the realizability condition,

$$\Delta_\rho(f) = \mathbb{E}\left[d_\rho\big(s^\star(X, X'), g_f(X, X')\big)\right].$$

For brevity, write $S = s^\star(X, X'), G = g_f(X, X')$. We first construct an integrable dominating function. For $S = 0$,

$$\frac{d_\rho(0, G)}{\rho} = \frac{-\log(1-\rho G)}{\rho}.$$

Using

$$-\log(1-x) \le \frac{x}{1-x}, \qquad 0 \le x < 1,$$

we obtain, for $0 < \rho \le 1/2$,

$$0 \le \frac{d_\rho(0, G)}{\rho} \le \frac{G}{1-\rho G} \le 2G.$$

For $S = 1$ and $G > 0$,

$$\frac{d_\rho(1, G)}{\rho} = -\log G + \frac{1-\rho}{\rho} \log \frac{1-\rho}{1-\rho G}.$$

Since $G \le 1$,

$$1 - \rho \le 1 - \rho G,$$

and hence

$$\log \frac{1-\rho}{1-\rho G} \le 0.$$

Because the KL divergence is nonnegative,

$$0 \le \frac{d_\rho(1, G)}{\rho} \le -\log G.$$

Therefore, for all sufficiently small ρ,

$$0 \le \frac{d_\rho(S, G)}{\rho} \le S(-\log G) + 2(1-S)G.$$

By assumption,

$$\mathbb{E}\left[S(-\log G) + (1-S)G\right] < \infty,$$

so the right-hand side above is integrable. In particular, this assumption also implies $G > 0$ almost surely on the event $S = 1$.

The dominated convergence theorem therefore yields

$$\lim_{\rho \downarrow 0} \frac{\Delta_\rho(f)}{\rho} = \mathbb{E}\left[\lim_{\rho \downarrow 0} \frac{d_\rho(S, G)}{\rho}\right] = \mathbb{E}\left[S\big(-\log G - (1 - G)\big) + (1 - S)G\right].$$

Equivalently,

$$\lim_{\rho \downarrow 0} \frac{\Delta_\rho(f)}{\rho} = \mathbb{E}\left[s^\star(X, X')\left(-\log g_f(X, X') - \big(1 - g_f(X, X')\big)\right) + \big(1 - s^\star(X, X')\big) g_f(X, X')\right].$$

The assumed integrability guarantees that this limit is finite. □